Projet Informatique

Année universitaire 2018–2019

Assembleur MIPS

Équipe pédagogique :
Nicolas Castagné, *François* Cayre, *Michel* Desvignes, *Kattel* Morin-Allory, *François* Portet, *Jérémy* Riffet

Avant-propos

La compilation est une notion fondamentale de l'informatique. L'assemblage, qui est un type de compilation particulier, consiste à concevoir des méthodes permettant de transformer des programmes textes écrits en langage assembleur en code binaire directement compréhensible par la machine cible. La compilation couvre les domaines des langages de programmation, de l'algorithmique, du génie logiciel et de l'architecture des calculateurs. Les techniques employées dans la réalisation d'un compilateur sont si générales qu'on les retrouve régulièrement dans la conception d'autres programmes informatiques.

Ce projet a pour but de réaliser un compilateur de programmes écrits en langage assembleur MIPS. Ce programme, que l'on appelle *Assembleur* et qui est beaucoup plus simple qu'un compilateur de langage de haut-niveau, sera écrit en langage C. Durant la réalisation de cet assembleur vous aborderez les notions d'analyse lexicale, d'analyse syntaxique, de gestion des erreurs, de fichiers objets et bien d'autres qui vous permettront d'enrichir votre culture générale en informatique et votre savoir faire en programmation C.

Table des matières

1 Introduction **2**

2 Le MIPS et son langage assembleur **3**
2.1 Exécution d'une instruction . . . 3
2.1.1 La mémoire . . . 3
2.1.2 Les registres . . . 4
2.1.3 Les instructions . . . 5
2.2 Le langage d'assemblage MIPS . . . 7
2.2.1 Les commentaires . . . 7
2.2.2 Les instructions machines . . . 7
2.2.3 Les directives . . . 8
2.2.4 Les modes d'adressage . . . 10
2.3 Instructions étudiées dans le projet . . . 11
2.3.1 Catégories d'instructions . . . 11
2.3.2 Détails des instructions à prendre en compte dans le projet . . . 13

3 Programmer un assembleur **17**
3.1 Compilateur . . . 17
3.2 Analyse lexicale . . . 18
3.3 Analyse grammaticale . . . 19
3.4 Analyse sémantique . . . 21
3.5 Code intermédiaire et optimisation/modification de code . . . 21
3.6 Génération du code binaire relogeable . . . 22
3.6.1 Nécessité d'un code relogeable . . . 22
3.6.2 Codage en position, *flat binary file* . . . 24
3.6.3 Code relogeable . . . 24
3.7 Informations nécessaires à la relocation . . . 26
3.7.1 Table de relocation . . . 26
3.7.2 Champ addend . . . 26
3.7.3 Modes de relocation du MIPS . . . 26
3.8 Format `elf` . . . 27

4 À propos de la mise en œuvre **28**
4.1 Quelques conseils sur la programmation . . . 28
4.1.1 Programmation incrémentale . . . 28
4.1.2 Se familiariser avec le domaine d'application . . . 28
4.1.3 Conception et développement . . . 29
4.1.4 Développement dirigé par les tests . . . 29
4.2 Quelques conseils sur le projet assembleur . . . 30
4.2.1 Réflexions sur le format des données . . . 30
4.2.2 Machine à états finis (*FSM*) . . . 31
4.2.3 Découper une chaîne de caractères en *token*, `strtok` pour la représentation des instructions MIPS . . . 32

5 Travail à réaliser **35**
5.1 Objectif général : . . . 35
5.1.1 Fichier objet binaire au format ELF . . . 36
5.1.2 Vérification du fichier objet . . . 36
5.2 Étapes de développement du programme . . . 37
5.2.1 [5 pts] Étape 1 : Analyse lexicale . . . 37
5.2.2 [5 pts] Étape 2 : Analyse syntaxique – 1 . . . 37
5.2.3 [5 pts] Étape 3 : Analyse syntaxique – 2 . . . 39
5.2.4 [5 pts] Étape 4 : Génération de code . . . 39
5.3 Bonus : extensions du programme . . . 39
5.4 Agenda et organisation du projet . . . 40

Bibliographie **41**

A Compilation d'un programme en assembleur MIPS **42**
A.1 Installation . . . 42
A.2 Compilation et étude des fichiers . . . 42
A.2.1 Assemblage . . . 42
A.2.2 Désassemblage . . . 42
A.2.3 Edition de lien . . . 44

B ELF : Executable and Linkable Format **45**
B.1 Fichier objet au format ELF . . . 45
B.2 Structure générale d'un fichier objet au format ELF et principe de la relocation . . . 46
B.3 Exemple de fichier relogeable . . . 49
B.4 Détail des sections . . . 50
B.4.1 L'en-tête . . . 50
B.4.2 La table des noms de sections (`.shstrtab`) . . . 50
B.4.3 La section table des chaînes (`.strtab`) . . . 51
B.4.4 La section `.text` . . . 51
B.4.5 La section `.data` . . . 52
B.4.6 La section table des symboles . . . 52
B.4.7 Les sections de relocation . . . 53
B.4.8 Autres sections . . . 56

C Spécifications détaillées des instructions **57**
C.1 Définitions et notations . . . 57

Chapitre 1

Introduction

L'objectif de ce projet informatique est de concevoir puis mettre en œuvre en langage C, un assembleur pour un microprocesseur MIPS.

Le rôle d'un assembleur est de transformer un programme écrit dans un langage informatique accessible à l'homme, **ici le langage assembleur MIPS**, en un programme décrivant la même série d'instructions mais cette fois-ci en **langage machine**, c'est-à-dire représenté en code binaire, directement compréhensible par le processeur. Dans notre cas, il s'agira d'un processeur MIPS 32 bits.

Le logiciel doit donc prendre en entrée un fichier texte contenant un programme écrit dans le langage d'assemblage de la machine MIPS, et produire un fichier objet binaire contenant la traduction en langage machine du programme assemblé au format ELF.

Pour ce projet nous considérerons en fait un microprocesseur simplifié, n'acceptant qu'un jeu réduit des instructions du MIPS.

Le chapitre 2 donne une présentation générale du microprocesseur et introduit le langage assembleur considéré et le sous-ensemble des instructions du MIPS à gérer dans le projet. Les chapitres 3, 4 et 5 présentent respectivement quelques notions sur la manière de programmer un assembleur, les considérations importantes pour la mise en œuvre, puis des informations sur l'organisation générale du projet.

L'intérêt pédagogique de ce projet informatique est multiple. Il permet tout d'abord de travailler sur un projet de taille importante sous tous ses aspects techniques (analyse d'un problème, conception puis mise en œuvre d'une solution, validation du résultat) mais aborde aussi les notions de gestion de projet et de respect d'un planning. Ce projet vous permettra également d'améliorer votre connaissance et maîtrise du langage C, qui est particulièrement utilisé pour la programmation scientifique et le développement industriel, ainsi que des outils de développement associés (systèmes Unix/Linux, outil Make, débogueur, etc.). Enfin, il illustre et met en pratique les connaissances relatives aux microprocesseurs, vues notamment dans le cours d'ordinateurs et microprocesseurs de première année et dans certains cours d'architecture ou de micro-électronique.

Chapitre 2

Le MIPS et son langage assembleur

MIPS, pour *Microprocessor without Interlocked Pipeline Stages*, est un microprocesseur RISC 32 bits. RISC signifie qu'il possède un jeu d'instructions réduit (*Reduced Instruction Set Computer*) mais qu'en contrepartie, il est capable de terminer l'exécution d'une instruction à chaque cycle d'horloge. Les processeurs MIPS ont été utilisés dans de nombreuses stations de travail et consoles de jeux (Silicon Graphics, Nintendo 64, Sony PlayStation 2...). De nos jours, ils sont surtout utilisés dans les systèmes embarqués (imprimantes, les routeurs, automobile ...).

2.1 Exécution d'une instruction

Les microprocesseurs RISC sont basés sur un modèle en pipeline pour exécuter les instructions. Cette structure permet d'exécuter chaque instruction en plusieurs cycles, mais de terminer l'exécution d'une instruction à chaque cycle. Cette structure en pipeline est illustrée sur la Figure 2.1. Il s'agit d'une architecture type en *load/store*. Le pipeline interagit d'un coté avec la mémoire (RAM) pour lire le programme et modifier les données du programme (*load/store* et d'un autre avec le microprocesseur et ses registres (mémoire à accès rapide) pour exécuter les instructions. L'extraction (*Instruction Fetch - IF*) va récupérer en mémoire l'instruction à exécuter. Le décodage (*Instruction Decode - ID*) interprète l'instruction et résout les adresses des registres. L'exécution (*Execute-EX*) utilise l'unité arithmétique et logique pour exécuter l'opération. L'accès en mémoire (*Memory - MEM*) est utilisé pour transférer le contenu d'un registre vers la mémoire ou vice-versa. Enfin, l'écriture registre (*Write Back - WB*) met à jour la valeur de certains registres avec le résultat de l'opération. Ce pipeline permet d'obtenir les très hautes performances qui caractérisent le MIPS. En effet, comme les instructions sont de taille constante et que les étages d'exécution sont indépendants, il n'est pas nécessaire d'attendre qu'une instruction soit complètement exécutée pour démarrer l'exécution de la suivante. Par exemple, lorsqu'une instruction atteint l'étage ID une autre instruction peut être prise en charge par l'étage IF. Dans le cas idéal, 5 instructions sont constamment dans le pipeline. Bien entendu, certaines contraintes impliquent des ajustements comme dans le cas où une instruction dépend de la précédente. Une des difficultés majeures de l'assembleur est de permettre de décharger le programmeur de ces contraintes et de les gérer lors de la compilation. C'est un problème que nous n'aborderons pas mais qu'il est nécessaire de connaître pour interpréter les résultats des assembleurs actuels.

2.1.1 La mémoire

Le microprocesseur MIPS possède une mémoire de 4 Go (2^{32} bits) adressable par octets. C'est dans cette mémoire qu'on charge la suite des instructions du microprocesseur contenues dans un programme binaire exécutable (ces instructions sont des mots de 32 bits). Pour exécuter un tel programme, le microprocesseur vient chercher séquentiellement les instructions dans cette mémoire, en se repérant grâce à un compteur programme (PC) contenant l'adresse en mémoire de la prochaine instruction à exécuter. Il est à noter que toutes les instructions sont alignées sur 4 octets.

L'adresse d'un octet en mémoire correspond au rang qu'il occupe dans le tableau des 4 Go qui la constitue. Ces adresses sont codées sur 32 bits, et sont contenues dans l'intervalle 0x00000000 à 0xFFFFFFFF.

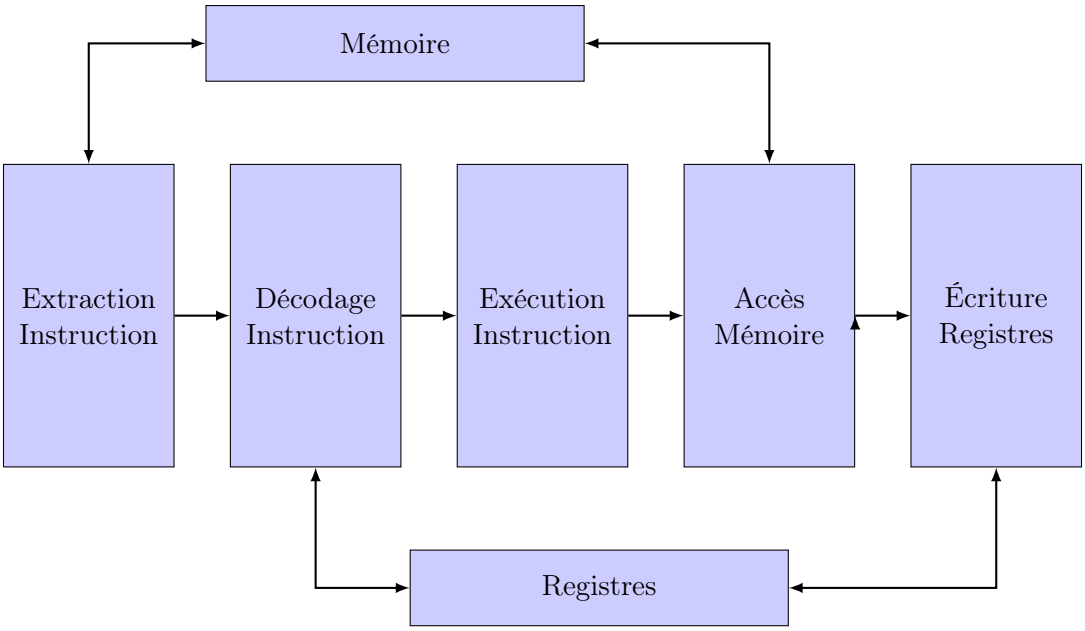


FIGURE 2.1 – Architecture interne simplifiée du microprocesseur RISC 32 bits

Pour stocker en mémoire des valeurs sur plusieurs octets, par exemple un mot sur 4 octets, deux systèmes existent (figure 2.2) :

- les systèmes de type *big endian* écrivent l'octet de poids le plus fort à l'adresse la plus basse. Les processeurs MIPS sont *big endian*, ainsi que les processeurs Motorola et certains ARM.
- les systèmes de type *little endian* écrivent l'octet de poids le plus faible à l'adresse la plus basse. Les processeurs Intel et AMD notamment sont *little endian*.

Adresse :	*Contenu de la mémoire :*	
	big endian	little endian
	...	...
0x00000004	0xFF	0xCC
	0xEE	0xDD
	0xDD	0xEE
0x00000007	0xCC	0xFF
	...	...

FIGURE 2.2 – Mode d'écriture en mémoire de la valeur hexadécimale *0xFFEEDDCC* pour un système *big endian* ou *little endian*. Le MIPS est un processeur de type *big endian* : l'octet de poids fort se trouve à l'adresse la plus basse.

2.1.2 Les registres

Les registres sont des emplacements mémoire spécialisés utilisés par les instructions et se caractérisant principalement par un temps d'accès rapide. La quasi-majorité des instructions n'agissent que sur des registres.

Les registres d'usage général

La machine MIPS dispose de 32 registres d'usage général (General Purpose Registers) de 32 bits chacun, dénotés `$0` à `$31`. Les registres peuvent également être identifiés par un mnemonic indiquant leur usage conventionel. Par exemple, le registre `$29` est noté `$sp`, car il est utilisé (par convention !) comme le pointeur de pile (sp pour *Stack Pointer*). Dans les programmes, un registre peut être désigné par son numéro aussi bien que son nom (par exemple, `$sp` équivaut à `$29`).

La figure 2.3 résume les conventions et restrictions d'usage que nous retiendrons pour ce projet.

Mnémonique	Registre	Usage
`$zero`	`$0`	Registre toujours nul, même après une écriture
`$at`	`$1`	*Assembler temporary* : registre réservé à l'assembleur
`$v0`, `$v1`	`$2`, `$3`	Valeurs retournées par une sous-routine
`$a0`-`$a3`	`$4`-`$7`	Arguments d'une sous-routine
`$t0`-`$t7`	`$8`-`$15`	Registres temporaires
`$s0`-`$s7`	`$16`-`$23`	Registres temporaires, préservés par les sous-routines
`$t8`, `$t9`	`$24`, `$25`	Deux registres temporaires de plus
`$k0`, `$k1`	`$26`, `$27`	kernel (réservés !)
`$gp`	`$28`	Global pointer (on évite d'y toucher !) [1]
`$sp`	`$29`	*Stack pointer* : pointeur de pile
`$fp`	`$30`	Frame pointer (pas utilisé dans le projet)
`$ra`	`$31`	*Return address* : utilisé par certains instructions (`JAL`) pour sauver l'adresse de retour d'un saut

FIGURE 2.3 – Conventions d'usage des registres MIPS.

Les registres spécialisés

En plus des registres généraux, plusieurs autres registres spécialisés sont utilisés par le MIPS :

- Le compteur programme 32 bits `PC`, qui contient l'adresse mémoire de la prochaine instruction. Il est incrémenté après l'exécution de chaque instruction, et modifié en cas de sauts ou branchements.
- Deux registres 32 bits `HI` et `LO` utilisés pour stocker le résultat de la multiplication ou de la division de deux données de 32 bits. Leur utilisation est décrite section 2.3.2.

D'autres registres existent, mais ils ne seront pas utilisés dans ce projet (`EPC` pour les exceptions, registres de valeurs flottantes, ...).

2.1.3 Les instructions

Le MIPS possède une large gamme d'instructions, plus de 280 ! Les spécifications des instructions étudiées dans ce projet sont données dans l'annexe C. Elles sont directement issues de la documentation du MIPS fournie par le *Software User's Manual* de *Architecture For Programmers Volume II* de *MIPS Technologies* [4]. Dans ce projet nous ne prendrons en compte qu'un nombre restreint d'instructions simples.

Nous donnons ici un exemple pour expliciter la spécification d'une instruction, l'opération addition (`ADD`). Dont la spécification, telle que donnée dans le manuel, est reportée ci dessous Figure 2.4. Toutes les instructions sont codées sur 32bits.

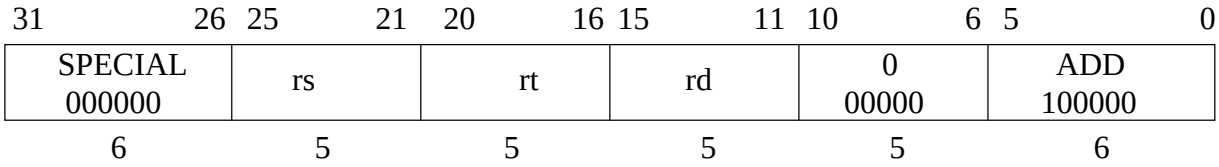


FIGURE 2.4 – Instruction ADD

Format: `ADD rd, rs, rt`

Purpose: `To add 32-bit integers. If an overflow occurs, then trap.`

Additionne deux nombres entiers sur 32-bits, si il y a un débordement, l'opération n'est pas effectuée.

Description: $GPR[rd] \leftarrow GPR[rs] + GPR[rt]$

```
The 32-bit word value in GPR rt is added to the 32-bit value in
GPR rs to produce a 32-bit result.
. If the addition results in 32-bit 2's complement arithmetic
  overflow, the destination register is not modified and an
  Integer Overflow exception occurs.
. If the addition does not overflow, the 32-bit result is placed
  into GPR rd.
```

No comment, juste un petit exercice pratique d'anglais .... Les descriptions données dans le manuel sont généralement très claires.

Restrictions: `None`

Operation:

```
temp <- (GPR[rs]31||GPR[rs]31..0) + (GPR[rt]31||GPR[rt]31..0)
if temp32 != temp31 then
   SignalException(IntegerOverflow)
else
   GPR[rd] <- temp
endif
```

ici l'opérateur `<-` est celui d'affectation, `||` est celui de concaténation, `GPR[rs]31` le bit de rang `31` du registre `rs` et `GPR[rt]31..0` le mot complet du registre `rs` (les bits de rang`0 à 31`).

Restriction: `Integer Overflow`

Programming Notes:

`ADDU performs the same arithmetic operation but does not trap on overflow.`

Exemple de codage pour les instructions `ADD` et `ADDI` :

```
ADD $2, $3, $4                  00641020
ADDI $2, $3, 200                206200C8
```

À vous de retrouver ceci à partir de la doc ! Un bon petit exercice pour bien comprendre...

2.2 Le langage d'assemblage MIPS

Pour programmer un MIPS ont utilise un langage assembleur spécifiquement dédié au MIPS. La syntaxe qui est présentée ici est volontairement moins permissive que celle de l'assembleur *GNU*. On se contente ici de présenter la syntaxe de manière intuitive.

Un programme se présente comme une liste d'unités, une unité tenant sur une seule ligne. Il est possible (et même recommandé pour aérer le texte) de rajouter des lignes blanches. Il y a trois sortes de lignes que nous allons décrire dans les sous-sections suivantes.

2.2.1 Les commentaires

C'est un texte optionnel non interprété par l'assembleur. Un commentaire commence sur une ligne par le caractère `#` et se termine par la fin de ligne.

Exemple

```
# Ceci est un commentaire. Il se termine à la fin de la ligne
ADD $2,$3,$4 # Ceci est aussi un commentaire, qui suit une instruction ADD
```

2.2.2 Les instructions machines

Elles ont la forme générale ci-dessous, les champs entre crochets indiquant des champs optionnels. Une ligne peut ne comporter qu'un champ étiquette, une opération peut ne pas avoir d'étiquette associée, ni d'opérande, ni de commentaire. Les champs doivent être séparés par des séparateurs qui sont des combinaisons d'espaces et/ou de tabulations.

```
[étiquette] [opération [opérandes]] [# commentaire]
```

Les sections suivantes présentent la syntaxe autorisée pour chacun des champs.

Le champ *étiquette*

C'est la désignation symbolique d'une adresse de la mémoire qui peut servir d'opérande à une instruction ou à une directive de l'assembleur. Une étiquette est une suite de caractères alphanumériques (sans espace) qui ne doit PAS commencer par un chiffre[2]. Cette chaîne est suivie par le caractère « :». Le nom de l'étiquette est la chaîne de caractères alphanumériques située à gauche du caractère « :». Plusieurs étiquettes peuvent être associées à la même opération ou à la même directive.

Une étiquette ne peut être définie qu'une seule fois dans une unité de compilation. Sa valeur lors de l'exécution est égale à son adresse d'implantation dans la mémoire après le chargement. Elle dépend donc de la section dans laquelle elle est définie et de sa position dans cette section (cf. section 2.2.3).

Exemple

```
etiq1:
_etiq2:
etiq3:  ADD $2,$3,$4  # les trois étiquettes repèrent la même instruction ADD
```

2. En réalité une étiquette peut contenir également les caractères : «.»,«_»,«$». Mais pour simplifier la conception de l'assembleur, seul «_» sera accepté.

Le champ *opération*

Il indique soit un des mnémoniques d'instructions du processeur MIPS, soit une des directives de l'assembleur.

Exemple

```
ADD $2,$3,$4   # le champ opération à la valeur ADD
.space 32      # le champ opération à la valeur .space
```

Le champ *opérandes*

Le champ *opérandes* a la forme : `opérandes = [op1] [,op2] [,op3]`

Ce sont les opérandes éventuels si l'instruction ou la directive en demande. S'il y en a plusieurs, ces opérandes sont séparés par des virgules.

Exemple

```
ADD $2,$3,$4   # les opérandes sont $2, $3 et $4
.space 32      # l'opérande est 32
```

2.2.3 Les directives

Une directive commence toujours par un point («.») et peut être suivie d'opérandes. Les directives sont des commandes à destinations de l'assembleur et non pas du microprocesseur MIPS. Elle permettent de spécifier comment le programme doit être assemblé. Nous considérerons 2 familles de directives : les directives de sectionnement du programme et les directives de définition de données.

Directive	**Description**
`.text`	Ce qui suit doit aller dans le segment TEXT
`.data`	Ce qui suit doit aller dans le segment DATA
`.bss`	Ce qui suit doit aller dans le segment BSS
`.set option`	Instruction à l'assembleur pour inhiber ou non certaine options. Dans notre cas seule l'option *noreorder* est considérée
`.word w1, ..., wn`	Met les n valeurs sur 32 bits dans des mots successifs (ils doivent être alignés!)
`.byte b1, ..., bn`	Met les n valeurs sur 8 bits dans des octets successifs
`.asciiz s1, ..., sn`	Met les n chaînes de caractères à la suite en mémoire. Chaque chaîne est terminée par $\backslash 0$.
`.space` n	Réserve n octets en mémoire. Les octets sont initialisés à zéro.

Directives de sectionnement

Bien que le processeur MIPS n'a qu'une seule zone mémoire contenant à la fois les instructions et les données (ce qui n'est pas le cas de tous les microprocesseurs), trois directives existent en assembleur pour spécifier les sections de code et de données.

- la section `.text` contient le code du programme (instructions) .
- la section `.data` est utilisée pour définir les données du programme.
- la section `.bss` permet de déclarer les données non initialisées. Ces données ne prennent ainsi pas de place dans le fichier binaire du programme. Elles ne seront effectivement allouées qu'au moment du chargement du programme où elles seront initialisées à zéro.

Les directives de sectionnement s'écrivent par leur nom de section : `.text`, `.data` ou `.bss`. Elles indiquent à l'assembleur d'assembler les lignes suivantes dans les sections correspondantes.

Remarque : Dans ce projet, les instructions doivent absolument se trouver dans une section `.text`, la section `.data` ne peut contenir que des directives de données et la section `.bss` uniquement la directive `.space`. Tout écart de ces contraintes entraînera une erreur et fera sortir du programme d'assemblage. Notez que ces contraintes ne sont pas toutes présentes dans la norme (elles dépendent de l'assembleur).

Les directives de définition de données

On distingue les données initialisées des données non initialisées.

Déclaration des données non initialisées Pouvoir réserver un espace sans connaître la valeur qui y sera stockée est une capacité importante de tout langage. Le langage assembleur MIPS fournit la directive suivante.

[*étiquette*] `.space` ***taille*** La directive `.space` permet de réserver un nombre d'octets égal à *taille* à l'adresse *étiquette*. Les octets sont initialisés à zéro.

```
toto: .space 13
```

La directive `.space` se trouve normalement dans une section de données `.bss`. Mais `.space` peut également se trouver dans une section data ou l'espace sera effectivement réservé sur le disque.

Déclaration de données initialisées L'assembleur permet de déclarer plusieurs types de données initialisées : des octets, des mots (32 bits), des chaînes de caractères, etc. Dans ce projet, on ne s'intéressera qu'aux directives de déclaration suivantes :

[*étiquette*] `.byte` ***valeur*** *valeur* peut être soit un entier signé sur 8 bits, soit une constante symbolique dont la valeur est comprise entre -128 et 127, soit une valeur hexadécimale dont la valeur est comprise entre 0x0 et 0xff. Par exemple, les lignes ci-dessous permettent de réserver deux octets avec les valeurs initiales -4 et 0xff sous forme hexadécimale. Le premier octet est à l'adresse Tabb de la mémoire, l'autre à l'adresse Tabb+1.

```
Tabb:   .byte -4
        .byte 0xff
```

[*étiquette*] `.word` ***valeur*** *valeur* peut être soit un entier signé sur 32 bits, soit une constante symbolique dont la valeur est représentable sur 32 bits. Par exemple, la ligne suivante permet de réserver un mot de 32 bits avec la valeur initiale 32767 à l'adresse Tabw de la mémoire. La deuxième ligne permet de stoker l'adresse du tableau `Tabw` en mémoire. Cette valeur ne sera déterminée que lors du chargement du programme en mémoire.

```
   Tabw:   .word 0x00007fff
address:   .word Tabw
```

[*étiquette*] `.asciiz` ***valeur*** *valeur* doit commencer et terminer par un `"`. Par exemple, la ligne suivante permet de réserver 19 octets pour une chaîne de 18 caractères « il a dit "bonjour"». Notez le caractère \ d'échappement qui permet d'inclure le `"` sans que l'assembleur ne l'interprète comme une frontière de chaîne.

```
   Tabc:   .asciiz "il a dit \"bonjour\""
```

D'autres caractères particuliers doivent également être échappés tels que : \ (noté `"\\"`), `'` (noté `"\'"`), newline (noté `"\n"`) etc.

2.2.4 Les modes d'adressage

Comme nous le verrons au chapitre 2.3, les instructions du microprocesseur MIPS ont de zéro à quatre opérandes. On appelle *mode d'adressage* d'un opérande la méthode qu'utilise le processeur pour déterminer où se trouve l'opérande, c'est-à-dire pour déterminer son **adresse**. Le langage assembleur MIPS contient 5 modes d'adressage décrit ci dessous.

Adressage registre direct

Dans ce mode, la valeur de l'opérande est contenue dans un registre et l'opérande est désigné par le nom du registre en question.

Exemple :

```
ADD $2, $3, $4   # les valeur des opérandes sont dans les registres 3 et 4
                 # le résultat est placé dans le registre 2
```

Adressage immédiat

La valeur de l'opérande est directement fournie dans l'instruction.

Exemple :

```
ADDI $2, $3, 200   # valeur immédiate entière signée sur 16 bits
ADDI $2, $3, 0x3f  # idem avec une valeur immédiate hexadécimale
ADDI $2, $3, X     # ajout $2 à la valeur (et non le contenu) de X (adresse mémoire)
```

Adressage indirect avec base et déplacement

Dans ce mode, interviennent un registre appelé *registre de base* qui contient une adresse mémoire, et une constante signée (décimale ou hexadécimale) appelée *déplacement*. La syntaxe associée par l'assembleur à ce mode est `offset(base)`.

Pour calculer l'adresse de l'opérande, le processeur ajoute au contenu du registre de base `base` la valeur sur 2 octets du déplacement `offset`.

Exemple :

```
LW $2, 200($3)      # $2 = memory[($3) + 200]
```

Adressage absolu aligné dans une région de 256Mo

Un opérande de 26 bits permet de calculer une adresse mémoire sur 32 bits. Ce mode d'adressage est réservé aux instructions de sauts (`J`, `JAL`).

Les 28 bits de poids faibles de l'adresse de saut sont contenus dans l'opérande décalé de 2 bits vers la gauche (car les instructions sont alignées tous les 4 octets). Les poids forts manquants sont pris directement dans le compteur programme. Un exemple est donné au paragraphe 2.3.2.

Exemple :

```
J ma_sous_routine  # l'adresse de saut est calculée à partir de l'opérande (l'adresse
                   # de l'étiquette ma_sous_routine) et de la valeur de PC
```

Adressage relatif

Ce mode d'adressage est utilisé par les instructions de branchement. Lors du branchement, l'adresse de branchement est déterminée à partir d'un opérande `offset` sur 16 bits. Cette adresse est d'abord décalée de 2 bits vers la gauche puis ajoutée au compteur PC courant. Par exemple, un offset codé 0xFD dans l'instruction correspond en réalité à un offset de 0x3F4 ! La valeur sur 18 bits est ensuite ajoutée au compteur programme pour déterminer l'adresse de saut.

Exemple :

```
                          # si $2==$3, branchement à l'adresse PC + 0x3F4
  BEQ $2, $3, 0x3F4       # code binaire correspondant ->104300FD
```

2.3 Instructions étudiées dans le projet

Cette section présente les instructions et les pseudo-instructions du MIPS qui devront être traitées par l'assembleur. Toutes les instructions MIPS ne seront pas traitées (en particulier les entrées-sorties, la gestion des valeurs flottantes...). La syntaxe des instructions en langage assembleur est donnée, ainsi qu'une description et le codage binaire des opérations.

2.3.1 Catégories d'instructions

Les processeurs MIPS possèdent des instructions simples de taille constante égale à 32 bits [3]. Ceci facilite notamment les étapes d'extraction et de décodage des instructions, réalisées chacune dans le pipeline en un cycle d'horloge. Les instructions sont toujours codées sur des adresses alignées sur un mot, c'est-à-dire divisibles par 4. Cette restriction d'alignement favorise la vitesse de transfert des données.

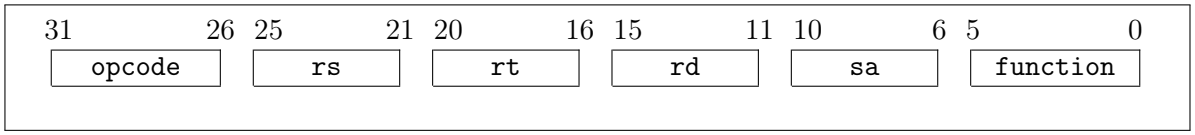
Il existe seulement trois formats d'instructions MIPS, *R-type*, *I-type* et *J-type*, dont les formes les plus courantes sont les suivantes :

```
R-instruction $rd, $rs, $rt
I-instruction $rt, $rs, immediate
J-instruction target
```

3. pour les séries R2000/R3000 auxquelles nous nous intéressons. Les processeurs récents sont sur 64 bits.

Les instructions de type R

Le codage binaire des instructions *R-type*, pour "register type", suit le format :

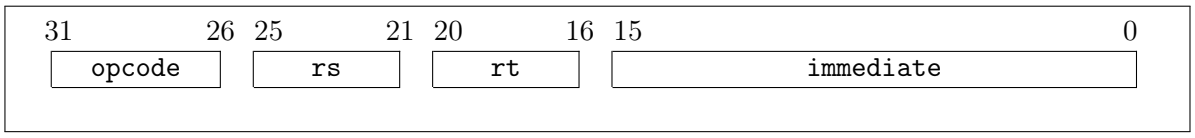


avec les champs suivants :

- le code binaire **opcode** (operation code) identifiant l'instruction. Sur 6 bits, il ne permet de coder que 64 instructions, ce qui même pour un processeur RISC est peu. Par conséquent, un champ additionnel **function** de 6 bits est utilisé pour identifier les instructions R-type.
- **rd** est le registre destination (valeur sur 5 bits, donc comprise entre 0 et 31, codant le numéro du registre)
- **rs** est le premier argument source
- **rt** est le second argument source
- **sa (shift amount)** est le nombre de bits de décalage, pour les instructions de décalage.
- **function** 6 bits additionnels pour le code des instructions R-type, en plus du champ **opcode**.

Les instructions de type I

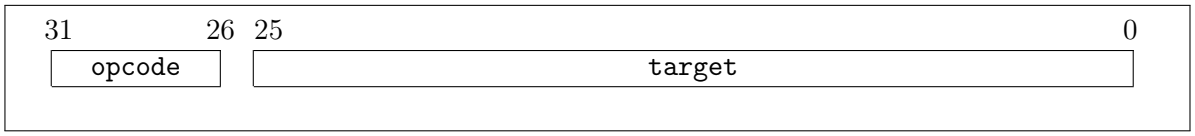
Le codage binaire des instructions *I-type*, pour "immediate type", suit le format :



avec **opcode** le code opération, **rt** le registre destination, **rs** le registre source et **immediate** une valeur numérique codée sur 16 bits.

Les instructions de type J

Le codage binaire des instructions *J-type*, pour "jump type", suit le format :



où **opcode** est le code opération et **target** une valeur de saut codée sur 26 bits.

2.3.2 Détails des instructions à prendre en compte dans le projet

Instructions arithmétiques

Mnemonic	Opérandes	Opération
ADD	$rd, $rs, $rt	$rd = $rs + $rt
ADDI	$rt, $rs, immediate	$rt = $rs + immediate
SUB	$rd, $rs, $rt	$rd = $rs - $rt
MULT	$rs, $rt	(HI, LO) = $rs * $rt
DIV	$rs, $rt	LO = $rs div $rt; HI = $rs mod $rt

ADD fait l'addition de deux registres rs et rt. Le résultat sur 32 bits de ces opérations est placé dans le registre rd.

ADDI est l'addition avec une valeur immédiate, SUB la soustraction. Le résultat sur 32 bits de ces opérations est stocké dans le registre rd. Les opérandes et le résultat sont des entiers signés sur 32 bits.

Pour la multiplication MULT, les valeurs contenues dans les deux registres rs et rt sont multipliées. La multiplication de deux valeurs 32 bits est un résultat sur 64 bits. Les 32 bits de poids fort du résultat sont placés dans le registre HI, et les 32 bits de poids faible dans le registre LO. Les valeurs de ces registres sont accessibles à l'aide des instructions MFHI et MFLO définies ci dessous.

La division DIV fournit deux résultats : le quotient de rs divisé rt est placée dans le registre LO, et le reste de la division entière dans le registre HI.

Les instructions logiques

Mnemonic	Opérandes	Opération
AND	$rd, $rs, $rt	$rd = $rs AND $rt
OR	$rd, $rs, $rt	$rd = $rs OR $rt
XOR	$rd, $rs, $rt	$rd = $rs XOR $rt

Les deux registres de 32 bits rs et rt sont combinés bit à bit selon l'opération logique effectuée. Le résultat est placé dans le registre 32 bits rd.

Les instructions de décalage

Mnemonic	Opérandes	Opération
ROTR	$rd, $rt, sa	$rd = $rt[sa-0] $rt[31-sa]
SLL	$rd, $rs, sa	$rd = $rt « sa
SRL	$rd, $rs, sa	$rd = $rt » sa

Le contenu du registre 32 bits rt est décalé à gauche pour SLL et à droite pour SRL de sa bits (en insérant des zéros). sa est une valeur immédiate sur 5 bits, donc entre 0 et 31. Pour ROTR le mot contenu dans le registre 32 bits rt subi une rotation par la droite. Le résultat est placé dans le registre rd.

Les instructions Set

Mnemonic	Opérandes	Opération
SLT	$rd, $rs, $rt	if $rs < $rt then $rd = 1, else $rd = 0

Le registre rd est mis à 1 si le contenu de rs est plus petit que celui de rt, à 0 sinon. Les valeurs rs et rt sont des entiers signés en complément à 2.

Les instructions Load/Store

Mnemonic	Opérandes	Opération
LW	$rt,offset($rs)	$rt = memory[$rs+offset]
SW	$rt, offset($rs)	memory[$rs+offset] = $rt
LUI	$rt,immediate	$rt = immediate « 16
MFHI	$rd	$rd = HI
MFLO	$rd	$rd = LO

— Load Word (LW) place le contenu du mot de 32 bits à l'adresse mémoire ($rs + offset) dans le registre rt. offset est une valeur signée sur 16 bits codée en complément à 2, elle est placée dans le champ immediate.

Exemple : LW $8, 0x60($10)

— Store Word (SW) place le contenu du registre rt dans le mot de 32 bits à l'adresse mémoire ($rs + offset). offset est une valeur signée sur 16 bits codée en complément à 2, elle est placée dans le champ immediate.

— Load Upper Immediate (LUI) place le contenu de la valeur entière 16 bits immediate dans les deux octets de poids fort du registre $rt et met les deux octets de poids faible à zéro.

— L'instruction Move from HI (MFHI) : Le contenu du registre HI est placé dans le registre rd. HI contient les 32 bits de poids fort du résultat 64 bits d'une instruction MULT ou le reste de la division entière d'une instruction DIV.

— L'instruction Move from LO (MFLO) est similaire à MFHI : le contenu du registre LO est placé dans le registre rd. LO contient les 32 bits de poids faible du résultat 64 bits d'une instruction MULT ou le quotient de la division entière d'une instruction DIV.

Les instructions de branchement et de saut

Mnemonic	Opérandes	Opération
BEQ	$rs, $rt, offset	Si ($rs = $rt) alors branchement
BNE	$rs, $rt, offset	Si ($rs != $rt) alors branchement
BGTZ	$rs, offset	Si ($rs > 0) alors branchement
BLEZ	$rs, offset	Si ($rs <= 0) alors branchement
J	target	PC=PC[31:28] target
JAL	target	GPR[31]=PC+8, PC=PC[31:28] target
JR	$rs	PC=$rs.

— BEQ effectue un branchement après l'instruction si les contenus des registres rs et rt sont égaux. L'offset signé de *18 bits* (16 bits décalés de 2) est ajouté à l'adresse de l'instruction de branchement pour déterminer l'adresse effective du saut.

— BNE effectue un branchement après l'instruction si les contenus des registres rs et rt sont différents. L'offset signé de *18 bits* (16 bits décalés de 2) est ajouté à l'adresse de l'instruction de branchement pour déterminer l'adresse effective du saut.

— BGTZ effectue un branchement après l'instruction si le contenu du registre rs est strictement positif. L'offset signé de *18 bits* (16 bits décalés de 2) est ajouté à l'adresse de l'instruction de branchement pour déterminer l'adresse effective du saut.

— BLEZ effectue un branchement après l'instruction si le contenu du registre rs est négatif ou nul. L'offset signé de *18 bits* (16 bits décalés de 2) est ajouté à l'adresse de l'instruction de branchement pour déterminer l'adresse effective du saut.

— J effectue un branchement aligné à 256 Mo dans la région mémoire du PC. Les *28 bits* de poids faible de l'adresse du saut correspondent au champ target, décalés de 2. Les 4 bits de poids fort restant correspondent au 4 bits de poids fort du compteur PC.

Exemple : Soit l'instruction `J ma_sous_routine`, localisée à l'adresse 0x56767296 et l'étiquette ma_sous_routine localisée à l'adresse 0x5AD52A8C. Quelle est la valeur de l'offset ?

Les quatre bits de poids fort de l'adresse de saut et du compteur PC doivent être les mêmes.

```
PC = 0x56767296 + 4 == 0x5676729A
0x5676729A --  0101 01100111011001111001010011010
0x5AD52A8C --  0101 10101101010100101010100001100
```

Le target sur 28 bits est donc l'adresse de saut sans les 4 bits de poids fort soit

```
adresse 28 bits:  1010110101010010101010001100
```

Comme toutes les instructions sont alignées sur des adresses multiples de 4, les deux bits de poids faible d'une instruction sont toujours 00. On peut donc décaler le champs `offset` de 2 bits, ce qui donne le target :

```
adresse 26 bits:  1010110101010010101010100011
```

La valeur finale du target à encoder est donc :

```
10101101010100101010100011 -- 0x2B54AA3
```

— `JAL` effectue un appel à une routine dans la région alignée de 256 Mo. Avant le saut, l'adresse de retour est placée dans le registre `$ra` (= `$31`). Il s'agit de l'adresse de l'instruction qui suit immédiatement le saut et où l'exécution reprendra après le traitement de la routine. Cette instruction effectue un branchement aligné à 256 Mo dans la région mémoire du PC. Les *28 bits* de poids faible de l'adresse du saut correspondent au champ `offset`, décalés de 2. Les poids forts restant correspondent au bits de poids fort de l'instruction.

— `JR` effectue un saut à l'adresse spécifiée dans `rs` Le contenu du registre 32 bits `rs` contient l'adresse du saut.

Les pseudo-instructions

Les pseudo-instructions ne sont pas définies parmi les instructions du processeur MIPS, mais uniquement au niveau de l'assembleur. Ces pseudo instructions permettent d'augmenter l'expressivité du langage afin de faciliter le travail du programmeur. La conséquence est une plus grande complexité pour la compilation. Lors de la compilation, l'assembleur doit les remplacer automatiquement par un équivalent (pas forcément unique) composé d'une ou plusieurs instructions en langage machine.

Mnemonic	Opérandes	Opération équivalente
NOP		SLL $0, $0, 0
LW	$rt, target	LUI $rt, upper_16(target) LW $rt, lower_16(target)($rt)
SW	$rt, target	LUI $at, upper_16(target) SW $rt, lower_16(target)($at)
MOVE	$rt, $rs	ADD $rt, $rs, $zero
NEG	$rt, $rs	SUB $rt, $zero, $rs
LI	$rt, immediate	ADDI $rt, $zero, immediate
BLT	$rt, $rs, target	SLT $1, $rt, $rs BNE $1, $zero, target

— `NOP` n'effectue aucun traitement, seul le compteur programme est incrémenté.

— `LW` et `SW` sont les instructions Load Word et Store Word vues plus haut mais prenants une étiquette en paramètre. `upper_16` et `lower_16` signifient respectivement de prendre les 16 bits de poids fort et de poids faible de l'adresse de l'étiquette sur 32 bits.
— `MOVE` copie le contenu du registre `$rs` dans le registre destination `$rt`.
— `NEG` copie l'opposé du contenu du registre `$rs` (`-$rs`) dans le registre `$rt`.
— `LI` place la valeur immédiate dans le registre destination `$rt`
— `BLT` permet un branchement suite à la comparaison directe de deux registres, alors qu'on ne peut comparer qu'à zéro avec les instructions. Si le contenu du registre `$rs` est inférieur à celui du registre `$rt`, le branchement vers `target` est effectué après l'instruction.

Chapitre 3

Programmer un assembleur

Le but de ce projet est de concevoir un compilateur de programmes écrits en assembleur. Ce type de compilateur est appelé *assembleur* (confusion entre le langage et le compilateur) et reste beaucoup plus simple à concevoir qu'un compilateur pour langage de haut niveau. Cependant, les principes généraux exposés dans cette section peuvent s'appliquer à tout type de compilateurs (p.ex. compilateur C, Pascal, etc.). Cette section donne quelques règles simples pour développer proprement un assembleur et quelques éléments de réflexion sur la façon de procéder.

3.1 Compilateur

D'une manière générale, un compilateur est un logiciel qui lit un programme écrit dans un premier langage — le langage *source* — et le traduit en un programme équivalent dans un autre langage — le langage *cible*. Dans notre cas, le langage source est l'assembleur MIPS et le langage cible est le code binaire MIPS. La transformation d'un programme source vers un programme cible nécessite plusieurs phases qui sont décrites figure 3.1.

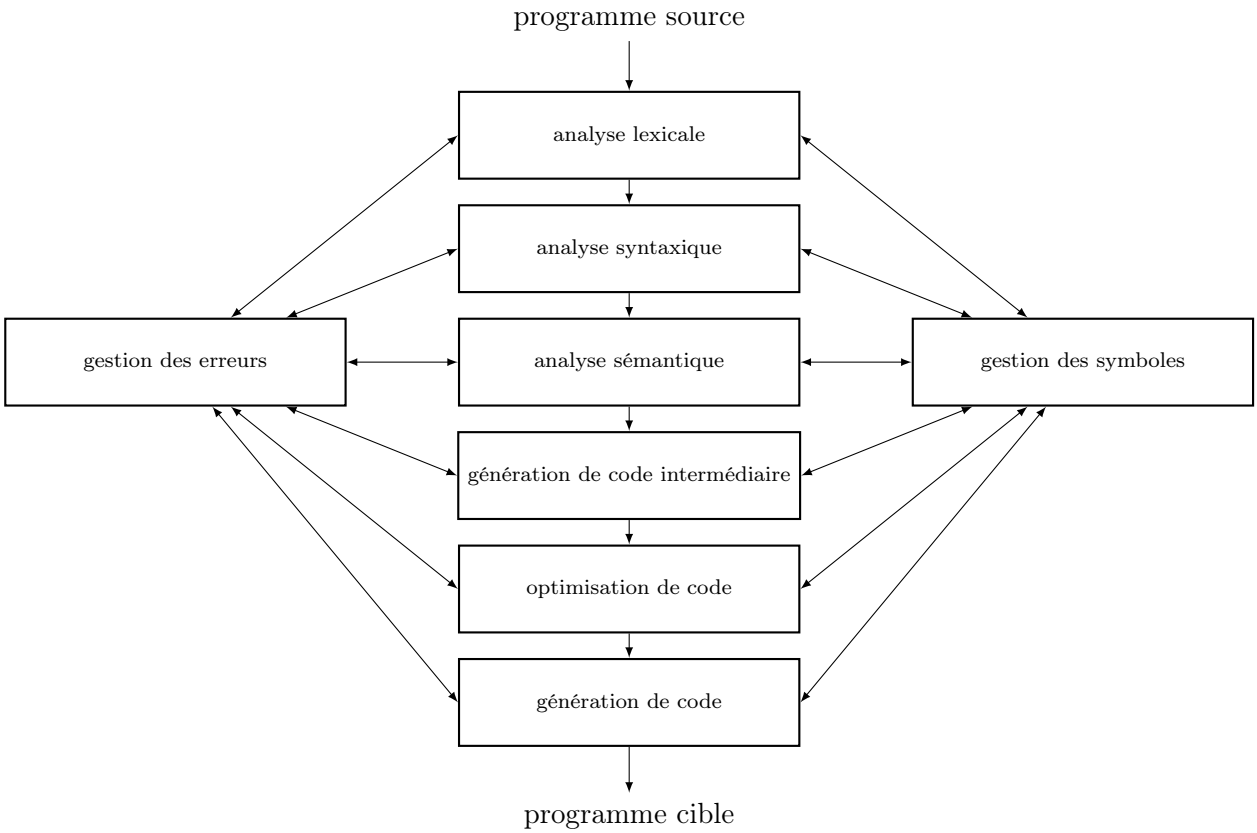


FIGURE 3.1 – Phases courantes d'un compilateur

Toutes les phases sont susceptibles de repérer des erreurs qui doivent être rapportées à l'utilisateur directement ou par un fichier de log. De plus, le compilateur doit maintenir une *liste des*

symboles qui enregistre les identifiants (p.ex. : nom de procédure, étiquette, constante, etc.) et complète l'information au fur et à mesure de l'avancement dans la compilation. Par exemple, dans le code suivant, l'adresse de l'étiquette `etiq` ne peut être connue qu'à la prochaine ligne. Le compilateur lit donc l'instruction `J etiq`, stocke `etiq` comme un nouveau symbole et complète les informations sur ce symbole au fil du traitement.

```
        J etiq
etiq:   .byte -4
        ADD $2, $3, $4
```

En résumé et pour citer J. S. Rohl [1], l'écriture d'un compilateur c'est définir des structures de données qui doivent être maintenues par le compilateur et les procédures par lesquelles ces structures sont créées et transformées.

Remarque : la compilation n'est pas concernée par *l'exécution* du programme (sauf dans quelques cas précis d'optimisation). Si le code source contient une addition, le compilateur se contente de traduire cette instruction en langage cible et certainement PAS de réaliser cette addition. Autrement dit, le compilateur doit pouvoir compiler n'importe quel code source qui respecte le langage même s'il est complément farfelu ou erroné à l'exécution (p.e., division par zéro).

La première phase du compilateur est l'*analyse lexicale*, c'est-à-dire extraire les mots du code source et vérifier qu'ils appartiennent bien au langage.

3.2 Analyse lexicale

L'analyse lexicographique peut se définir comme l'analyse des mots (ou «lexèmes») contenus dans un langage. Dans notre cas, il s'agit de reconnaître les lexèmes contenus dans le programme source en langage assembleur. Les lexèmes lus peuvent être de nature différente : des mnémoniques, des noms de registres, des nombres, des étiquettes, etc. À titre d'exemple, on montre le découpage du petit programme assembleur suivant en lexème :

```
# Un commentaire...
etiq:  .byte - 4
    ADD $2,$3,$4
    J 0xABCD
```

Le résultat de l'analyse peut se présenter sous la forme suivante (`NL` signifie Nouvelle Ligne codée par le caractère '$\backslash n$' en langage C) :

```
[COMMENT      ]   Un commentaire...
[NL           ]   \n
[SYMBOLE      ]   etiq
[DEUX_PTS     ]   :
[DIRECTIVE    ]   .byte
[VAL_DECIMAL  ]   -4
[NL           ]   \n
[SYMBOLE      ]   ADD
[REGISTRE     ]   $2
[VIRGULE      ]   ,
[REGISTRE     ]   $3
[VIRGULE      ]   ,
[REGISTRE     ]   $4
[NL           ]   \n
[SYMBOLE      ]   J
[VAL_HEXA     ]   0xABCD
```

Cette opération se réalise typiquement en mettant en œuvre un automate à états finis comme celui de la figure 3.2. À chaque étape (décodage d'un lexème), on part de l'état initial `Init` et on est amené vers les différents états de l'automate suivant le prochain caractère rencontré. On boucle alors sur l'état à chaque nouveau caractère jusqu'à ce que ce caractère ne soit pas permis par l'état. On est alors amené vers l'état terminal `Term` où l'on identifie et stocke le lexème avant de retourner à l'état `Init`.

Le schéma présenté est donné à titre d'exemple, il doit être adapté à vos besoins !

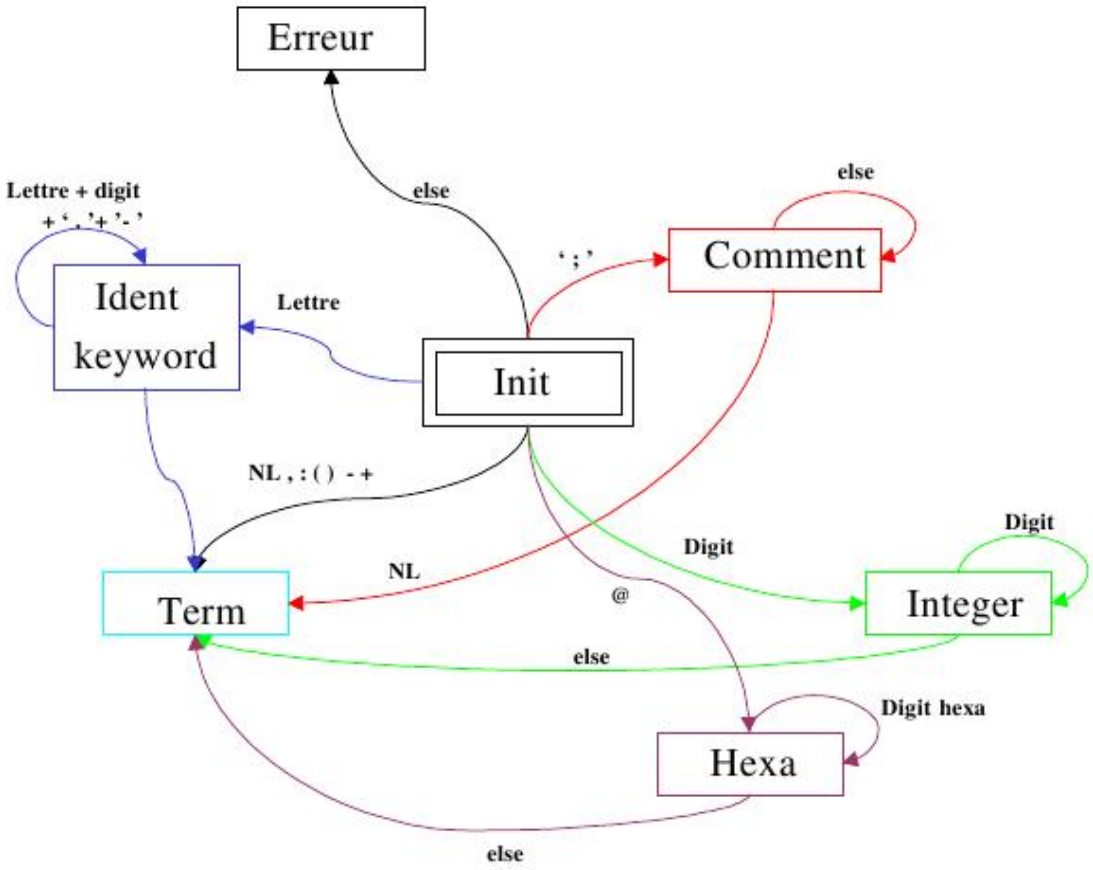


FIGURE 3.2 – Exemple d'automate à états finis

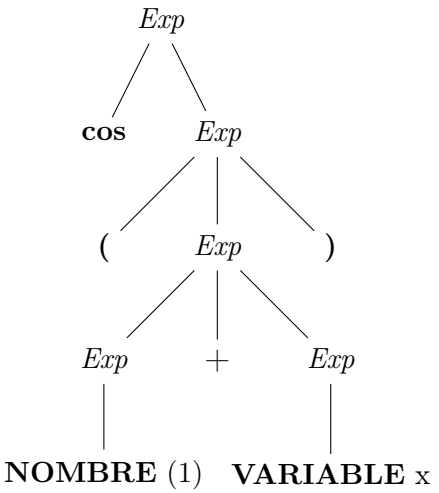
3.3 Analyse grammaticale

L'analyse grammaticale vérifie que les lexèmes extraits lors de l'analyse lexicale forment une séquence valide par rapport au langage assembleur MIPS, autrement dit, que le code respecte la syntaxe MIPS.

Le rôle de l'analyseur syntaxique est de vérifier la conformité syntaxique et de produire l'arbre d'analyse qui sera par la suite exploitable informatiquement. En général, on explicite la syntaxe d'un langage de programmation par une grammaire (d'où le nom d'analyse grammaticale). Une grammaire est composée de

$Exp \rightarrow$ NOMBRE
$Exp \rightarrow$ VARIABLE
$Exp \rightarrow$ cos Exp
$Exp \rightarrow Exp + Exp$
$Exp \rightarrow$ (Exp)
...

(a) Grammaire réduite des expressions arithmétiques.



(b) Arbre d'analyse de *cos(1+x)*.

FIGURE 3.3 – Exemple de grammaire et d'arbre d'analyse de *cos(1+x)*.

— symboles terminaux (les lexèmes) dans un alphabet A,
— de variables (symboles intermédiaires) qui appartiennent à l'alphabet V,
— règles de grammaire $v \rightarrow w$, où $v \in V$ et $w \in A \cup V$

La figure 3.3.a donne un exemple de grammaire pour l'analyse d'expressions arithmétiques. Les termes en italique sont les variables tandis que les autres sont les symboles terminaux.

L'analyse proprement dite consiste à vérifier si chaque lexème d'entrée peut être associé à un symbole terminal en respectant le chemin de production induit par la grammaire. Cette analyse est soit faite de manière descendante (c'est-à-dire trouver le chemin en partant de la racine qui peut expliquer la séquence de lexèmes d'entrées) ou ascendante (trouver le chemin qui mène à la racine à partir de la séquence de lexèmes). Le chemin ainsi trouvé permet de constituer l'arbre d'analyse. La figure 3.3.b montre l'arbre d'analyse de l'expression $cos(1+x)$ ou chaque symbole terminal est affiché en gras. Bien entendu la performance des analyseurs dépend de la complexité des grammaires qui peuvent comprendre un certain nombre d'ambiguïtés, de récursions ou de dépendances contextuelles (p.ex : le symbole '*' en C). L'analyse grammaticale est un ancien et vaste domaine de l'informatique que nous n'aurons pas le loisir d'aborder dans ce projet. Sachez seulement que ce domaine a fourni des outils d'analyse syntaxique tels que Yacc [1] ou Cup [2] pour faciliter l'analyse de la plupart des langages informatiques.

Cependant, la grammaire correspondant au langage MIPS est tellement simple qu'elle nous permettra de nous passer d'outils et de créer notre propre analyseur linéaire spécifique pour le langage MIPS. Cette grammaire peut être recomposée à partir de la syntaxe des différentes instructions, avec leurs opérandes et modes d'adressages décrits dans l'annexe C.

1. http://dinosaur.compilertools.net/
2. http://www2.cs.tum.edu/projects/cup/

3.4 Analyse sémantique

L'analyse sémantique se concentre sur la vérification du sens du programme. Sa tâche concerne notamment la vérification de type de variables dans des expressions (p. ex. : opérateur appliqué à un opérande incompatible), le contrôle d'unicité (p. ex. : un terme déclaré plusieurs fois) ou encore la vérification de la cohérence du programme (p.ex., utilisation d'une variable non initialisée). Le langage assembleur étudié ici étant très simple cette analyse prendra une place relativement peu importante (p.ex. : pas de contrôle de flot d'exécution par exemple). Par conséquent, cette analyse ne nécessitera pas forcément un module indépendant mais pourra être effectuée dans les autres modules (analyse grammaticale et génération de code). Par exemple, la détection d'utilisation multiple d'un nom pour une étiquette pourra être mise en œuvre au sein de la lecture des symboles.

3.5 Code intermédiaire et optimisation/modification de code

L'optimisation de code consiste à modifier le programme source pour le rendre plus efficace ou moins consommateur en ressource. Les compilateurs, tels que GCC, sont capables de modifier un programme pour les rendre plus compact (gain de mémoire) et plus efficace en termes de rapidité d'exécution et d'accès mémoire. Il existe une trop grande variété de types d'optimisation pour les décrire tous en détail. Un exemple d'optimisation classique est le remplacement de toutes les opérations de division ou multiplication par une puissance de deux par des décalages binaires.

Un autre exemple d'optimisation consiste à réécrire une suite non optimale ou inutile d'opérations en une suite d'opérations plus efficace. Dans l'exemple suivant, la deuxième instruction est inutile car $t1 et $t2 sont déjà égaux. Notez cependant que si une étiquette était présente en face de la deuxième instruction celle-ci ne devrait pas être supprimée car elle pourrait être utilisée dans une boucle. De même, la troisième et la quatrième instructions n'ont aucun effet (pas de modification de valeur et écriture dans `$zero`) et peuvent être éliminées.

Exemple :

```
MOVE $t1, $t2
MOVE $t2, $t1
ADD $t1,$t1,$0
ADDI $zero,$t1, 120
```

Parmi la multitude d'optimisations possibles, une technique simple d'amélioration du code est l'optimisation à lucarne (*peephole optimization*). Celle-ci consiste à améliorer une courte séquence d'instructions (appelée lucarne) en une séquence plus efficace. Une des caractéristiques de l'optimisation à lucarne est que chaque amélioration est susceptible d'engendrer d'autres optimisations. Cela implique que plusieurs passes sont nécessaires pour atteindre un code optimal. L'optimisation à lucarne est particulièrement adaptée à l'élimination de code mort (instructions ne pouvant jamais être exécutées), aux simplifications algébriques (une opération logique est souvent plus rapide qu'une multiplication) et au remplacement de tous motifs de code prédéfini (p. ex. : les deux `MOVE` de l'exemple ci-dessus). Nous verrons que cette dernière capacité pourra être mise à profit dans la gestion des pseudo-instructions.

La génération de code intermédiaire est présente dans la plupart des compilateurs. Cette étape permet une grande ré-utilisabilité du compilateur pour la génération de code sur différentes cibles car, dans ce cas, seul l'étape de génération de code final est à réécrire. C'est aussi pour cette raison que la plupart des optimisations sont effectuées sur le code intermédiaire qui sert de pivot entre le langage source et le langage cible. Dans le cadre de l'assembleur, l'étape de génération de code intermédiaire est souvent absente étant donné sa grande simplicité. En effet, certains compilateurs,

tel que le GCC de GNU, utilisent même un langage assembleur comme code intermédiaire. Par conséquent, nous n'utiliserons pas de code intermédiaire et travaillerons directement sur le code assembleur d'entrée pour générer le code de sortie.

3.6 Génération du code binaire relogeable

Durant les phases d'analyse lexicale et grammaticale, l'information doit être stockée dans une structure afin de l'exploiter ultérieurement pour la génération du code binaire. En effet, à partir de cette structure, il devient possible de coder les instructions les unes après les autres et de déterminer les adresses correspondantes. Par exemple, dans le cas d'instructions dont les opérandes sont des registres, des valeurs immédiates ou des adresses absolues, les instructions peuvent être codées directement, indépendamment du reste du programme.

Par contre, lorsqu'un opérande est une étiquette, i.e. l'adresse d'une autre instruction ou d'une donnée, les choses se compliquent un peu. En fait, il serait nécessaire de calculer l'adresse effective de l'opérande, puis de l'intégrer dans le codage binaire des instructions du programme. Or ceci n'est pas possible car on ne connaît pas, **au moment de l'assemblage**, les adresses mémoires auxquelles seront implantées les différentes sections `.text`, `.data` et `.bss`. En effet, celles-ci ne seront connues qu'au moment du chargement du programme où le *chargeur* devra *loger* le programme en mémoire. Les assembleurs génèrent donc des instructions *incomplètes* avec les informations permettant au chargeur de les compléter. Le code devient ainsi *relogeable*.

3.6.1 Nécessité d'un code relogeable

Le rôle de l'assembleur est, à partir du fichier source en langage assembleur, de générer le code binaire de l'ensemble des instructions et des données composant le programme. C'est ce code binaire qui sera chargé en mémoire puis exécuté lorsque le programme sera lancé.

Certaines instructions peuvent être codées directement, indépendamment du reste du programme et de l'adresse à laquelle elle seront chargées en mémoire. C'est notamment le cas des opérandes des instructions ne mettant en jeu que des registres ou des valeurs immédiates. Par contre, quand les opérandes des instructions sont des étiquettes, comme dans le cas d'un saut ou d'un accès à la zone `.data`, le code binaire de ces instructions ne peut pas être produit à l'assemblage. En effet, les adresses définitives référencées par les étiquettes ainsi que les adresses des sections ne seront connues que lors du chargement du programme en mémoire.

Les exemples suivants illustrent les problèmes rencontrés dans ce cas.

Exemple 1:

```
.text
ADDI $3,$0,12345   # met 12345 dans le registre 3
SW $3, X           # écrit le contenu de $3 à l'adresse X

.data
X: .word 0         # réservation d'un mot initialisé à 0 dans data
```

Dans cet exemple, l'instruction ADDI ne pose aucun problème. Par contre le SW [a] dépend de l'adresse à laquelle correspond l'étiquette `"X"`. Or cette adresse ne peut être connue avant que l'adresse d'implantation des sections `.text` et `.data` ne soient elles mêmes connues. Or celles-ci ne seront connues qu'au chargement du programme. Mais le programme ne pourra jamais être chargé s'il n'est pas assemblé... Le problème est donc : Comment coder l'instruction SW de l'exemple si l'offset ne peut pas être calculé ?

a. Il s'agit ici de la *pseudo-instruction* SW dont le deuxième opérande est une adresse (32 bits) et non de l'instruction qui attend un offset sur 16 bits. Elle sera remplacée par l'assembleur par les instructions LUI et SW.

Exemple 2:

```
.data
X: .byte 0xAB   # réservation d'un octet initialisé à 0xAB
Y: .word Z      # réservation d'un mot initialisé à l'adresse
                # référencée par l'étiquette Z
Z: .word 0      # réservation d'un mot initialisé à 0xAB
```

Dans cet exemple, la valeur qui suit l'étiquette `Y` est en fait l'adresse référencée par l'étiquette `Z`. Comme dans le cas précédent, il est nécessaire de connaître l'adresse d'implantation de la section `.data` pour déterminer les valeurs de `X, Y, Z`.

Exemple 3:

```
JAL fonction  # Appel de la procédure "fonction"
NOP           # On ne fait rien pour cause de \emph{Delay Slot} (cf section \ref{sec:delay_s
B end         # Retour ici après la procédure, on se branche sur la fin du programme
NOP           # delay slot

fonction :    # début de la procédure
ADD $t1,$0,$0
JR $31        # fin de procédure fonction, retour à l'appelant

end:
```

Ici, deux étiquettes sont utilisées par des instructions de Branchement et de saut. Ces deux cas ne sont cependant pas équivalents :

- Le codage du branchement nécessite de calculer le décalage entre l'instruction `B` et l'étiquette `end`. Mais puisque cette étiquette se situe dans la même section que le branchement lui-même, ce décalage peut être calculé lors de l'assemblage. Mais ce n'est pas toujours le cas, car cette étiquette aurait très bien pu être déclarée dans un autre fichier assembleur. Ce décalage n'aurait été dans ce cas calculable qu'au moment de l'édition de liens.
- Le problème est différent pour l'instruction `JAL`. En effet, l'adresse de destination du saut est calculée à partir de la valeur de l'opérande (ici l'étiquette) **ET** de la valeur de `PC`, soit l'adresse de l'instruction `JAL` + 4 [a].
  Donc, même si la position de l'étiquette `fonction` est ici connue *par rapport* à l'instruction `JAL`, il manque la position effective en mémoire de cette instruction pour pouvoir déterminer son codage complet.

a. Souvenez-vous que lors de l'exécution d'une instruction, le compteur PC a déjà été incrémenté de 4 octets.

En résumé, **le code d'un programme n'est déterminé de manière absolue qu'à partir du moment où les adresses mémoires auxquelles seront implantées les différentes sections (.text, .data ...) sont connues.**

3.6.2 Codage en position, *flat binary file*

Une solution est d'assembler le programme en définissant à l'avance, l'adresse mémoire de la section `.text`. Par exemple si `.text` est implanté en 0x5000, la section `.data` sera ensuite placée à la suite et ainsi de suite pour toutes les autres sections. De cette manière, il devient possible de coder l'ensemble des cas des exemples 1, 2 et 3.

Le programme peut ainsi être chargé directement en mémoire **UNIQUEMENT** *à l'adresse prévue* et exécuté en l'état. Par contre le programme n'est plus valide dès que les adresses utilisées ne sont plus disponibles. C'est notamment le cas si la machine ne dispose pas de suffisamment d'espace memoire ou si les plages d'adresses sont déjà utilisées par d'autres programmes.

3.6.3 Code relogeable

Une autre solution consiste à créer des fichiers objets dits *relogeables* qui contiennent des informations permettant de **modifier le code binaire au moment du chargement du programme**,

juste avant l'exécution, pour l'adapter dynamiquement aux plages d'adresse qui lui auront été allouées. On appelle cette opération la *relocation*.

Le principe des codes relogeables repose sur le fait que si les adresses effectives des instructions et étiquettes ne sont pas connues au moment de l'assemblage, leur position *relative* au début de la section où elles se trouvent l'est ! Ainsi dans les exemples ci-dessous :

— l'étiquette `X` de l'exemple 1 correspond en fait à l'adresse de la section `.data`
— l'étiquette `Y` (resp. `Z`) de l'exemple 2 correspond à l'adresse de la section `.data` + 4 octets (resp. 8 octets).
— l'étiquette `fonction` (resp. `end`) de l'exemple 3 correspond à l'adresse de la section `.text` + 16 octets (resp. 24 octets) et l'instruction `JAL` est implantée en début de section.

Lors de l'assemblage, il suffit donc de générer une version *'relative'* du code binaire et d'inclure dans le fichier objet toutes les informations qui permettent d'adapter ce code lors de son positionnement en mémoire. Un fichier relogeable contient donc au moins :

— le code des sections assemblées avec des informations liées au positions relatives des instructions et données par rapport au début de chaque section.
— une *table de relocation* indiquant, pour chaque section, les positions des adresses relatives à mettre à jour au moment du chargement et le mode de relocation (définissant les calculs à effectuer pour déterminer les bonnes adresses effectives).
— une *table des symboles* contenant les labels des étiquettes (chaînes de caractères) associées à leur adresse relative (pour les symboles définis localement).

Le chargement, avant l'exécution du programme est alors précédé d'une phase de relocation :

1. le fichier objet est lu pour déterminer le nombre d'octets nécessaires (taille des sections) pour copier les instructions et données dans la mémoire.
2. l'éditeur de liens (ou sur certains systèmes le chargeur dynamique) détermine à quelles adresses vont être placées les différentes sections composant le programme
3. Le relocation proprement dite a alors lieu. À partir des adresses d'implantation déterminées par le système et des informations de la table de relocation, elle convertit les *adresses relatives* des étiquettes en *adresses absolues*. Selon les modes de relocation à utiliser, cette conversion est plus ou moins simple.
4. le code ainsi modifié est prêt a être exécuté.

Le principal intérêt des codes relogeables est d'être portable d'une machine à une autre puisqu'ils ne sont pas spécifiques à une configuration mémoire. Ils sont également moins encombrants, en particulier dans le cas de programme multi-fichiers ou utilisant des librairies car les procédures ne sont codées qu'une seule fois.

Compilation séparée

Les fichiers relogeables permettent la compilation séparée, c'est-à-dire la création d'un fichier objet qui fait partie d'un programme plus vaste utilisant plusieurs fichiers objets qui seront rassemblés par l'édition de liens.

Au moment de l'assemblage, il est possible que certains symboles soient indéfinis (par exemple, procédures définies dans un autre fichier) et c'est l'éditeur de liens qui ira chercher leur définition. Pour chaque symbole indéfini, l'assembleur va donc laisser un "trou" à l'endroit où ce symbole est référencé et noter dans le fichier binaire à quel symbole correspond ce trou, ainsi que l'action à exécuter pour l'adresse finale au moment où il aura connaissance de l'adresse de ce symbole.

3.7 Informations nécessaires à la relocation

3.7.1 Table de relocation

Les informations nécessaires à la relocation sont définies dans les tables de relocation qui seront incluses par l'assembleur dans le fichier objet. Une table est associée à chaque section contenant des symboles à reloger. La table associée à la section `.text` (resp. `.data`) est appelée `.rel.text` (resp. `.rel.data`). Chaque ligne d'une table de relocation est définie par les lignes suivantes :

- `offset` : la position de l'entrée à modifier, en nombre d'octets par rapport au début de la section à laquelle la table est associée.
- `type` : le mode de relocation
- `value` : le symbole par rapport auquel il faudra faire la relocation. Dans le cas de symboles locaux, `value` est *l'index* de la zone contenant le symbole et dans le cas de symboles globaux (éventuellement non définis à l'assemblage) c'est l'index du symbole en question.

Dans l'exemple 2, l'entrée à reloger est un mot à l'adresse `Y` (soit 4 octets après le début de la section `.data`) et le symbole à reloger est `Z` qui se trouve aussi dans la section `.data`. La table associée est donc :

```
[.rel.data]
Offset        Type          Value
00000004      R_MIPS_32     .data
```

Dans l'exemple 3, l'entrée à reloger est l'instruction `JAL`, la première de la section `.text` et le symbole recherché est l'étiquette `fonction` qui se trouve également dans la section `.text`. La table associée est donc :

```
[.rel.text]
Offset        Type          Value
00000000      R_MIPS_26     .text
```

3.7.2 Champ addend

Une dernière donnée pour la relocation d'une entrée est le `addend`, c'est-à-dire la valeur binaire présente dans l'espace qui va être modifié par le relocation. Cette valeur, contenue dans le fichier objet, va être récupérée par l'éditeur de lien ou le chargeur dynamique et utilisée avec les informations de la table de relocation pour déterminer le nouveau code de l'instruction ou de la donnée. Elle sera finalement écrasée par le nouveau code au moment du chargement en mémoire.

Selon le mode de relocation, le addend correspond aux 32/26/16 bits de poids faible de l'entrée (toujours sur 32 bits) à reloger.

Dans le cas de symboles définis mais dont on ne connaît pas l'adresse absolue, `addend` est en générale l'adresse relative du symbole par rapport au début de section. Dans le cas des symboles indéfinis (référence à un symbole défini dans un autre fichier), l'assembleur, n'ayant pas d'adresse relative laisse généralement le addend à zéro.

3.7.3 Modes de relocation du MIPS

Le *mode de relocation* détermine le calcul spécifique à effectuer pour reloger une entrée (instruction ou donnée). Beaucoup de modes existent mais nous nous restreindrons à ceux définis ci-dessous :

R_MIPS_32 Ce mode est utilisé pour reloger une donnée en section `.data`. Les 32 bits de l'entrée sont modifiés

R_MIPS_26 Ce mode est utilisé pour reloger une instruction de saut de type `J`. Les 6 bits de poids fort (*opcode*) ne seront jamais modifiés par la relocation.

R_MIPS_HI16 &**R_MIPS_LO16** Ces deux modes fournissent les informations pour la relocation d'instructions successives de type I. Dans notre cadre, ils seront utilisés pour les instructions d'accès mémoire (`SW, LB...`). Seuls les 16 bits de poids faibles seront modifiés par la relocation. Des entrées de ces types apparaissent toujours par paire dans une table de relocation : chaque relocation **R_MIPS_HI16** est immédiatement suivie d'une relocation de type **R_MIPS_LO16**[3].

Les notations utilisées pour la description des calculs de relocations sont les suivantes :

— **P**lace désigne l'adresse de l'entrée à reloger, c'est-à-dire l'adresse allouée par l'éditeur ou le chargeur à la section plus la valeur `offset` associée à l'entrée, c'est à dire l'adresse "finale" de l'élément à reloger.

— **A**ddend désigne la valeur à **a**jouter pour calculer la valeur du champ à reloger (c'est la valeur du champ avant relogement). Selon le mode de relocation il s'agit des 32/26/16 bits de poids faible de l'entrée (en fait l'extension signée de `addend`)

— **S**ymbole désigne l'adresse "finale" du symbole par rapport auquel on reloge l'instruction. Si le symbole est de type `STT_SECTION` alors **S** prend la valeur de l'adresse du segment auquel le symbole appartient. Sinon, l'adresse est calculée à l'aide du champs `st_value` qui est ajouté à l'adresse de début de section à laquelle il appartient.

— **AHL** est une valeur de 32 bits calculée à partir des `addend` des deux entrées de relocation **R_MIPS_HI16** et **R_MIPS_LO16** consécutives. Si `AHI` et `ALO` sont les `addend` d'une paire d'entrées `HI16` et `LO16` de 16 bits chacun alors `AHL = (AHI << 16) + (short)ALO`. Par exemple si `AHI=0xABCD` et `ALO=0x12` alors `AHL=0xABCD0012`.

Avec ceci les calculs de relocation sont :

Mode	Adresse du 1er bit à modifier	Nb de bits à modifier	Valeur à écrire
`R_MIPS_32`	P	32	S+A
`R_MIPS_26`	P+6 bits	26	((A<<2)|(P & 0xF0000000)+S)>>2
`R_MIPS_HI16`	P+16 bits	16	((AHL+S)- ((short)(AHL+S)))>>16
`R_MIPS_LO16`	P+16 bits	16	(short)(AHL+S)

Remarque : Dans le cadre de ce projet, **nous supposerons que l'adresse d'implantation des sections `.text`, `.data` et `.bss` sont toujours à 0**. Cela sera au *linker* puis au *loader* d'attribuer une adresse finale aux sections.

3.8 Format `elf`

Plusieurs formats de fichier relogeable existent pour différent systèmes d'exploitation. Dans ce projet nous utiliserons le format ELF qui est celui utilisés par les systèmes UNIX. Pour la description de ce format veuillez vous referer à l'annexe B.

3. En réalité une entrée **R_MIPS_HI16** peut être suivie de plus d'une entrées **R_MIPS_LO16**.

Chapitre 4

À propos de la mise en œuvre

4.1 Quelques conseils sur la programmation

4.1.1 Programmation incrémentale

La conduite d'un projet de programmation de taille «importante» dans un contexte de travail en équipe (qui n'offre pas que des avantages !) et multitâches (autres cours en parallèles) nécessite une méthodologie qui vous permettra d'éviter les erreurs (ou les résoudre facilement) et de gagner en efficacité. Par ailleurs, le développement de programmes nécessite de nos jours de plus en plus de réactivité aux modifications de toutes sortes (p.ex. : demandes des clients, changement d'équipe, bugs, évolutions technologiques) ce qui a conduit à des méthodes de conception s'écartant des schémas classiques de conception/implémentation pour adopter un processus plus souple, plus facile à modifier en cours de développement. On pourra utilement s'inspirer de la programmation incrémentale qui consiste principalement à :

- Séparer le projet en modules indépendants de petites tailles en fonction des fonctionnalités désirées du programme.
- Chercher des solutions simples pour les réaliser (ce qui ne veut pas dire les SEULES solutions que vous connaissez).
- Concevoir les tests des modules AVANT leur écriture (concevoir les tests avant permet de bien réfléchir sur le comportement attendu et de détecter les erreurs au plus tôt).
- Intégrer la génération de traces pour faciliter le débogage.
- Commenter le code pendant l'écriture du code (après, c'est trop tard).
- Bien définir les responsabilités de chaque membre de l'équipe (p.ex. : écriture des tests, des structures de données, des rapports, etc.).
- Discuter du projet avant chaque phase de travail.
- ! Se mettre d'accord sur les standards de programmation [14] ! (p.ex. : organisation des dossiers, include, makefile, éditeurs, commentaires, nom des variables, etc.)
- ...

À toute fin utile vous pouvez consulter le site de la communauté de l'*eXtreme Programming*[1] qui fourmille de conseils intéressants.

4.1.2 Se familiariser avec le domaine d'application

Réalisez des expériences pour mieux appréhender les différents aspects du projet et préparer des fichiers de tests que vous utiliserez pour corriger et valider votre programme.

- Écrivez des programmes en assembleur, qui vous permettront de mieux appréhender ce langage. Vous pouvez commencer par quelques instructions puis compliquer la chose en gérant des données, des branchements (instructions conditionnelles), un appel à une procédure, un tableau, ...
- Constituez-vous une base de programmes tests en langage assembleur couvrant les différentes instructions, les modes d'adressage, etc.
- Ces fichiers pourront servir de tests pendant le projet pour déboguer vos programmes.
- Déterminez manuellement, à partir des spécifications, le codage de quelques instructions.

1. http://www.extremeprogramming.org

— Vous pouvez ensuite jouez avec les différents outils fournis pour compiler et exécuter du code assembleur (cf. annexe A). Toujours instructif !

4.1.3 Conception et développement

Après avoir pris connaissance du langage assembleur MIPS, vous pouvez vous lancer dans la conception du programme : identification des différentes tâches, choix des structures de données intermédiaires, découpage modulaire du code (quels fichiers ? quelles fonctions ? quelles entrées-sorties ?), etc.

Par exemple, sous quelle forme vont être représentés les lexèmes (quelle structure de données) ? Quelles sont les fonctionnalités nécessaires pour réaliser l'analyse syntaxique ?

Il faut prévoir une décomposition du développement de manière à pouvoir tester et corriger le programme au fur et à mesure que vous l'écrivez. Sinon, vous risquez d'avoir un programme très difficile à corriger ou vous risquez de devoir réécrire de grandes portions de code. La programmation modulaire permet en outre d'avoir un code plus concis et donc plus facile à déboguer.

Programmez de manière **défensive**. Pour les cas que votre programme ne devrait jamais rencontrer générez un message compréhensible du type `Erreur Ouverture Fichier : fonction analyse lexicale` et arrêtez proprement le programme, afin de pouvoir déboguer plus facilement. Placez des traces d'exécutions dans vos programmes de manière à pouvoir suivre le déroulement du programme. Il est fortement recommandé de se familiariser avec l'utilisation d'un débogueur (`valgrind` et `gdb`).

Pensez à concevoir le programme de manière à pouvoir le modifier facilement. Par exemple, il peut être intéressant de prévoir une représentation des données sous forme de structure pour pouvoir ajouter facilement des champs, ou encore d'utiliser des fonctions pour isoler et réutiliser au maximum du code indépendant. De cette manière, une correction/amélioration sur une fonction générale sera bénéfique pour l'ensemble du code.

De manière générale, on code d'abord les cas les plus généraux et les plus simples, avant de coder les cas particuliers et compliqués. Gardez-comme ligne directrice d'avoir le plus tôt possible **un programme qui fonctionne**, même s'il ne gère pas tout. Ensuite, améliorez-le au fur et à mesure.

4.1.4 Développement dirigé par les tests

Quand un programme dépasse les quelques lignes, qu'il est conçu par plusieurs personnes et qu'il a un objectif d'utilisation générale (c.-à-d., d'être utilisé par des clients), plusieurs problèmes vont devoir être résolus. Comment parvenir à faire évoluer un code de plusieurs milliers de lignes sans effets de bord ? Comment corriger un bug sur le code de Dupond qui est parti en week-end prolongé à Ibiza ?

Mise à part une gestion rigoureuse, un bon moyen de s'assurer, dès la conception, du bon fonctionnement d'un bout de code est de prévoir des tests. Dans un programme informatique, chaque fonction prend en entrée un certain nombre de paramètres, effectue un calcul et renvoie une valeur et peut modifier les paramètres d'entrées. Durant la conception de cette fonction (avant le codage), le programmeur a en tête quelques scénarios d'utilisation. Le développement dirigé par les tests consiste à utiliser ces scénarios pour *tester* la validité de la fonction. Un *test*, consiste donc à fournir des valeurs en entrée à une fonction et à vérifier que le résultat est bien celui attendu. L'extrait de code ci-dessous fournir un exemple de test.

Dans ce test, écrit avant le codage de la fonction, le programmeur est sûr de ne pas avoir oublié de prendre en compte les cas particuliers (le zéro) et les cas d'erreurs (nombres négatifs). Sans ces

```
/* test factorielle */
#include "factoriel.h"  /* prototype de la fonction factorielle:
                           int fact(int)*/

void main(){
        /* test d'un cas isole*/
        if(fact(3)==6) printf("1-OK\n") else printf("1-KO\n")

        /* test du zero*/
        if(fact(0)==1) printf("2-OK\n") else printf("2-KO\n")

        /* test des nombres negatifs : renvoie -1 lorsque erreur*/
        if(fact(-2)==-1) printf("3-OK\n") else printf("3-KO\n")

}
```

FIGURE 4.1 – Extrait de code illustrant le test de la fonction factorielle

tests, il y aurait eu de fortes chances pour que ces cas aient été oubliés lors du codage [2] ce qui aurait pu avoir des conséquences sur le reste du programme (par exemple beaucoup de temps perdu pour retrouver l'origine d'un bug). Par ailleurs, si la fonction doit être réécrite — d'impératif en récursif par exemple — les tests n'ont pas besoin d'êtres modifiés, ils sont réutilisables à l'infini.

4.2 Quelques conseils sur le projet assembleur

4.2.1 Réflexions sur le format des données

La compilation se compose de plusieurs phases de traitement bien définies (cf. figure 3.1) qui vont produire de l'information nouvelle (p.ex., le code final) ou enrichir de l'information produite par les traitements précédents (p.ex., ajouter l'adresse à une instruction traitée par l'analyse syntaxique). Le choix de représentation des structures de données doit donc se tourner tout naturellement vers des solutions permettant de lier les informations à propos d'une même instruction. Pour cela, les structures sont un choix tout indiqué. Par ailleurs, les contraintes de temps liées au projet (livrables à rendre fréquemment) vont faire qu'il ne sera pas possible de prévoir d'avance toutes les subtilités du programme. Ainsi, il est fort possible que les structures de données soient amenées à être modifiées au cours du projet. Là aussi, les structures sont un choix pertinent. En effet, il est très facile d'ajouter un champ à une structure avec un minimum d'impact sur le code original. Par exemple, la fonction `void affiche_personne(char *nom, int age, int poids)` sera beaucoup plus dure à modifier que `void affiche_personne(Personne p)` (ou `Personne` est une structure ayant les champs `nom`, `age` et `poids`) si on décide après coup d'afficher aussi le numéro de sécurité sociale. Dans le premier cas, il faudra modifier tous les endroits ou la fonction est utilisée en ajoutant un argument alors que dans le deuxième cas il sera simplement nécessaire d'ajouter un champ dans la structure et de modifier le code dans la fonction `affiche_personne`. D'une manière générale, essayez de toujours concevoir un code facile à faire évoluer.

2. si, si, souvenez-vous de votre première année...

Les différentes phases de la compilation vont toutes manipuler le programme original mais à un niveau différent. L'analyse lexicale va manipuler des lexèmes, l'analyse syntaxique des motifs syntaxiques, la génération de code des instructions et la plupart de ces phases vont devoir gérer les symboles contenus dans le code source. Concernant l'analyse lexicale, le nombre de lexèmes n'étant pas connus à l'avance on pourra éviter l'utilisation d'un tableau en se référant au type abstrait vu en première année (liste, pile, file). L'analyse syntaxique est typiquement conçue comme manipulant les instructions sous forme d'arbres syntaxiques auxquels sont appliqués des transformations et ajoutés des informations au fur et à mesure des traitements (on parle d'arbres décorés). Cependant, la syntaxe de l'assembleur est suffisamment simple est contrainte pour utiliser une approche moins élaborée, plus proche de vos connaissances en programmation et algorithmique. En effet, les motifs des instructions sont de trois types (cf. 2.3.1) avec un nombre fixe d'opérandes. Une instruction pourra donc être représentée par une liste composée des chacun des éléments (étiquette, opérateur, opérandes...) voire une structure. Enfin, il conviendra de prendre en compte la table des symboles qui pourrait être codée par une table de hachage. Quel que soit le mode de représentation choisi, il sera important d'écrire les fonctions qui permettront l'accès, la modification, la création, l'affichage et la destruction de ces données[3].

4.2.2 Machine à états finis (*FSM*)

Une machine à états finis (ou FSM pour *finite state machine* en anglais) décrit les états dans lesquels un automate est autorisé à être. Une FSM décrit aussi quoi faire lorsque l'on atteint un état donné, ou même ce qu'il faut faire lorsque l'on passe de tel à tel autre état. Pour les différentes analyses, nous allons être amené à construire des automates. Par exemple, la première phase de l'assembleur est de vérifier que le code donné en entrée contient uniquement des éléments acceptés par le langage et de les identifier. Il faut notamment pouvoir reconnaître que la chaîne de caractères `0123` est une valeur octale et que `0x123` est une valeur hexadécimale, que `ADDI` est un symbole et que `.text` est une directive. Un moyen brutal serait de comparer les caractères du fichier avec toutes les chaînes possibles du langage (avec par exemple `strcmp`). Ceci est bien entendu impossible car : les possibilités sont trop importantes, les tailles des chaînes peuvent varier, il est nécessaire de bien identifier le début et la fin des chaînes d'intérêt pour éviter les recouvrements (p.ex. : trouver `.text` dans le commentaire `# la section .text`)...

Heureusement, le langage assembleur est complètement déterministe et il est possible de connaître la composition de chaque élément terminal du langage. Par exemple, on sait qu'une valeur hexadécimale est toujours préfixée par `0x` suivi d'un certain nombre de caractères $\in$ `[0,9]` $\cup$ `[a,b,c,d,e,f]` alors qu'une valeur octale n'est composée que de chiffres inférieurs à 8. En tournant les choses de cette façon le problème devient de trouver des *motifs* de caractères dans le texte et non plus des mots prédéfinis. Un autre problème vient du fait que les motifs peuvent partager des caractéristiques communes qui impliquent qu'il faut avoir lu un certain nombre de caractères avant de reconnaître un motif (p.ex. : tant que l'on n'a pas lu le 'x' après un zéro on ne sait pas si on lit un nombre hexadécimal ou octal).

Une façon d'aborder le problème est de représenter les motifs et leur parcours par un FSM. Succinctement, l'automate est composé d'états qui dans notre cas représentent la catégorie courante de la chaîne de caractères lue et de transitions entre états étiquetés par les caractères que l'on va lire. L'exemple de la figure 4.2 montre comment on peut représenter un automate faisant la différence entre nombres décimaux, octaux ou hexadécimaux.

Cet automate lit les caractères un par un jusqu'à arriver à l'état terminal ou erreur. Ainsi, en prenant l'exemple de la chaîne `0567`, l'automate passe successivement par `INIT`, `pref hexa`, et reste

3. Ainsi que les tests, bien entendu...

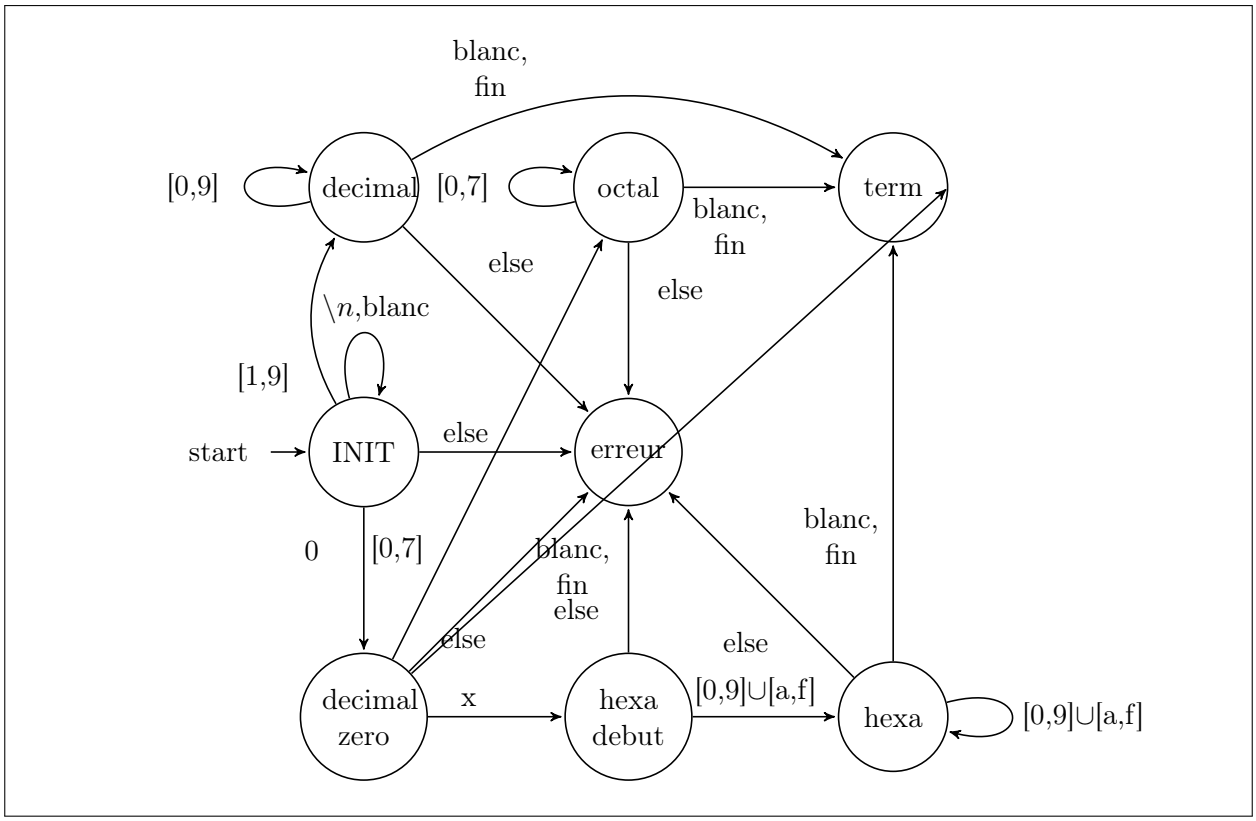


FIGURE 4.2 – Exemple d'automate faisant la différence entre une valeur décimale, octale et hexadécimale

dans `octal` jusqu'à la fin donnant ainsi la catégorie de la chaîne. La figure 4.3 donne une traduction en langage C de l'automate (il existe bien d'autre moyens de traduire l'automate en C).

Dans cette traduction, le fichier est parcouru caractère par caractère (boucle `while`). L'automate est tout d'abord dans l'état `INIT`. La lecture de chaque caractère peut provoquer une transition vers un autre état (par exemple si `c` est un chiffre), laisser l'automate dans l'état présent (p.ex., si `c` est un saut de ligne), faire terminer la tâche (p.ex., `EOF` *End Of File*) ou encore détecter une erreur. Ce programme est capable de traiter des fichiers textes volumineux en peu de temps.

4.2.3 Découper une chaîne de caractères en *token*, `strtok` pour la représentation des instructions MIPS

Votre assembleur doit vérifier un programme d'entrée et traduire chaque instruction en code machine. Cependant, où cette connaissance sera-t-elle stockée ? Comment la représenter ? Comment y accéder ?

La solution de coder *en dur* les instructions une par une dans le code est bien évidement à rejeter. D'une manière générale, on ne mélange pas la connaissance opérationnelle (c.-à-d., le code) et les données. La meilleure option pour le projet est de représenter les instructions dans un fichier texte séparé et de charger les instructions au début de l'exécution de l'assembleur. Le flot peut être ensuite parcouru pour chercher les informations avec la fonction standard `strtok`. L'extrait de code ci-dessous illustre son utilisation avec une chaîne de caractères.

La chaîne de caractères `"Dupond 20 76"` est séparée en éléments délimités par les espaces.

```c
/* definition des etats*/
enum {INIT, DECIMAL_ZERO, DEBUT_HEXA, HEXA, DECIMAL, OCTAL};
/* mise en oeuvre de l'automate*/
int main() {
    int c;              /*caractere analyse courant*/
    int S=INIT;         /*etat de l'automate*/
    FILE *pf;           /*pointeur du fichier à analyser*/

    if((pf=fopen("nombres.txt","rt"))==NULL) {
        perror("erreur d'ouverture fichier");return 1;}

    while(EOF!=(c=fgetc(pf))) {
        switch(S) {
        case INIT:
            i=0;
            if(isdigit(c)) {    /* si c'est un chiffre*/
                S = (c=='0')? DECIMAL_ZERO : DECIMAL;
            }
            else if (isspace(c)) S=INIT;
            else if (c==EOF) return 0; /* fin de fichier*/
            else return erreur_caractere(string,i,c);
            break;
        case DECIMAL_ZERO: /*reperage du prefixe de l'hexa*/
            if (c == 'x' || c == 'X') S=HEXA;
            else if (isdigit(c) && c<'8') S=OCTAL;/* c'est un octal*/
            else if (c==EOF || isspace(c)){ S=INIT;
                printf("la chaine est sous forme édcimale\n");
            }
            else return  erreur_caractere(string,i,c);
            break;
        case DEBUT_HEXA: /* il faut au moins un chiffre apres x*/
            if(isxdigit(c)) S=HEXA;
            else return erreur_caractere(string,i,c);
            break;
        case HEXA: /* tant que c'est un chiffre hexa*/
            if(isxdigit(c)) S=HEXA;
            else if(c==EOF || isspace(c)) { S=INIT;
                printf("la chaine est sous forme éhexadcimale\n");
            }
            else return erreur_caractere(string,i,c);
            break;
        case DECIMAL: /*tant que c'est un chiffre*/
            if(isdigit(c))  S=DECIMAL;
            else if(c==EOF || isspace(c)) {  S=INIT;
                printf("la chaine est sous forme édcimale\n");
            }
            else  return erreur_caractere(string,i,c);
            break;
        case OCTAL: /*tant que c'est un chiffre*/
            if(isdigit(c)&& c<'8') S=OCTAL;
            else if(c==EOF || isspace(c)) { S=INIT;
                printf("la chaine est sous forme octale\n");
            }
            else  return erreur_caractere(string,i,c);
            break;
        }
    }
    return 0;
}
```

FIGURE 4.3 – Exemple de traduction en C de l'automate de la figure 4.2

```
/* test strtok */
#include <string.h>/* prototype de la fonction strtok*/
#include <stdio.h>
typedef struct {char* nom; int age; int poids;} Personne;

void main(){
char *token;
char *texte = strdup("Dupond␣20␣\t␣76");
char *delimiteur = "␣";
Personne pers;

/*renvoie un pointeur vers "Dupond". */
printf("%s\n", pers.nom=strdup(strtok(texte, delimiteur)));

/* renvoie l'entier "20". */
printf("%d\n", pers.age=atoi(strtok(NULL, delimiteur)));

/* renvoie l'entier "76". */
printf("%d\n", pers.age=atoi(strtok(NULL, delimiteur)));
}
```

FIGURE 4.4 – Extrait de code illustrant l'usage de `strstok()`

Chaque appel à strtok() renvoie le prochain élément. Ainsi, on peut stocker des données sous forme de chaîne de caractères dans un fichier et lire (charger) ces informations dans des structures de données au démarrage du programme pour un accès rapide en mémoire (l'accès au fichier est lent). Des informations complètes sont accessibles dans le `man` de `strtok`.

Chapitre 5

Travail à réaliser

La page web du projet informatique est disponible à l'adresse suivante : http://tdinfo.phelma.grenoble-inp.fr/2Aproj/. Veuillez vous y référer pour tous ce qui concerne l'organisation et l'évaluation.

5.1 Objectif général :

À la fin de ce projet vous devrez avoir réalisé un assembleur qui prend en entrée un fichier source `file.s` et génère un fichier objet au format ELF : `file.o`

```
# allons au ru
.set noreorder
.text
    Lw $t0 , lunchtime
    LW $6, -200($7)
    ADDI $t1,$zero,8
boucle:
    BEQ $t0 , $t1 , byebye
    NOP
    addi $t1 , $t1 , 1
    J boucle
    NOP
byebye:
    JAL viteviteauru

.data
lunchtime:
    .word 12
    .word menu
    .asciiz "ils␣disent␣:␣\"au␣ru!\""
.bss
menu:
    .space 24
```

FIGURE 5.1 – Exemple de fichier assembleur `miam_sujet.s`

L'assembleur sera appelé en tapant sous Linux la commande :

```
as-mips source_filename
```

où `source_filename` est le nom du fichier texte contenant le programme à assembler.

5.1.1 Fichier objet binaire au format ELF

Le fichier binaire sera composé du codage binaire des sections `.text` puis `.data` puis `.bss`. Afin de rendre se code relogeable et éditable, le fichier binaire contiendra également la table des symboles ainsi que toutes les informations de relocation. L'écriture d'un fichier ELF nécessite de comprendre que ceux-ci sont composés d'un ensemble de sections (éventuellement vides) :

- Entête
- Table des entêtes de sections
- Table des noms de sections (".text", ".data", ".rel.text",...)
- Table des chaînes (noms des symboles)
- Table des symboles (informations sur les symboles)
- Section de données (.text, .data, .bss)
- Tables de relocations (.rel.text, .rel.data)

Une bibliothèque interagissant avec les librairies C standards (libelf, gelf) sera fournie pour permettent la lecture et l'écriture d'un fichier ELF. Pour comparer votre assembleur à un assembleur professionnel, vous pouvez utiliser l'utilitaire `as`. Par exemple la commande `mips-as -o miam_sujet.o miam_sujet.s` créera un fichier objet au format ELF. Un fois le fichier object créé vous pouvez observer les correspondances entre le code binaire et le programme en utilisant l'utilitaire `objdump`. Ainsi la commande `mips-objdump -d -j .text miam_sujet.o` désassemblera le fichier objet et affichera le contenu de la section text ce qui vous permettra de comprendre comment gnu effectue son assemblage.

```
miam_sujet.o:     format de fichier elf32-tradbigmips

Déassemblage de la section .text :

00000000 <boucle-0x10>:
   0: 3c010000  lui at,0x0
   4: 8c280000  lw t0,0(at)
   8: 8ce6ff38  lw a2,-200(a3)
   c: 20090008  addi t1,zero,8

00000010 <boucle>:
  10: 11090004  beq t0,t1,24 <byebye>
  14: 00000000  nop
  18: 21290001  addi t1,t1,1
  1c: 08000004  j 10 <boucle>
  20: 00000000  nop

00000024 <byebye>:
  24: 0c000000  jal 0 <boucle-0x10>
```

5.1.2 Vérification du fichier objet

Pour valider entièrement votre fichier binaire, vous pouvez regarder le contenu à l'aide d'un éditeur de fichier binaire, ou en utilisant l'utilitaire `objdump`. par exemple, la commande `mips-objdump -s -r -t -j .text -j .data -j .bss miam_sujet.o` affiche le contenu binaire d'un fichier. Ce contenu est représenté Figure 5.2.

Les premières lignes donnent la composition de la table des symboles encodée dans le fichier. On peut noter que même le nom des sections est inclus dans cette table. Chaque entrée est constitué de :

1. son décalage par rapport au début de la section (lunchtime est au début de la section data tandis que byebye est à 0x20 du début de la section text)
2. sa portée (ici l signifie local)
3. son type de symbole (ici d signifie debug)
4. la section à laquelle appartient le symbole
5. la taille du symbole
6. le nom du symbole

On trouve ensuite les informations de relocation pour chaque section. L'explication de chaque information peut être trouvée annexe B.

Puis le contenu binaire de chaque section est affiché à raison de 16 octets par ligne. On remarquera que la section bss n'est pas affichée ce qui est normal étant données qu'elle n'est crée qu'à l'exécution.

5.2 Étapes de développement du programme

Le projet peut être découpé en 4 étapes principales que vous aurez à achever dans un délai imparti. Au début de chaque étape, une séance de tutorat sera utilisée pour la préparation puis quelques séances de codages seront consacrées à la mise en œuvre.

5.2.1 [5 pts] Étape 1 : Analyse lexicale

But

À la fin de cette étape, vous devrez avoir validé l'analyse lexicale.

Marche à suivre

1. Définition des catégories de lexèmes et du format interne
2. Écriture des tests
3. Écriture de l'automate à états finis
4. Génération des traces
5. Gestion des erreurs

5.2.2 [5 pts] Étape 2 : Analyse syntaxique – 1

But

À la fin de cette étape, vous devrez avoir validé le chargement des instructions et le décodage des directives.

```
miam_sujet.o:     format de fichier elf32-tradbigmips

SYMBOL TABLE:
00000000 l    d  .text 00000000 .text
00000000 l    d  .data 00000000 .data
00000000 l    d  .bss 00000000 .bss
00000000 l       .data 00000000 lunchtime
00000010 l       .text 00000000 boucle
00000024 l       .text 00000000 byebye
00000000 l       .bss 00000000 menu


RELOCATION RECORDS FOR [.text]:
OFFSET   TYPE              VALUE
00000000 R_MIPS_HI16       .data
00000004 R_MIPS_LO16       .data
0000001c R_MIPS_26         .text
00000024 R_MIPS_26         viteviteauru


RELOCATION RECORDS FOR [.data]:
OFFSET   TYPE              VALUE
00000004 R_MIPS_32         .bss


RELOCATION RECORDS FOR [.bss]: (none)

Contenu de la section .text :
 0000 3c010000 8c280000 8ce6ff38 20090008  <....(.....8 ...
 0010 11090004 00000000 21290001 08000004  ........!).....
 0020 00000000 0c000000                    ........
Contenu de la section .data :
 0000 0000000c 00000000 696c7320 64697365  ........ils dise
 0010 6e74203a 20226175 20727521 2200      nt : "au ru!".
```

FIGURE 5.2 – La sortie de la commande `objdump` correspondant au code de la figure 5.1

Marche à suivre

1. Définition des catégories d'instructions, des types d'adressage et du format interne
2. Définition des structures de représentation des instructions
3. Écriture des tests
4. Chargement du dictionnaire d'instructions
5. Définition de la table des symboles
6. Décodage des instructions

7. Décodage des directives
8. Décodage des définitions de symbole
9. Vérification des instructions (nombre opérandes)
10. Génération des traces
11. Gestion des erreurs

5.2.3 [5 pts] Étape 3 : Analyse syntaxique – 2

But

À la fin de cette étape, vous devrez avoir validé la vérification de la syntaxe des instructions et la génération des entrées de relocation.

Marche à suivre

1. Écriture des tests
2. Décodage des opérandes et des modes d'adressage
3. Génération de la table des symboles
4. Vérification des instructions
5. Vérification des directives
6. Génération des entrées de relocation
7. Mécanisme de réécriture
8. Génération des traces
9. Gestion des erreurs

5.2.4 [5 pts] Étape 4 : Génération de code

But

À la fin de cette étape, vous devrez avoir validé la génération de code.

Marche à suivre

1. Écriture des tests
2. Calcul des adresses des instructions
3. Définition et génération de la table de relocation
4. Génération du fichier ELF
5. Tests de fonctionnement de l'assembleur

5.3 Bonus : extensions du programme

Plusieurs extensions possibles du programme peuvent être envisagées, telle que la prise en compte des *float*, des variables globales, des fichiers multiples, etc. Une extension intéressante peut être d'inclure certaines pseudo-instructions (cf. 2.3.2). Toute extension menée de manière satisfaisante amènera un bonus dans la notation.

5.4 Agenda et organisation du projet

Consulter le site web du projet info : http://tdinfo.phelma.grenoble-inp.fr/2Aproj/

Bibliographie

[1] J. S. Rohl *An introduction to compiler writing* London : Macdonald and Jane's ; New York : American Elsevier, 1975

[2] B. W. Kernighan et D. M. Ritchie *Le langage C, Norme ANSI* `http://http://cm.bell-labs.com/cm/cs/cbook/`

[3] Bradley Kjell. *Programmed Introduction to MIPS Assembly langage.* `http://chortle.ccsu.edu/AssemblyTutorial/TutorialContents.html`

[4] *MOPS32 Architectur For Programmers Volume II, Revision2.50.* 2001-2003,2005 MIPS Technologies Inc.
`http://www.mips.com/products/product-materials/processor/mips-architecture/`

[5] *Executable and Linkable Format (ELF).* Tools Interface Standard (TIS). Portable Formats Specification, Ver 1.1. `http://www.skyfree.org/linux/references/references.html`

[6] *64-bit ELF Object File Specification.*
`http://techpubs.sgi.com/library/manuals/4000/007-4658-001/pdf/007-4658-001.pdf`

[7] *System V Application Binary Interface - MIPS® RISC Processor.* `http://www.caldera.com/developers/devspecs`

[8] Amblard P., Fernandez J.C., Lagnier F., Maraninchi F., Sicard P. et Waille P. *Architectures Logicielles et Matérielles.* Dunod, collection Siences Sup., 2000. ISBN 2 10 004893 7.

[9] Dean Elsner, Jay Fenlason and friends. *Using as, the* GNU *Assembler.* Free Software Foundation, January 1994. `http://www.gnu.org/manual/gas-2.9.1/`

[10] David Alex Lamb. Construction of a peephole optimizer, *Software : Practice and Experience,* ***11(6)****, pages 639–647, 1981.*

[11] FSF. *GCC online documentation.* Free Software Foundation, January 1994. `http://gcc.gnu.org/onlinedocs/`

[12] Steve Chamberlain, Cygnus Support. *Using ld, the* GNU *Linker.* Free Software Foundation, January 1994. `http://www.gnu.org/manual/ld-2.9.1/`

[13] Linux Assembly `http://linuxassembly.org/`

[14] Linus Torvalds. *Linux Kernel Coding Style.*
`https://computing.llnl.gov/linux/slurm/coding_style.pdf`

[15] Bernard Cassagne. *Introduction au langage C.*
`http://www-clips.imag.fr/commun/bernard.cassagne/Introduction_ANSI_C.html`

Annexe A

Compilation d'un programme en assembleur MIPS

Ce chapitre décrit les principales commandes dont vous aurez besoin pour assembler un programme MIPS, générer les fichiers ELF qui vous serviront de tests et les analyser.

Nous allons utiliser l'assembleur `as` de GNU, qui peut être compilé pour traiter les programmes en assembleur MIPS et générer des fichiers binaires compatibles avec un processeur MIPS, et ce même si votre machine a effectivement un autre processeur.

A.1 Installation

Par défaut, `gcc` et d'autres outils de la librairie binutils (assembleur `as`, linker `ld`, désassembleur `objdump`,...) sont compilés pour la machine et le système d'exploitation que vous utilisez. À PHELMA, il s'agit essentiellement de PC Intel Pentium ou Athlon sous Linux.

Pour notre projet, il est nécessaire de compiler ces outils pour qu'ils soient adaptés au microprocesseur MIPS. On parle de "cross-compilation". Tout a été fait à l'école, bien sûr, et il vous suffira d'utiliser les commandes `mips-gcc`, `mips-as`, `mips-ld`, `mips-readelf`, etc. `gcc` (sans le préfixe `mips-`) sera bien sûr encore disponible pour que vous puissiez développer votre simulateur en langage C !

A.2 Compilation et étude des fichiers

Soit le programme `exemple.s` présenté figure A.1, écrit en langage assembleur MIPS [1].

A.2.1 Assemblage

Pour assembler ce fichier et créer un fichier objet binaire relogeable au format ELF, vous avez deux solutions :

1. Ou bien utiliser directement l'assembleur `mips-as`

   ```
   mips-as exemple.s -o exemple.o
   ```

2. Utiliser `mips-gcc` [2] avec l'option `-c` :

   ```
   mips-gcc -c exemple.s -o exemple.o
   ```

Vérifiez, celà revient exactement au même !

A.2.2 Désassemblage

Pour étudier le contenu du fichier `exemple.o` ainsi généré, vous pouvez utiliser différents outils :

1. Au passage, vous aurez un exemple de code assembleur avec appel d'une procédure...
2. `mips-gcc` est en fait un alias de la commande `mips64-gcc -mabi=32 -mgp32 -march=R3k`
Les options permettent de spécifier que l'on souhaite utiliser une architecture 32 bits, avec des *general purpose register* de 32 bits, pour un processeur de type MIPS R3000. Aspirine ?

```
.text
.globl __start
__start:
   li $a0, 1 # fib(n): parameter n
   move $v0, $a0 # n < 2 => fib(n) = n
   blt $a0, 2, done
   li $t0, 0 # second last Fibâ number
   li $v0, 1 # last Fibâ number
   fib: add $t1, $t0, $v0 # compute next Fibâ number in sequence
   move $t0, $v0 # update second last
   move $v0, $t1 # update last
   sub $a0, $a0, 1 # more work to do?
   bgt $a0, 1, fib # yes: iterate again
   done: sw $v0, result # no: store result, done
.data
result: .word 0x11111111
```

```
## Fichier exemple.s : un programme principal, avec appel d'une procédure Min
.text
.globl __start                # Pas obligé pour vous...

__start:

##-------------------------------------------------------------------------
# Programme principal
  li   $2, 10                 # $2 <- 10
  li   $3, 0xff               # $3 <- 0xff

# appel de la procedure min
  jal  Min                    # $ra (return adress) <- pc+1; puis saut à Min
  sw   $4, 0x1000($zero)      # ecrit le resultat a l'adresse 0x1000
  j    Exit                   # saut à la fin du programme


##-------------------------------------------------------------------------
# fonction Min qui calcule le minimum de $2 et $3 et le stocke dans $4.
# On pourrait aussi utiliser la pile...
Min:
  blt   $2, $3, Then          # si $2 >= $3 saut à Then
  move  $4, $3                # $4 <- $3 (else)
  j     End                   # saut à Fin
  Then: move  $4, $2          # $4 <- $2 (then)
  End:                        # fin de la procedure min
  jr    $ra                   # saut à l'adresse contenue dans $ra,
                              # (mise à jour auto lors de l'appel avec jal)
                              # on retourne ainsi dans le programme principal

Exit:
## The End
```


FIGURE A.1 – Fichier `exemple.s` en langage assembleur MIPS.

- `od -t xC exemple.o` permet de voir le contenu du fichier binaire. Différentes options d'affichage sont possible. Ici `-t xC` affiche les valeurs octet par octet en hexadécimal. `man od` pour plus d'informations.
- `mips-objdump exemple.o` désassembleur standard.
- `mips-readelf exemple.o` permet de voir les différentes sections, symboles, etc. d'un fichier ELF.

```
mips-gcc -S toto.c -o toto.s
```

A.2.3 Edition de lien

L'édition de lien (commande `ld`) permet normalement de créer, à partir d'un ou plusieurs fichiers objets `.o`, un unique fichier binaire exécutable sur une machine MIPS (ou avec un simulateur !). Par exemple :

```
mips-ld exemple.o -o exemple.bin
```

Dans le cadre de ce projet, on ne s'intéresse pas à l'édition de liens ! On se contentera de charger des fichiers ELF relogeables.

Cependant, la commande `ld` pourra tout de même être utilisée pour lier deux fichiers objets. La commande `ld` avec l'option `-r` permet en effet de produire à partir de plusieurs fichiers relogeables, un seul fichier relogeable. Ainsi la commande suivante lie ensemble `toto1.o` et `toto2.o` dans un seul fichier relogeable `toto.o` :

```
mips-ld -r toto1.o toto2.o -o toto.o
```

Cette forme d'édition de lien, appelée édition de lien statique, recopie le code des 2 fichiers en entrée dans le fichier en sortie. Vous pouvez faire un essai, avec par exemple le code d'une procédure dans un fichier et l'appel dans un autre. De jolis tests de relocation en perspective !

Annexe B

ELF : Executable and Linkable Format

Ce chapitre décrit "brièvement" le format ELF (*Executable and Linkable Format*) et explique comment en extraire les informations nécessaires pour le projet. Le format ELF est le format des fichiers objets dans la plupart des systèmes d'exploitation de type UNIX (GNU/Linux, Solaris, BSD, Androïd...). Il est conçu pour assurer une certaine portabilité entre différentes plates-formes. Il s'agit d'un format standard pouvant supporter l'évolution des architectures et des systèmes d'exploitation.

Le format ELF est manipulable, en lecture et écriture, par utilisation d'une bibliothèque C de fonctions d'accès `libelf`. Cette librairie suffit pour lire et écrire des fichiers ELF, même quand ELF n'est pas le format utilisé par le système d'exploitation (par exemple, on peut l'utiliser sous MacOs X).

Dans ce chapitre, nous nous limitons aux fichiers objets, dits *à lier*, fichiers contenant du *code binaire relogeable* (ou *translatable*), c'est-à-dire du code binaire dont l'affectation en mémoire n'est déterminée qu'au moment de l'exécution. Ce chapitre décrit tout d'abord la structure d'un fichier objet, puis s'intéresse à la notion de relocation.

B.1 Fichier objet au format ELF

Les fichiers ELF sont composés d'un ensemble de sections (éventuellement vides) comme indiqué par la figure B.1.

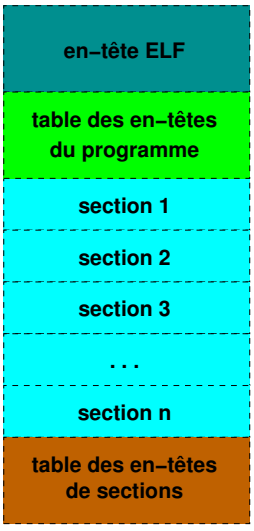


FIGURE B.1 – Structure d'un fichier ELF

:

— En-tête du fichier ELF
— Table des entêtes de sections
— Table des noms de sections (".text", ".data", ".rel.text",...)
— Table des chaînes (noms des symboles)
— Table des symboles (informations sur les symboles)

— Section de données (.text, .data, .bss)
— Tables de relocations (.rel.text, .rel.data)

Les seules sections qui contiennent des données sont les sections :

— `.text` qui contient l'ensemble des instructions exécutables du programme.
— `.data` qui contient les données initialisées du programme.
— `.bss` qui contient les données non initialisées du programme. Ces données ne prennent pas de place dans le fichier ELF : seules les tailles des zones mémoire à réserver sont spécifiées et ces zones sont remplies avec des zéros au début de l'exécution du programme.

Les autres sections servent à décrire le programme et le rendre portable et relogeable.

B.2 Structure générale d'un fichier objet au format ELF et principe de la relocation

Un fichier objet à lier au format ELF est formé d'un en-tête donnant des informations générales sur la version, la machine, etc., puis d'un certain nombre de pointeurs et de valeurs décrits ci-dessous. Ce que nous appelons ici pointeur est en fait une valeur représentant un déplacement en nombre d'octets par rapport au début du fichier. Les tailles, elles aussi, sont exprimées en nombre d'octets. La figure B.2 donne une idée de la structure d'un fichier au format ELF. Le fichier est constitué d'un *en-tête* donnant les caractéristiques générales du fichier, puis d'un certain nombre de *sections* contenant différentes formes de données, ces sections étant détaillées par la suite (par exemple, section "`.text`", section "`.data`", section des "relocations en zone text", section de la table des symboles, etc).

Lors de la fabrication d'un fichier objet, les instructions du programme sont logées dans une section binaire correspondant à une zone `.text` alors que certains des opérandes de ces instructions peuvent appartenir à une section binaire différente, par exemple la zone `.data`. Lors du chargement du fichier en mémoire en vue de son exécution (de sa simulation dans notre cas), ces différentes zones sont placées en mémoire par le chargeur. Ce n'est qu'après cette étape que les adresses des opérandes seront connues. Il faut donc mettre à jour la partie adresse effective des instructions afin que ces dernières accèdent correctement aux opérandes (notons qu'avant cette mise à jour, les adresses des opérandes dans les instructions sont des adresses relatives définies par rapport au début de la zone `.data` du fichier initial). Par ailleurs, cette situation peut être généralisée si ce fichier objet fait partie d'un programme plus vaste utilisant plusieurs fichiers objets susceptibles d'être rassemblés en un seul programme par l'éditeur de liens entre fichiers. Par exemple, un opérande en zone `.data` peut être utilisée par des instructions contenues dans des fichiers objets différents. Cette remarque est également valable pour les zones `.text`, par exemple l'appel d'une fonction déclarée dans un fichier externe. Il faut donc indiquer pour chaque section `.text` concernée, un moyen de retrouver l'adresse de cet opérande. Pour cette raison chaque section `.text` ou `.data` possède une section associée dite de relocation (`.rel.text` et `.rel.data`) contenant les informations nécessaires aux calculs des adresses. Comme explicité précédemment, toutes les adresses non définies avant le chargement, sont finalement mise à jour à l'issue du chargement. La partie du code des instructions correspondant à l'adresse des opérandes en question ne peut donc pas être figée au moment de la compilation du fichier objet initial mais seulement à l'issue du chargement du programme. Du fait de cette relocation, les fichiers ELF sont dits "relogeables".

Dans ce projet, pour des raisons de simplicité, nous n'aborderons pas le traitement de l'édition de liens entre plusieurs fichiers ELF. On se contentera de réaliser la simulation d'un fichier objet simple mais susceptible de nécessiter une relocation lors du chargement dans la mémoire du simulateur. Dans la suite de ce chapitre, nous donnons un exemple précis permettant d'illustrer en détails les

principes de la relocation.

Entête d'un fichier ELF L'en-tête est décrite par le type C `struct ELF32_Ehdr` dont voici la description des principaux champs :

- `e_ident` : identification du format et des données indépendantes de la machine permettant d'interpréter le contenu du fichier (Cf. exemple dans le paragraphe suivant).
- `e_type` : un fichier relogeable a le type `ET_REL` (constante égale à 1).
- `e_machine` : le processeur MIPS qui nous intéresse est identifié par la valeur `EM_MIPS` (constante égale à 8).
- `e_version` : la version courante est 1.
- `e_ehsize` : taille de l'en-tête en nombre d'octets.
- `e_shoff` : pointeur sur la *table des en-têtes de sections* (ou plus simplement "*table des sections*"). Chaque entrée de cette table est l'en-tête d'une section qui décrit la nature de cette section, et donne sa localisation dans le fichier (voir ci-dessous). La table des sections est appelée *shdr* dans la documentation ELF. Dans le reste du format ELF, les sections sont généralement désignées par l'index de leur en-tête dans cette table. Le premier index (qui est 0) est une sentinelle qui désigne la section "non définie".
- `e_shnum` : nombre d'entrées dans la table des sections.
- `e_shentsize` : taille d'une entrée de la table des sections.
- `e_shstrndx` : index de la table des noms de sections (dans la table des en-têtes de sections).
- Les autres champs de l'en-tête ne sont pas utilisés dans le cadre du projet. On les met à 0.

En-têtes des sections Un en-tête de section est décrit par le type C `struct Elf32_shdr`. Il définit les champs suivants :

- `sh_name` : index dans la table des noms de sections (section "`.shstrtab`").
- `sh_type` : type de la section. Les types qu'on utilise dans ce projet sont les suivants :
  - `SHT_PROGBITS` (constante 1) : type des sections "`.text`" et "`.data`". Ce type indique que la section contient une suite d'octets correspondant aux données ou aux instructions du programme.
  - `SHT_NOBITS` (constante 8) : type de la section "`.bss`" (données non initialisées). La section ne contient aucune donnée. Elle sert essentiellement à déclarer la taille occupée par les données non initialisées (voir ci-dessous).
  - `SHT_SYMBTAB` (constante 2) : type des tables des symboles. Dans le projet, on en utilise une seule, appelée "`.symtab`".
  - `SHT_STRTAB` (constante 3) : type des tables de chaînes. Dans le projet, deux sections ont ce type : la table des noms de sections, appelée "`.shstrtab`", et la table des noms de symboles, appelée "`.strtab`" (elles sont souvent désignées par "table des chaînes").
  - `SHT_REL` (constante 9) : type des tables de relocations. Dans le projet, on aura : "`.rel.text`" pour les relocations en zone `.text` et "`.rel.data`" pour les relocations en zone `.data`.
  - `SHT_REGINFO` : type spécifique aux fichiers ELF de processeurs MIPS 32 bits, pour la section "`.reginfo`" (Register Information Section).
- `sh_offset` : pointeur sur le début de la section dans le fichier.
- `sh_size` : taille qu'occupera la section une fois chargée en mémoire (en octets). Si le type de la section n'est pas `SHT_NOBITS`, la section doit correspondre effectivement `sh_size` octets dans le fichier, puisque la section sera chargée "telle quelle" en mémoire. Si le type de la section est `SHT_NOBITS`, ce champ sert à déclarer la taille de la zone non initialisée qui sera finalement allouée en mémoire.

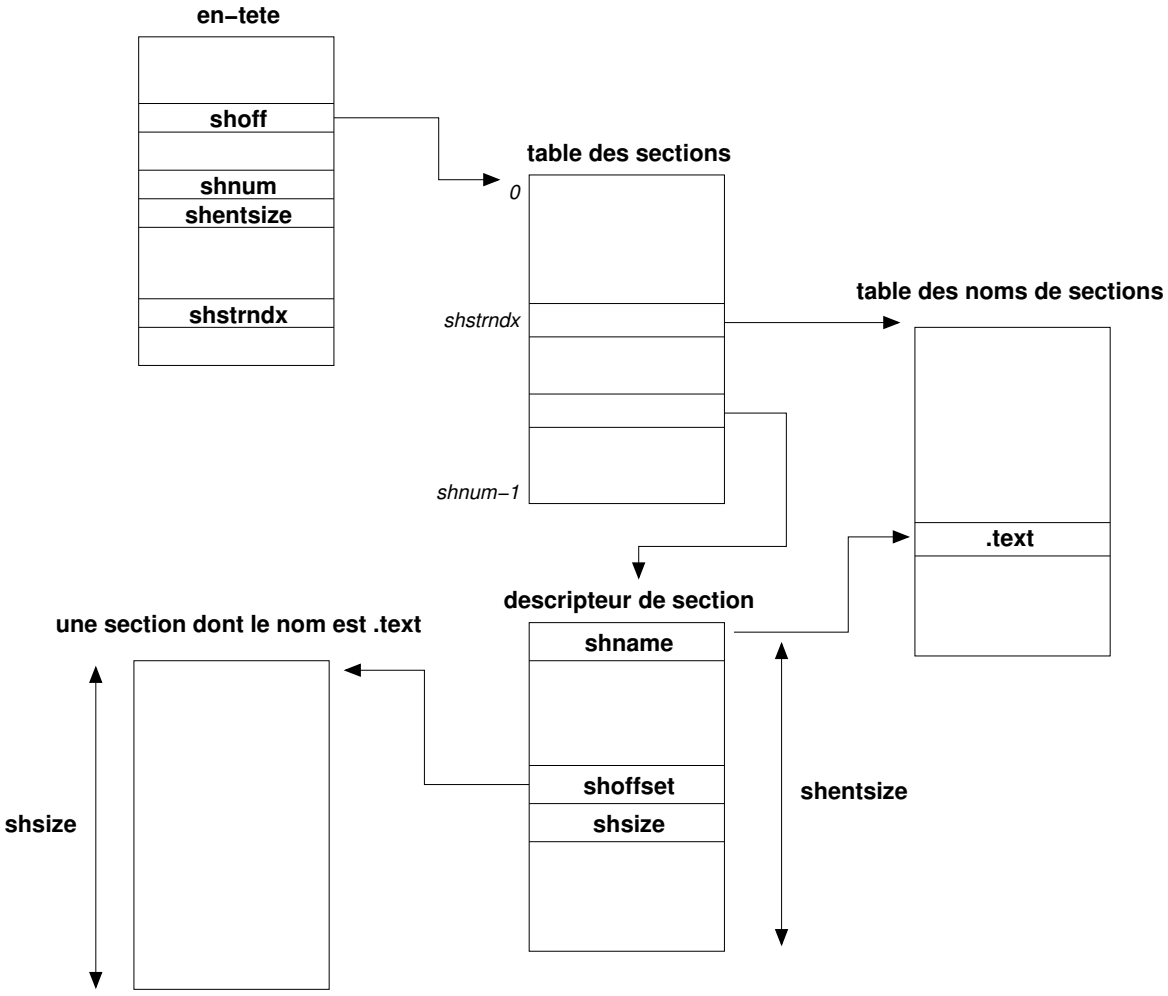


FIGURE B.2 – Structure d'un fichier relogeable au format ELF

- `sh_addralign` : contrainte d'alignement sur l'adresse finale de la zone en mémoire. L'adresse finale doit être un multiple de ce nombre.
- `sh_entsize` : certaines sections ont des entrées de taille fixe. Cet entier donne donc cette taille. Dans le projet, seules les sections de type `SHT_SYMBTAB` et `SHT_REL` sont concernées par ce champ.
- `sh_link` et `sh_info` : ont des interprétations qui dépendent du type de la section. Dans tous les cas, le champs `sh_link` est l'index d'un en-tête de section (dans la table des en-têtes de sections). Dans le cadre du projet, on a :

`sh_type`	`sh_link`	`sh_info`
`SHT_REL`	l'index de la table de symbole associée	l'index (dans la table des en-têtes de sections) de la section à reloger
`SHT_SYMTAB`	l'index de la table des noms de symboles	l'index (dans la table des symboles) du premier symbole global [1]

Remarque le contenu d'un fichier objet *exécutable* ressemble à celui d'un fichier objet relogeable. Il est formé de segments au lieu de sections et on y trouve ainsi une table des segments (au lieu d'une table des sections). Dans l'en-tête, les informations décrivant la table des segments sont données par les champs dont les noms commencent par `ph` au lieu de `sh`.

B.3 Exemple de fichier relogeable

Pour étudier le format ELF en détaillant le format de chacune des sections considérées dans le projet, considérons le programme en langage d'assemblage `reloc_miam.s` donné Figure B.3.

FIGURE B.3 – Programme assembleur nécessitant une relocation.

La commande `mips-as reloc_miam.s -o reloc_miam.o` produit un fichier objet `reloc_miam.o` dont nous donnons en figure B.4 le contenu affiché par la commande Unix `od -t xC reloc_miam.o`.

FIGURE B.4 – Résultat de `od -t xC reloc_miam.o`.

C'est assez difficile à lire... On va donc utiliser les programmes standards `objdump` et `readelf` sous Linux pour analyser les fichiers objets. Il s'agit d'outils capable d'analyser la séquence de bits du fichier pour y retrouver la structure d'un fichier objet, et d'afficher en clair les informations intéressantes. La commande `readelf -a reloc_miam.o` produit :

1. On verra en effet en sous-section B.4.6 que tous les symboles globaux doivent être rassemblés à la fin de la table des symboles.

B.4 Détail des sections

B.4.1 L'en-tête

```
0000000 7f 45 4c 46 01 02 01 00 00 00 00 00 00 00 00 00
0000020 00 01 00 08 00 00 00 01 00 00 00 00 00 00 00 00
0000040 00 00 00 b8 00 00 00 01 00 34 00 00 00 00 00 28
0000060 00 0b 00 08
```

Les 16 premiers octets constituent le champ `e_ident` (taille définie par la constante `EI_IDENT`). Les quatre premiers identifient le format : notons que `0x45, 0x4c` et `0x46` sont les codes ASCII des caractères `'E', 'L', 'F'`. Le cinquième octet `01` donne la classe du fichier, ici `ELFCLASS32`, ce qui signifie que les adresses sont exprimées sur 32 bits. Le sixième donne le type de codage des données, ici `2`, ce qui signifie que les données sont codées en complément à deux, avec les bits les plus significatifs occupant les adresses les plus basses (big endian).

Par ailleurs on repère : la taille de l'en-tête (0x34) ; le déplacement par rapport au début du fichier donnant accès à la table des sections (0xb8 octets = 184 en décimal) ; la taille d'une entrée de la table des sections (0x28 = 40 octets), le nombre d'entrées dans la table des sections (0x0b = 11). L'entrée numéro 8 dans la table des sections est celle du descripteur de la table des noms de sections `.shstrtab` (celle-ci est indispensable pour savoir quelle section décrit un autre descripteur).

La commande `readelf -S reloc_miam.o` permet d'obtenir l'en-tête des sections du ELF et de savoir ainsi quelles sont les sections qui le cosntitue. On peut ainsi voir qu'il existe une section `.text` de type `.PROGBITS` (bits appartenant à un programme) se trouvant à l'offset `0x34` et faisant `0x20` octets.

```
There are 11 section headers, starting at offset 0xbc:

Section Headers:
  [Nr] Name              Type            Addr     Off    Size   ES Flg Lk Inf Al
  [ 0]                   NULL            00000000 000000 000000 00      0   0  0
  [ 1] .text             PROGBITS        00000000 000034 000020 00  AX  0   0  4
  [ 2] .rel.text         REL             00000000 000394 000018 08      9   1  4
  [ 3] .data             PROGBITS        00000000 000054 00000c 00  WA  0   0  4
  [ 4] .rel.data         REL             00000000 0003ac 000008 08      9   3  4
  [ 5] .bss              NOBITS          00000000 000060 000010 00  WA  0   0  1
  [ 6] .reginfo          MIPS_REGINFO    00000000 000060 000018 01      0   0  4
  [ 7] .pdr              PROGBITS        00000000 000078 000000 00      0   0  4
  [ 8] .shstrtab         STRTAB          00000000 000078 000042 00      0   0  1
  [ 9] .symtab           SYMTAB          00000000 000274 0000c0 10     10   6  4
  [10] .strtab           STRTAB          00000000 000334 00005e 00      0   0  1
Key to Flags:
  W (write), A (alloc), X (execute), M (merge), S (strings)
  I (info), L (link order), G (group), x (unknown)
  O (extra OS processing required) o (OS specific), p (processor specific)
```

FIGURE B.5 – Table des section : résultat de `readelf -S reloc_miam.o`.

B.4.2 La table des noms de sections (`.shstrtab`)

La table des noms de section est accessible par la commande `readelf -hex-dump=8 reloc_miam.o`. En effet, la section `.shstrtab` est à l'indice 8. La table des noms de sections est du type table de

chaînes de caractères. Elle contient les noms suivants :

```
0000 002e7379 6d746162 002e7374 72746162  .symtab .strtab
0010 002e7368 73747274 6162002e 72656c2e  .shstrtab .rel.
0020 74657874 002e7265 6c2e6461 7461002e text .rel.data .
0030 62737300 2e726567 696e666f 002e7064 bss .reginfo .pd
0040 7200                                 r.
```

C'est dans cette table que `readelf` va chercher les noms des sections pour remplir la table donnée figure B.5. Le format ELF impose certaines règles à respecter. La première chaîne doit être la chaîne vide (`0x00`). Chaque chaîne doit se terminer par le caractère de code ASCII 0 (comme en "C").

B.4.3 La section table des chaînes (`.strtab`)

Cette section (qui porte l'index 10) rassemble tous les noms des symboles utilisés (directement ou implicitement) dans le code. Les règles concernant cette section sont les mêmes que pour `.shstrtab` ci-dessus.

```
0000 002e7465 7874002e 64617461 002e6273  .text .data .bs
0010 73002e72 6567696e 666f002e 70647200 s .reginfo .pdr.
0020 6c756e63 6874696d 6500626f 75636c65 lunchtime boucle
0030 00627965 62796500 7461626c 65617500  byebye tableau
0040 64656275 745f636f 75727300 61647265 debut_cours adre
0050 7373655f 6c756e63 6874696d 6500     sse_lunchtime
```

C'est dans cette table que `readelf` va chercher les noms des symboles, aucun symbole n'est codé "en dur" dans les autres sections.

B.4.4 La section `.text`

La portion de fichier objet correspondant à la zone TEXT (les instructions) est le résultat de la commande `readelf -hex-dump=8 reloc_miam.o` :

```
0000 20090008 3c080000 8d080004 11090004 ....<...........
0010 00000000 21290001 08000003 00000000 ....!)..........
```

Une version plus lisible de la zone TEXT peut être obtenue avec `mips-objdump -d -section=.text reloc_miam.o` :

```
Disassembly of section .text:

00000000 <boucle-0xc>:
   0: 20090008  addi t1,zero,8
   4: 3c080000  lui t0,0x0
   8: 8d080004  lw t0,4(t0)

0000000c <boucle>:
   c: 11090004  beq t0,t1,20 <byebye>
  10: 00000000  nop
  14: 21290001  addi t1,t1,1
  18: 08000003  j c <boucle>
  1c: 00000000  nop
```

On peut par exemple constater que l'instruction `Lw $t0 , lunchtime` a été remplacé par `lui $t0 , 0x0 == 3c080000` et `lw $t0 , 0x4($t0) == 8d080004`.

Pour chaque instruction, les deux premiers octets codent le numéro de l'instruction et le registre $t0 ; les deux derniers correspondent à la valeur numérique de `lunchtime`. En fait la valeur numérique pour `lui` devrait être les 16 bits de poids fort de `lunchtime` et la valeur numérique pour `lw` devrait être les 16 bits de poids faible de `lunchtime`. Cependant, la valeur de `lunchtime` n'est pas connue. En effet, `lunchtime` est une adresse est cette adresse ne sera connue que lors du chargement final en mémoire, les 16 bits sont donc à zéro ou une valeur temporaire (ici 4 pour lw) en attendant d'être replacés par une valeur lors du chargement grâce à une donnée de relocation `rel.text` qui est associée à cette instruction (cf. section B.4.7).

B.4.5 La section `.data`

La portion de fichier objet correspondant à la zone DATA (les données initialisées) est donnée par la commande `readelf -hex-dump=3 reloc_miam.o` :

```
0000  00000008 0000000c 00000004
```

La valeur 8 indiquée (pointée par) `debut_cours` est codée sur les 4 premiers octets (c'est un `.word`). La valeur 12 (`0x0000000c == 12`) indiquée (pointée par) `lunchtime` est codée sur les 4 octets suivants (c'est un `.word`). Les 4 octets suivants devraient contenir la valeur immédiate d'adresse correspondant à l'étiquette `lunchtime`. Mais comme l'adresse de `lunchtime` ne sera connue que lors du chargement final en mémoire, les 4 octets codent une valeur, 4, qui est en fait une donnée de translation. Cette information permettra, avec celles contenues dans la zone `.rel.data`, de calculer l'adresse finale de `lunchtime` et donc la valeur de `adresse_lunchtime`.

B.4.6 La section table des symboles

La table des symboles d'un fichier objet est donnée par la commande `readelf -s reloc_miam.o` :

```
Symbol table '.symtab' contains 12 entries:
   Num:    Value  Size Type    Bind   Vis      Ndx Name
     0: 00000000     0 NOTYPE  LOCAL  DEFAULT  UND
     1: 00000000     0 SECTION LOCAL  DEFAULT    1 .text
     2: 00000000     0 SECTION LOCAL  DEFAULT    3 .data
     3: 00000000     0 SECTION LOCAL  DEFAULT    5 .bss
     4: 00000000     0 SECTION LOCAL  DEFAULT    6 .reginfo
     5: 00000000     0 SECTION LOCAL  DEFAULT    7 .pdr
     6: 00000004     0 NOTYPE  LOCAL  DEFAULT    3 lunchtime
     7: 0000000c     0 NOTYPE  LOCAL  DEFAULT    1 boucle
     8: 00000020     0 NOTYPE  LOCAL  DEFAULT    1 byebye
     9: 00000000     0 NOTYPE  LOCAL  DEFAULT    5 tableau
    10: 00000000     0 NOTYPE  LOCAL  DEFAULT    3 debut_cours
    11: 00000008     0 NOTYPE  LOCAL  DEFAULT    3 adresse_lunchtime
```

Le type décrivant une entrée de la table des symboles est `struct Elf32Sym`. Une entrée occupe 16 octets, il y a ici 12 entrées. La première entrée est à zéro (elle sert de sentinelle). Les entrées de numéros 1 à 5 sont des symboles spéciaux qui représentent en fait respectivement les sections `.text`, `.data`, `.bss`, `.reginfo` et `.pdr`. On remarque en effet qu'elles ont le type `SECTION` (constante 3). Le champ `Ndx` de ces entrées indique dans quelle section le symbole est défini (par exemple `tableau` est

défini dans la section `5 == .bss`). Le champ `Value` indique l'offset a appliquer à partir de la section des symboles pour trouver l'adresse des symboles. Les entrées numéros 6 à 11 sont de vrais symboles de l'utilisateur, correspondants aux étiquettes `lunchtime`, `boucle`, `byebye` , `tableau`, `debut_cours` et `adresse_lunchtime`.
Les différents champ de la structure `Elf32Sym` sont les suivants :

- `st_name` : index du symbole dans la table des noms de symboles. Lorsque le symbole est de type `STT_SECTION`, aucun nom ne lui est associé. La valeur de l'index est alors 0 (qui désigne donc la chaîne vide, d'après les contraintes sur `.strtab`).
- `st_value` : il vaut 0 pour un symbole non défini ; pour un symbole défini localement, `st_value` est le déplacement (en octets) par rapport au début de la zone de définition.
- `st_shndx` : indique l'index (dans la table des en-têtes de sections) de la section où le symbole est défini. Si le symbole n'est défini dans aucune section (symbole externe), cet index vaut 0. Ainsi **l'index 0 de la table des en-têtes de section est une sentinelle qui sert à marquer les symboles externes**. Si le symbole est de type `STT_SECTION`, cet index est directement l'index de la section.
- `st_size` et `st_other` : non utilisés dans le projet (`st_size` sert à associer une taille de donnée au symbole).
- `st_info` : ce champ codé sur un octet sert en fait à coder 2 champs : le champ "bind" et le champ "type". Les macros suivantes (définie dans les fichiers d'en-tête du format ELF) permettent d'encoder ou de décoder le champ `st_info` :

  ```
  #define ELF32_ST_BIND(i)   ((i)>>4)               /* de info vers bind */
  #define ELF32_ST_TYPE(i)   ((i)&0xf)              /* de info vers type */
  #define ELF32_ST_INFO(b,t) (((b)<<4)+((t)&0xf))   /* de (bind,type) vers info */
  ```

  Le champ "bind" indique la portée du symbole. Dans le cadre du projet, on considère les 2 cas suivants :
  - `STB_LOCAL` (constante 0) indique que le symbole est défini localement et non exporté.
  - `STB_GLOBAL` (constante 1) indique que le symbole est soit défini et exporté, soit non défini et importé. Il faut noter qu'à l'édition de lien, il est interdit à 2 fichiers distincts de définir deux symboles de même nom ayant la portée `STB_GLOBAL`.

  Dans le cadre du projet, on ne considère que les 2 cas suivants pour le champ "type" :
  - `STT_SECTION` (constante 3) indiquant que le symbole désigne en fait une section.
  - `STT_NOTYPE` (constante 0), dans les autres cas.

Le format ELF impose certaines contraintes sur l'ordre des symboles dans la table : le premier symbole n'est pas utilisé. Tous les symboles globaux doivent être regroupés à la fin de la table.

B.4.7 Les sections de relocation

Les données de relocation `.rel.text` décrivent comment réécrire certains champs d'instruction incomplets le codage des instructions de la zone `.text`. Ces instructions sont partiellement codées, car contenant des références à des symboles (étiquettes) dont on ne peut pas connaître l'adresse au moment de la fabrication du fichier objet. L'instruction est codée en réservant les 4 octets nécessaires au stockage de la valeur d'adresse mais la valeur indiquée dans le champ est provisoire. Cependant, cette valeur est très importante, comme nous allons le voir. Elle dépend du mode de relocation, et permet de calculer la valeur finale du champs. Les données de relocation `.rel.date`, selon le même principe, décrivent comment réécrire certaines valeurs dans la section data.

Format de la table de relocation

Une entrée de la table de relocation est représentée par le type `struct Elf32_Rel` :

- `r_offset` : contient un décalage en octets par rapport au début de la section. Cette valeur indique la position dans la zone à reloger de l'instruction qui contient un champ incomplet.
- `r_info` : est un champ codé sur 4 octets composé en fait de deux champs :.
  - Le champ "`sym`" est l'index du symbole à reloger dans la table des symboles.
  - Le champ "`type`" indique le mode de calcul de la relocation. Dans le cadre du projet, il prend par exemple les valeurs `R_MIPS_32` (constante 2) ou `R_MIPS_LO16` (constante 6).

Les macros de codage/décodage du champ `r_info` défini dans les fichiers d'en-têtes de la librairie ELF sont :

```
#define ELF32_R_SYM(i)    ((i)>>8)                      /* info vers sym */
#define ELF32_R_TYPE(i)   ((unsigned char)(i))          /* info vers type */
#define ELF32_R_INFO(s,t) (((s)<<8)+(unsigned char)(t)) /* (sym,type) vers info */
```

Modes de calcul de la relocation

Détaillons maintenant le mode de calcul de la relocation, c'est-à-dire la valeur que l'éditeur de lien (ou le chargeur de notre simulateur) met finalement dans les champs incomplets des instructions. Les notations sont les suivantes :

V désigne la **V**aleur "finale" du champ accueillant le relogement.
P désigne la **P**lace, c'est à dire l'adresse "finale" de l'élément à reloger.
S désigne l'adresse "finale" du **S**ymbole par rapport auquel on reloge.
A désigne la valeur à **a**jouter pour calculer la valeur du champ à reloger (c'est la valeur du champ avant relogement).
AHL désigne un autre type de valeur à **a**jouter qui est calculée à partir de la valeur provisoire A[2].

Les adresses finales des sections `.text`, `.data`, `.bss` sont normalement déterminées lors de l'édition de liens. Dans notre simulateur, elles sont calculées lors du chargement des sections en mémoire suivant les contraintes définies section **??**.

Le mode de calcul dépend du type de relocation, codé dans le champs `type` de `r_info`. Les principaux modes de calcul sont :

- `R_MIPS_32` (constante 2) : la valeur mise à l'adresse P vaut $V = S + A$. Ce mode sert pour les adressages directs.
- `R_MIPS_26` (constante 4) : le calcul se décompose en plusieurs étapes :
  - calcul de l'adresse de saut (comme décrit section 2.3.2) : $(P\&0xf0000000)+S$
  - ou logique avec A décalé de 2 à gauche : $(A << 2) \mid ((P\&0xf0000000)+S)$
  - résultat décalé de 2 à droite : $V = ((A << 2) \mid ((P\&0xf0000000)+S)) >> 2$

  Ce mode sert pour les adressages absolus alignés sur 256Mo (pour les *J-instructions*).
- `R_MIPS_HI16` (constante 5) : la valeur mise à l'adresse P vaut $V = (AHL + S - (short)(AHL + S)) >> 16$. Ce mode sert pour remplacement des accès à la section data par étiquette (`lw, sw, lb, sb`...). Une relocation `R_MIPS_HI16` est toujours suivi d'une relocation `R_MIPS_LO16` car la valeur AHL est calculée par $(AHI << 16) + (short)(ALO)$ ou AHI est le A de l'instruction ayant une relocation `R_MIPS_HI16` et ALO est le A de l'instruction ayant une relocation `R_MIPS_LO16`.
- `R_MIPS_LO16` (constante 6) : la valeur mise à l'adresse P vaut $AHL + S$. Ce mode sert pour les adressages immédiats avec une valeur sur 16 bits.

La section de relocation `.rel.data`

La relocation est peut-être plus simple à comprendre en regardant le résultat de la commande `readelf -r reloc_miam.o`. Sur l'exemple, la table des relocations en data texte est :

```
Relocation section '.rel.data' at offset 0x3ac contains 1 entries:
 Offset     Info    Type            Sym.Value  Sym. Name
00000008  00000202 R_MIPS_32         00000000   .data
```

Ici on va chercher à calculer $V = S + A$. Le champ `info` vaut 0x00000202, donc se décompose en :

```
sym = (info)>>8 = 0x00000202 >> 8 = 0x00000002 = 2
type = (unsigned char)(info) = (unsigned char)(0x00000202) = 0x02
```

Le champ `type` vaut 2, ce qui signifie que le mode de relocation est `R_MIPS_32`. Il faut donc déterminer A et S. Le champ `sym` vaut également 2 ce qui signifie qu'il s'agit de l'adresse de la zone `.data`. On a donc S = adresse de `.data` = $0x0$. A est la valeur du champ présent dans le code à l'adresse de l'instruction. Le champ `r_offet` vaut `0x8`, si on se réfère à la section B.4.5, la valeur sur 32 bits à l'adresse 8 de la section data est 0x00000004. Nous avons donc $A = 0x4$ et $V = S + A = 0x0 + 0x4 = 0x4$. On va trouver que le contenu de P sera égal à l'adresse finale de la zone `.data` + 4. C'est normal, on fait ici référence à `lunchtime`, qui est le deuxième mots de 32 bits de la zone data.

La section de relocation `.rel.text`

Pour la zone TEXT, la relocation fonctionne de la même manière :

```
Relocation section '.rel.text' at offset 0x394 contains 3 entries:
 Offset     Info    Type            Sym.Value  Sym. Name
00000004  00000205 R_MIPS_HI16       00000000   .data
00000008  00000206 R_MIPS_LO16       00000000   .data
00000018  00000104 R_MIPS_26         00000000   .text
```

Pour la première entrée le champs `r_offet` vaut `0x4`, ce qui signifie que l'instruction à reloger est la deuxième instruction de la zone `.text`. Effectivement, il s'agit bien de `lui $t0, 0x0`.

Le champ `info` vaut 0x00000205, donc se décompose en :

```
sym = (info)>>8 = 0x00000205 >> 8 = 0x00000002 = 2
type = (unsigned char)(info) = (unsigned char)(0x00000205) = 0x05
```

Le champ `type` vaut 5, ce qui signifie que le mode de relocation est `R_MIPS_HI16`. Il faut donc déterminer S et AHL. Pour calculer ce dernier il faut récupérer AHI et ALO à partir des instructions. Le AHI est le A de l'instruction en adresse $0x4$ soit `lui $t0, 0x0 == 3c080000`. Dans le cas d'une relocation `R_MIPS_HI16` se sont les 16 bits de poids faible que l'on récupère soit $AHI = A = 0x0000$. Le ALO est le A de l'instruction en adresse $0x8$ soit `lw $t0,4($t0) == 8d080004`. Dans le cas d'une relocation `R_MIPS_LO16` se sont les 16 bits de poids faible que l'on récupère soit $ALO = A = 0x0004$. Nous pouvons donc calculer $AHL = (AHI << 16) + (short)(ALO) = 0x0 << 16 + (short)0x4 = 0x000004$

Le champ `sym` vaut 2 et correspond au début de la zone `.data`. Ici S = `Val.-sym` = 00000000.

V vaut donc $V = (AHL + S - (short)AHL + S) >> 16 = (0x04 + 0x0 - (short)0x4 + 0x0) >> 16 = (0x0) >> 16 = 0x0$, les 16 bits de poids faible de l'instruction prennent donc la valeur 0000. C'est bien ce qui était codé à la fin de 3c080000

Essayez maintenant d'effectuer la relocation des entrées 2 et 3 de la table.

B.4.8 Autres sections

Il reste trois sections à décrire :

- La section `.bss` contient uniquement la taille à réserver en mémoire pour les données non initialisées.
- La section `.reginfo` (*Register Information Section*) indique l'usage des registres dans le fichier objet. On ne s'y intéressera pas dans le projet.
- La section `.pdr` (*Procedure Descriptor*). On ne s'y intéressera pas dans le projet.

Annexe C

Spécifications détaillées des instructions

Cette annexe contient les spécifications des instructions étudiées dans ce projet. Elles sont directement issues de la documentation du MIPS fournie par le *Architecture For Programmers Volume II* de *MIPS Technologies* [4].

C.1 Définitions et notations

Commençons par rappeller quelques définitions et notations utiles.

Octet/Mot Un *octet* (byte en anglais) est une suite de 8 bits qui constitue la plus petite entité que l'on peut adresser sur la machine. La concaténation de deux octets forme un *demi-mot* (half-word) de 16 bits, et la concaténation de quatre octets, ou de deux demi-mots, forme un *mot* (word) de 32 bits. Les bits sont numérotés de la droite (poids faible) vers la gauche (poids fort) de 0 à 7 pour l'octet, de 0 à 15 pour un demi-mot et de 0 à 31 pour un mot.

0	1	1	0	1	1	1	0

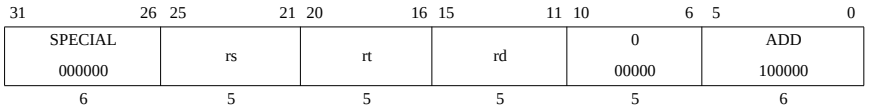
FIGURE C.1 – Octet

Représentation hexadécimale d'un octet/mot On représente par 0xij la valeur d'un octet dont les 4 bits de poids fort valent i et les 4 bits de poids faible j (avec i, $j \in ([0...9, A...F])$. Par exemple, la valeur de l'octet de la figure C.1 s'écrit 0x6E en hexadécimal. Pour un demi-mot ou un mot, on aura respectivement 4 ou 8 chiffres hexadécimaux, chacun représentant 4 bits.

Codage binaire d'un entier non signé Quand on parle d'un entier non signé codé sur n bits ou plus simplement d'un entier codé sur n bits, il s'agit de sa représentation en base 2 sur n bits, donc d'une valeur entière comprise entre 0 et $2^n - 1$. Un entier codé sur un octet a donc une valeur comprise entre 0 et 255 correspondant aux images binaires 0x00 à 0xFF, un entier codé sur un demi-mot a une valeur comprise entre 0 et 65535 correspondant aux images binaires 0x0000 à 0xFFFF, et un entier codé sur un mot a une valeur comprise entre 0 et 4294967295 correspondant aux images binaires 0x00000000 à 0xFFFFFFFF. Les adresses du processeur de la machine MIPS sont des entiers non signés sur 32 bits.

Codage binaire d'un entier signé Les entiers signés sont représentés en complément à 2. Le codage sur n bits du nombre i est la représentation en base 2 sur n bits de $2^n + i$, si $-2^{n-1} \leq i \leq -1$, et de i, si $0 \leq i \leq 2^{n-1} - 1$. Un entier signé codé sur un octet est compris entre -128 à 127 correspondant aux images binaires 0x80 à 0x7F. Un entier signé sur un demi-mot est compris entre -32768 et 32767 correspondant à l'intervalle binaire 0x8000 à 0x7FFF. Enfin, un entier signé sur un mot est compris entre -2147483648 et 2147483647 correspondant à l'intervalle binaire 0x80000000 à 0x7FFFFFFF. On remarque que le bit de plus fort poids d'un octet/mot/long mot représentant un entier négatif est toujours égal à 1 alors qu'il vaut 0 pour un nombre positif (c'est le bit de signe).

Add Word — ADD



Format: `ADD rd, rs, rt` **MIPS32**

Purpose:

To add 32-bit integers. If an overflow occurs, then trap.

Description: `GPR[rd] ← GPR[rs] + GPR[rt]`

The 32-bit word value in GPR *rt* is added to the 32-bit value in GPR *rs* to produce a 32-bit result.

- If the addition results in 32-bit 2's complement arithmetic overflow, the destination register is not modified and an Integer Overflow exception occurs.
- If the addition does not overflow, the 32-bit result is placed into GPR *rd*.

Restrictions:

None

Operation:

```
temp ← (GPR[rs]31||GPR[rs]31..0) + (GPR[rt]31||GPR[rt]31..0)
if temp32 ≠ temp31 then
   SignalException(IntegerOverflow)
else
   GPR[rd] ← temp
endif
```

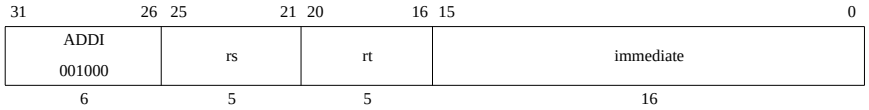
Exceptions:

Integer Overflow

Programming Notes:

ADDU performs the same arithmetic operation but does not trap on overflow.

Add Immediate Word — ADDI



Format: `ADDI rt, rs, immediate` **MIPS32**

Purpose:

To add a constant to a 32-bit integer. If overflow occurs, then trap.

Description: `GPR[rt] ¨ GPR[rs] + immediate`

The 16-bit signed *immediate* is added to the 32-bit value in GPR *rs* to produce a 32-bit result.

- If the addition results in 32-bit 2's complement arithmetic overflow, the destination register is not modified and an Integer Overflow exception occurs.
- If the addition does not overflow, the 32-bit result is placed into GPR *rt*.

Restrictions:

None

Operation:

```
temp ← (GPR[rs]31||GPR[rs]31..0) + sign_extend(immediate)
if temp32 ≠ temp31 then
   SignalException(IntegerOverflow)
else
   GPR[rt] ← temp
endif
```

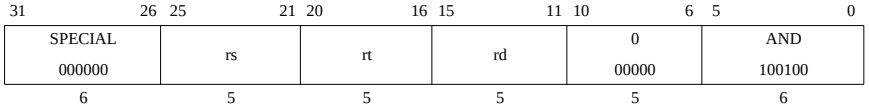
Exceptions:

Integer Overflow

Programming Notes:

ADDIU performs the same arithmetic operation but does not trap on overflow.

And AND



Format: `AND rd, rs, rt` **MIPS32**

Purpose:

To do a bitwise logical AND

Description: `GPR[rd] ← GPR[rs] AND GPR[rt]`

The contents of GPR *rs* are combined with the contents of GPR *rt* in a bitwise logical AND operation. The result is placed into GPR *rd*.

Restrictions:

None

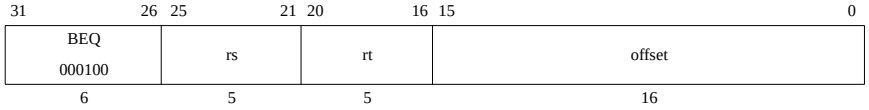
Operation:

```
GPR[rd] ← GPR[rs] and GPR[rt]
```

Exceptions:

None

Branch on Equal BEQ



Format: `BEQ rs, rt, offset` **MIPS32**

Purpose:

To compare GPRs then do a PC-relative conditional branch

Description: `if GPR[rs] = GPR[rt] then branch`

An 18-bit signed offset (the 16-bit *offset* field shifted left 2 bits) is added to the address of the instruction following the branch (not the branch itself), in the branch delay slot, to form a PC-relative effective target address.

If the contents of GPR *rs* and GPR *rt* are equal, branch to the effective target address after the instruction in the delay slot is executed.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:      target_offset ← sign_extend(offset || 0^2)
           condition ← (GPR[rs] = GPR[rt])
I+1:    if condition then
           PC ← PC + target_offset
        endif
```

Exceptions:

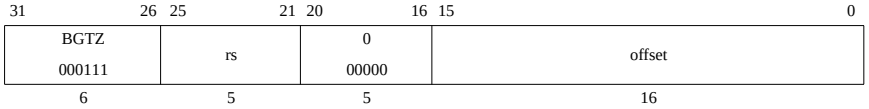
None

Programming Notes:

With the 18-bit signed instruction offset, the conditional branch range is ± 128 Kbytes. Use jump (J) or jump register (JR) instructions to branch to addresses outside this range.

BEQ r0, r0 offset, expressed as B offset, is the assembly idiom used to denote an unconditional branch.

Branch on Greater Than Zero — BGTZ



Format: `BGTZ rs, offset` **MIPS32**

Purpose:

To test a GPR then do a PC-relative conditional branch

Description: `if GPR[rs] > 0 then branch`

An 18-bit signed offset (the 16-bit *offset* field shifted left 2 bits) is added to the address of the instruction following the branch (not the branch itself), in the branch delay slot, to form a PC-relative effective target address.

If the contents of GPR *rs* are greater than zero (sign bit is 0 but value not zero), branch to the effective target address after the instruction in the delay slot is executed.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:    target_offset ← sign_extend(offset || 0^2)
      condition ← GPR[rs] > 0^GPRLEN
I+1:  if condition then
         PC ← PC + target_offset
      endif
```

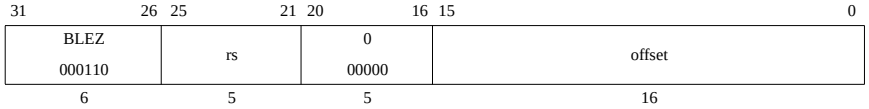
Exceptions:

None

Programming Notes:

With the 18-bit signed instruction offset, the conditional branch range is ± 128 KBytes. Use jump (J) or jump register (JR) instructions to branch to addresses outside this range.

Branch on Less Than or Equal to Zero — BLEZ



Format: `BLEZ rs, offset` **MIPS32**

Purpose:

To test a GPR then do a PC-relative conditional branch

Description: `if GPR[rs] ≤ 0 then branch`

An 18-bit signed offset (the 16-bit *offset* field shifted left 2 bits) is added to the address of the instruction following the branch (not the branch itself), in the branch delay slot, to form a PC-relative effective target address.

If the contents of GPR *rs* are less than or equal to zero (sign bit is 1 or value is zero), branch to the effective target address after the instruction in the delay slot is executed.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:    target_offset ← sign_extend(offset || 0^2)
      condition ← GPR[rs] ≤ 0^GPRLEN
I+1:  if condition then
         PC ← PC + target_offset
      endif
```

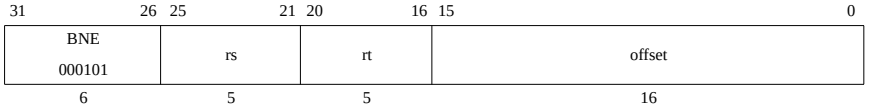
Exceptions:

None

Programming Notes:

With the 18-bit signed instruction offset, the conditional branch range is ± 128 KBytes. Use jump (J) or jump register (JR) instructions to branch to addresses outside this range.

Branch on Not Equal — BNE



Format: `BNE rs, rt, offset` **MIPS32**

Purpose:

To compare GPRs then do a PC-relative conditional branch

Description: `if GPR[rs] ≠ GPR[rt] then branch`

An 18-bit signed offset (the 16-bit *offset* field shifted left 2 bits) is added to the address of the instruction following the branch (not the branch itself), in the branch delay slot, to form a PC-relative effective target address.

If the contents of GPR *rs* and GPR *rt* are not equal, branch to the effective target address after the instruction in the delay slot is executed.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:    target_offset ← sign_extend(offset || 0^2)
      condition ← (GPR[rs] ≠ GPR[rt])
I+1:  if condition then
         PC ← PC + target_offset
      endif
```

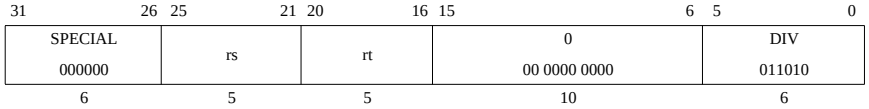
Exceptions:

None

Programming Notes:

With the 18-bit signed instruction offset, the conditional branch range is ± 128 KBytes. Use jump (J) or jump register (JR) instructions to branch to addresses outside this range.

Divide Word — DIV



Format: `DIV rs, rt` **MIPS32**

Purpose:

To divide a 32-bit signed integers

Description: `(HI, LO) ← GPR[rs] / GPR[rt]`

The 32-bit word value in GPR *rs* is divided by the 32-bit value in GPR *rt*, treating both operands as signed values. The 32-bit quotient is placed into special register *LO* and the 32-bit remainder isplaced into special register *HI*.

No arithmetic exception occurs under any circumstances.

Restrictions:

If the divisor in GPR *rt* is zero, the arithmetic result value is **UNPREDICTABLE**.

Operation:

```
q  ← GPR[rs]31..0 div GPR[rt]31..0
LO ← q
r  ← GPR[rs]31..0 mod GPR[rt]31..0
HI ← r
```

Exceptions:

None

Programming Notes:

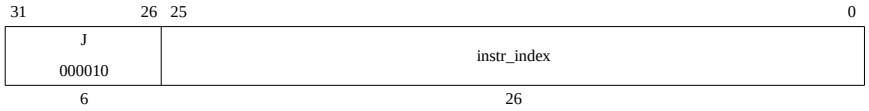
No arithmetic exception occurs under any circumstances. If divide-by-zero or overflow conditions are detected and some action taken, then the divide instruction is typically followed by additional instructions to check for a zero divisor and/or for overflow. If the divide is asynchronous then the zero-divisor check can execute in parallel with the divide. The action taken on either divide-by-zero or overflow is either a convention within the program itself, or more typically within the system software; one possibility is to take a BREAK exception with a *code* field value to signal the problem to the system software.

As an example, the C programming language in a UNIX® environment expects division by zero to either terminate the program or execute a program-specified signal handler. C does not expect overflow to cause any exceptional condition. If the C compiler uses a divide instruction, it also emits code to test for a zero divisor and execute a BREAK instruction to inform the operating system if a zero is detected.

In some processors the integer divide operation may proceed asynchronously and allow other CPU instructions to execute before it is complete. An attempt to read *LO* or *HI* before the results are written interlocks until the results are ready. Asynchronous execution does not affect the program result, but offers an opportunity for performance improvement by scheduling the divide so that other instructions can execute in parallel.

Historical Perspective:

In MIPS 1 through MIPS III, if either of the two instructions preceding the divide is an MFHI or MFLO, the result of the MFHI or MFLO is UNPREDICTABLE. Reads of the HI or LO special register must be separated from subsequent instructions that write to them by two or more instructions. This restriction was removed in MIPS IV and MIPS32 and all subsequent levels of the architecture.



Format: `J target` **MIPS32**

Purpose:

To branch within the current 256 MB-aligned region

Description:

This is a PC-region branch (not PC-relative); the effective target address is in the “current” 256 MB-aligned region. The low 28 bits of the target address is the *instr_index* field shifted left 2 bits. The remaining upper bits are the corresponding bits of the address of the instruction in the delay slot (not the branch itself).

Jump to the effective target address. Execute the instruction that follows the jump, in the branch delay slot, before executing the jump itself.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:
I+1:PC ← PC_GPRLEN-1..28 || instr_index || 0^2
```

Exceptions:

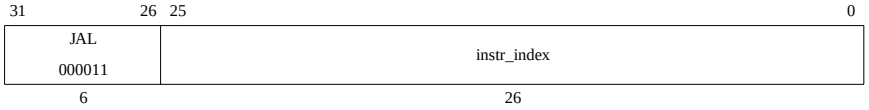
None

Programming Notes:

Forming the branch target address by catenating PC and index bits rather than adding a signed offset to the PC is an advantage if all program code addresses fit into a 256 MB region aligned on a 256 MB boundary. It allows a branch from anywhere in the region to anywhere in the region, an action not allowed by a signed relative offset.

This definition creates the following boundary case: When the jump instruction is in the last word of a 256 MB region, it can branch only to the following 256 MB region containing the branch delay slot.

Jump and Link — JAL



Format: `JAL target` **MIPS32**

Purpose:

To execute a procedure call within the current 256 MB-aligned region

Description:

Place the return address link in GPR 31. The return link is the address of the second instruction following the branch, at which location execution continues after a procedure call.

This is a PC-region branch (not PC-relative); the effective target address is in the "current" 256 MB-aligned region. The low 28 bits of the target address is the *instr_index* field shifted left 2 bits. The remaining upper bits are the corresponding bits of the address of the instruction in the delay slot (not the branch itself).

Jump to the effective target address. Execute the instruction that follows the jump, in the branch delay slot, before executing the jump itself.

Restrictions:

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

```
I:  GPR[31]← PC + 8
I+1:PC     ← PC_GPRLEN-1..28 || instr_index || 0^2
```

Exceptions:

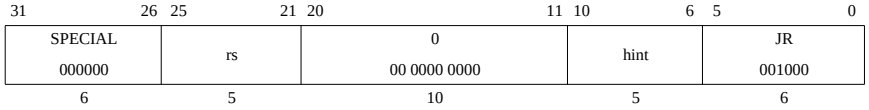
None

Programming Notes:

Forming the branch target address by catenating PC and index bits rather than adding a signed offset to the PC is an advantage if all program code addresses fit into a 256 MB region aligned on a 256 MB boundary. It allows a branch from anywhere in the region to anywhere in the region, an action not allowed by a signed relative offset.

This definition creates the following boundary case: When the branch instruction is in the last word of a 256 MB region, it can branch only to the following 256 MB region containing the branch delay slot.

Jump Register — JR



Format: `JR rs` **MIPS32**

Purpose:

To execute a branch to an instruction address in a register

Description: `PC ← GPR[rs]`

Jump to the effective target address in GPR *rs*. Execute the instruction following the jump, in the branch delay slot, before jumping.

For processors that implement the MIPS16e ASE, set the *ISA Mode* bit to the value in GPR *rs* bit 0. Bit 0 of the target address is always zero so that no Address Exceptions occur when bit 0 of the source register is one

Restrictions:

The effective target address in GPR *rs* must be naturally-aligned. For processors that do not implement the MIPS16e ASE, if either of the two least-significant bits are not zero, an Address Error exception occurs when the branch target is subsequently fetched as an instruction. For processors that do implement the MIPS16e ASE, if bit 0 is zero and bit 1 is one, an Address Error exception occurs when the jump target is subsequently fetched as an instruction.

In release 1 of the architecture, the only defined hint field value is 0, which sets default handling of JR. In Release 2 of the architecture, bit 10 of the hint field is used to encode an instruction hazard barrier. See the JR.HB instruction description for additional information.

Processor operation is **UNPREDICTABLE** if a branch, jump, ERET, DERET, or WAIT instruction is placed in the delay slot of a branch or jump.

Operation:

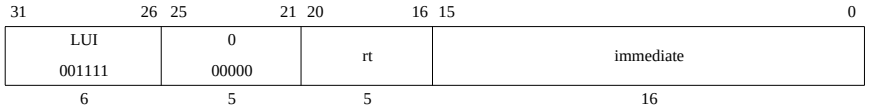
```
I:  temp ← GPR[rs]
I+1:if Config1_CA = 0 then
       PC ← temp
    else
       PC ← temp_GPRLEN-1..1 || 0
       ISAMode ← temp_0
    endif
```

Exceptions:

None

Programming Notes:

Software should use the value 31 for the *rs* field of the instruction word on return from a JAL, JALR, or BGEZAL, and should use a value other than 31 for remaining uses of JR.



Format: `LUI rt, immediate` **MIPS32**

Purpose:

To load a constant into the upper half of a word

Description: `GPR[rt] ← immediate || 0^16`

The 16-bit *immediate* is shifted left 16 bits and concatenated with 16 bits of low-order zeros. The 32-bit result is placed into GPR *rt*.

Restrictions:

None

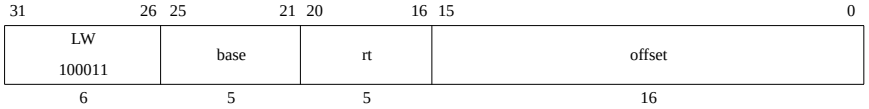
Operation:

```
GPR[rt] ← immediate || 0^16
```

Exceptions:

None

Load Word — LW



Format: `LW rt, offset(base)` **MIPS32**

Purpose:

To load a word from memory as a signed value

Description: `GPR[rt] ← memory[GPR[base] + offset]`

The contents of the 32-bit word at the memory location specified by the aligned effective address are fetched, sign-extended to the GPR register length if necessary, and placed in GPR *rt*. The 16-bit signed *offset* is added to the contents of GPR *base* to form the effective address.

Restrictions:

The effective address must be naturally-aligned. If either of the 2 least-significant bits of the address is non-zero, an Address Error exception occurs.

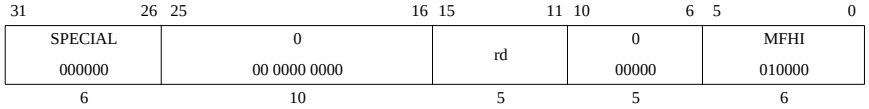
Operation:

```
vAddr ← sign_extend(offset) + GPR[base]
if vAddr1..0 ≠ 0^2 then
   SignalException(AddressError)
endif
(pAddr, CCA)← AddressTranslation (vAddr, DATA, LOAD)
memword← LoadMemory (CCA, WORD, pAddr, vAddr, DATA)
GPR[rt]← memword
```

Exceptions:

TLB Refill, TLB Invalid, Bus Error, Address Error, Watch

Move From HI Register — MFHI



Format: `MFHI rd` **MIPS32**

Purpose:

To copy the special purpose *HI* register to a GPR

Description: `GPR[rd] ← HI`

The contents of special register *HI* are loaded into GPR *rd*.

Restrictions:

None

Operation:

```
GPR[rd] ← HI
```

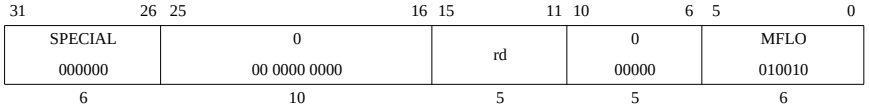
Exceptions:

None

Historical Information:

In the MIPS I, II, and III architectures, the two instructions which follow the MFHI must not moodify the HI register. If this restriction is violated, the result of the MFHI is **UNPREDICTABLE**. This restriction was removed in MIPS IV and MIPS32, and all subsequent levels of the architecture.

Move From LO Register — MFLO



Format: `MFLO rd` **MIPS32**

Purpose:

To copy the special purpose *LO* register to a GPR

Description: `GPR[rd] ← LO`

The contents of special register *LO* are loaded into GPR *rd*.

Restrictions: None

Operation:

```
GPR[rd] ← LO
```

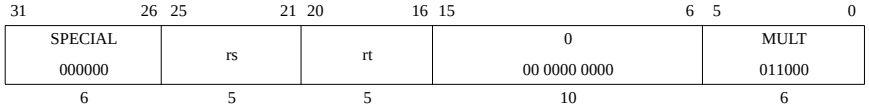
Exceptions:

None

Historical Information:

In the MIPS I, II, and III architectures, the two instructions which follow the MFHI must not moodify the HI register. If this restriction is violated, the result of the MFHI is **UNPREDICTABLE**. This restriction was removed in MIPS IV and MIPS32, and all subsequent levels of the architecture.

Multiply Word — MULT



Format: `MULT rs, rt` **MIPS32**

Purpose:

To multiply 32-bit signed integers

Description: `(HI, LO) ← GPR[rs] ×GPR[rt]`

The 32-bit word value in GPR *rt* is multiplied by the 32-bit value in GPR *rs*, treating both operands as signed values, to produce a 64-bit result. The low-order 32-bit word of the result is placed into special register *LO*, and the high-order 32-bit word is splaced into special register *HI*.

No arithmetic exception occurs under any circumstances.

Restrictions:

None

Operation:

```
prod   ← GPR[rs]31..0 × GPR[rt]31..0
LO     ← prod31..0
HI     ← prod63..32
```

Exceptions:

None

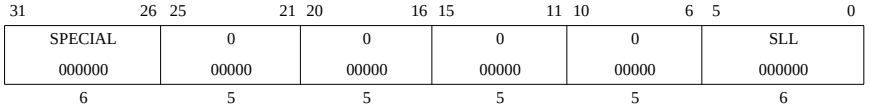
Programming Notes:

In some processors the integer multiply operation may proceed asynchronously and allow other CPU instructions to execute before it is complete. An attempt to read *LO* or *HI* before the results are written interlocks until the results are ready. Asynchronous execution does not affect the program result, but offers an opportunity for performance improvement by scheduling the multiply so that other instructions can execute in parallel.

Programs that require overflow detection must check for it explicitly.

Where the size of the operands are known, software should place the shorter operand in GPR *rt*. This may reduce the latency of the instruction on those processors which implement data-dependent instruction latencies.

No Operation NOP



Format: `NOP` **Assembly Idiom**

Purpose:

To perform no operation.

Description:

NOP is the assembly idiom used to denote no operation. The actual instruction is interpreted by the hardware as SLL r0, r0, 0.

Restrictions:

None

Operation:

None

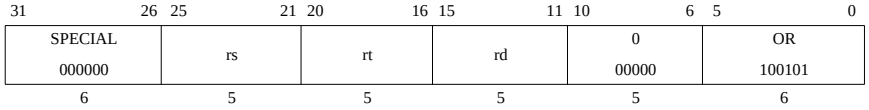
Exceptions:

None

Programming Notes:

The zero instruction word, which represents SLL, r0, r0, 0, is the preferred NOP for software to use to fill branch and jump delay slots and to pad out alignment sequences.

Or OR



Format: `OR rd, rs, rt` **MIPS32**

Purpose:

To do a bitwise logical OR

Description: `GPR[rd] ← GPR[rs] or GPR[rt]`

The contents of GPR *rs* are combined with the contents of GPR *rt* in a bitwise logical OR operation. The result is placed into GPR *rd*.

Restrictions:

None

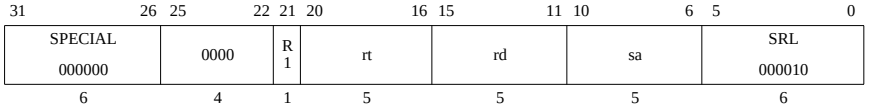
Operation:

```
GPR[rd] ← GPR[rs] or GPR[rt]
```

Exceptions:

None

Rotate Word Right — ROTR



Format: `ROTR rd, rt, sa` **SmartMIPS Crypto, MIPS32 Release 2**

Purpose:

To execute a logical right-rotate of a word by a fixed number of bits

Description: `GPR[rd] ← GPR[rt] ↔(right) sa`

The contents of the low-order 32-bit word of GPR *rt* are rotated right; the word result is placed in GPR *rd*. The bit-rotate amount is specified by *sa*.

Restrictions:

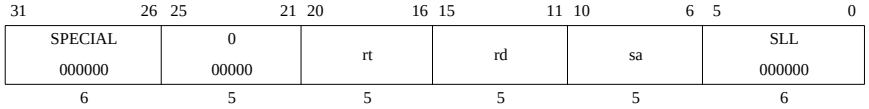
Operation:

```
if ((ArchitectureRevision() < 2) and (Config3SM = 0)) then
   UNPREDICTABLE
endif
s      ← sa
temp   ← GPR[rt]s-1..0 || GPR[rt]31..s
GPR[rd]← temp
```

Exceptions:

Reserved Instruction

Shift Word Left Logical — SLL



Format: `SLL rd, rt, sa` **MIPS32**

Purpose:

To left-shift a word by a fixed number of bits

Description: `GPR[rd] ← GPR[rt] << sa`

The contents of the low-order 32-bit word of GPR *rt* are shifted left, inserting zeros into the emptied bits; the word result is placed in GPR *rd*. The bit-shift amount is specified by *sa*.

Restrictions:

None

Operation:

```
s      ← sa
temp   ← GPR[rt](31-s)..0 || 0s
GPR[rd]← temp
```

Exceptions:

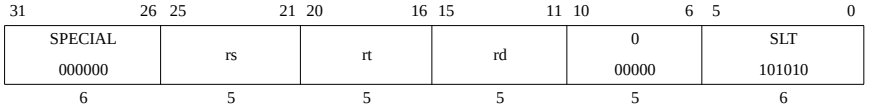
None

Programming Notes:

SLL r0, r0, 0, expressed as NOP, is the assembly idiom used to denote no operation.

SLL r0, r0, 1, expressed as SSNOP, is the assembly idiom used to denote no operation that causes an issue break on superscalar processors.

Set on Less Than SLT



Format: `SLT rd, rs, rt` **MIPS32**

Purpose:

To record the result of a less-than comparison

Description: `GPR[rd] ← (GPR[rs] < GPR[rt])`

Compare the contents of GPR *rs* and GPR *rt* as signed integers and record the Boolean result of the comparison in GPR *rd*. If GPR *rs* is less than GPR *rt*, the result is 1 (true); otherwise, it is 0 (false).

The arithmetic comparison does not cause an Integer Overflow exception.

Restrictions:

None

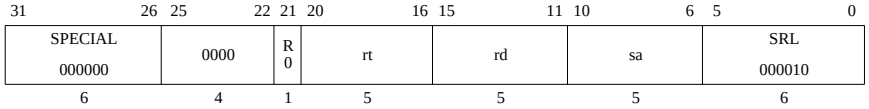
Operation:

```
if GPR[rs] < GPR[rt] then
   GPR[rd] ← 0^(GPRLEN-1) || 1
else
   GPR[rd] ← 0^GPRLEN
endif
```

Exceptions:

None

Shift Word Right Logical SRL



Format: `SRL rd, rt, sa` **MIPS32**

Purpose:

To execute a logical right-shift of a word by a fixed number of bits

Description: `GPR[rd] ← GPR[rt] >> sa (logical)`

The contents of the low-order 32-bit word of GPR *rt* are shifted right, inserting zeros into the emptied bits; the word result is placed in GPR *rd*. The bit-shift amount is specified by *sa*.

Restrictions:

None

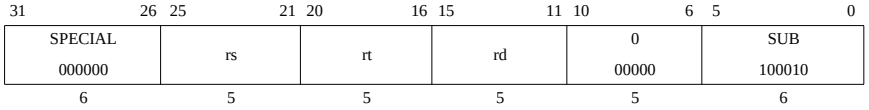
Operation:

```
s    ← sa
temp ← 0^s || GPR[rt]_31..s
GPR[rd]← temp
```

Exceptions:

None

Subtract Word — SUB



Format: `SUB rd, rs, rt` **MIPS32**

Purpose:

To subtract 32-bit integers. If overflow occurs, then trap

Description: `GPR[rd] ← GPR[rs] - GPR[rt]`

The 32-bit word value in GPR *rt* is subtracted from the 32-bit value in GPR *rs* to produce a 32-bit result. If the subtraction results in 32-bit 2's complement arithmetic overflow, then the destination register is not modified and an Integer Overflow exception occurs. If it does not overflow, the 32-bit result is placed into GPR *rd*.

Restrictions:

None

Operation:

```
temp ← (GPR[rs]31||GPR[rs]31..0) – (GPR[rt]31||GPR[rt]31..0)
if temp32 ≠ temp31 then
   SignalException(IntegerOverflow)
else
   GPR[rd] ← temp31..0
endif
```

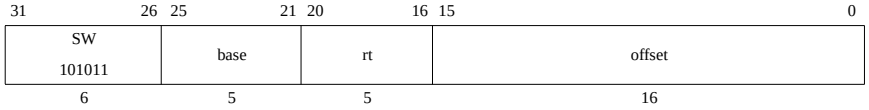
Exceptions:

Integer Overflow

Programming Notes:

SUBU performs the same arithmetic operation but does not trap on overflow.

Store Word — SW



Format: `SW rt, offset(base)` **MIPS32**

Purpose:

To store a word to memory

Description: `memory[GPR[base] + offset] ← GPR[rt]`

The least-significant 32-bit word of GPR *rt* is stored in memory at the location specified by the aligned effective address. The 16-bit signed *offset* is added to the contents of GPR *base* to form the effective address.

Restrictions:

The effective address must be naturally-aligned. If either of the 2 least-significant bits of the address is non-zero, an Address Error exception occurs.

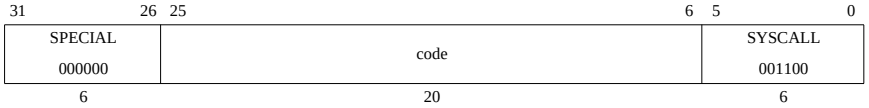
Operation:

```
vAddr ← sign_extend(offset) + GPR[base]
if vAddr1..0 ≠ 02 then
   SignalException(AddressError)
endif
(pAddr, CCA)← AddressTranslation (vAddr, DATA, STORE)
dataword← GPR[rt]
StoreMemory (CCA, WORD, dataword, pAddr, vAddr, DATA)
```

Exceptions:

TLB Refill, TLB Invalid, TLB Modified, Address Error, Watch

System Call SYSCALL



Format: `SYSCALL` **MIPS32**

Purpose:

To cause a System Call exception

Description:

A system call exception occurs, immediately and unconditionally transferring control to the exception handler.

The *code* field is available for use as software parameters, but is retrieved by the exception handler only by loading the contents of the memory word containing the instruction.

Restrictions:

None

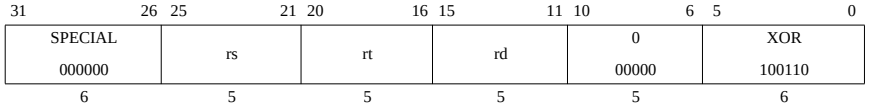
Operation:

```
SignalException(SystemCall)
```

Exceptions:

System Call

Exclusive OR XOR



Format: `XOR rd, rs, rt` **MIPS32**

Purpose:

To do a bitwise logical Exclusive OR

Description: `GPR[rd] ← GPR[rs] XOR GPR[rt]`

Combine the contents of GPR *rs* and GPR *rt* in a bitwise logical Exclusive OR operation and place the result into GPR *rd*.

Restrictions:

None

Operation:

```
GPR[rd] ← GPR[rs] xor GPR[rt]
```

Exceptions:

None