

RustOS

Système d'exploitation en Rust

Orphée Antoniadis

Projet de Bachelor - Prof. Florent Glück

Hepia ITI 3ème année

Semestre de Printemps 2017-2018

Résumé

Le but de ce projet est d'étudier le langage Rust, en particulier son utilisation pour l'implémentation d'un système d'exploitation de type *bare metal*. Le langage Rust se révèle particulièrement intéressant en tant que digne successeur de C : beaucoup plus robuste que ce dernier et potentiellement tout aussi rapide. La première partie du projet sera de comprendre les paradigmes de programmation utilisés par Rust ainsi que ses caractéristiques principales. Dans un deuxième temps, il s'agira d'implémenter un système d'exploitation très simple, similaire à celui réalisé au cours logiciel « Programmation système avancée » mais écrit en Rust plutôt qu'en C.

Table des matières

1	Introduction	8
1.1	Contexte	8
1.2	Objectif	8
2	Analyse	9
3	Conception	10
3.1	Environnement de développement	10
3.2	Technologies	10
3.3	Architecture	11
4	Rust	12
5	Exécution du <i>kernel</i>	13
5.1	Compilation	13
5.2	<i>Linking</i>	13
5.3	<i>Boot</i>	14
6	Gestion mémoire	16
6.1	Introduction	16
6.2	Segmentation	17
6.3	Pagination	22
6.4	Allocation dynamique	25
7	Périphériques	28
7.1	Ports	28
7.2	Interruptions et Exceptions	28
7.3	VGA	33
7.4	<i>Timer</i>	35
7.5	Clavier	36
8	Système de fichiers	37
8.1	Introduction	37
8.2	Structure	37
9	Tâches utilisateur	38
10	Résultats	39
11	Discussions	40
11.1	Problèmes rencontrés	40
11.2	Améliorations possibles	40
12	Conclusion	41
13	Références	42

Table des figures

1	Strucutre du fichier ELF	14
2	<i>Boot</i> d'une machine à base de BIOS	14
3	Exemple d'adressage mémoire	16
4	Protection mémoire avec un MMU	16
5	Translation d'adresse	17
6	Conversion d'une adresse logique en adresse linéaire	18
7	Structure d'un sélecteur de segment	18
8	Exemple d'une GDT	19
9	Structure d'une entrée dans la GDT	20
10	Modèle de segmentation de type <i>flat</i>	20
11	Descripteur de GDT	21
12	Structure d'une <i>Page Entry</i>	22
13	Exemple de pagination à 3 niveaux	23
14	Répertoire de pages adressant le <i>kernel</i> au début de la RAM	24
15	Répertoire de pages adressant le <i>kernel</i> à la fin de la RAM	24
16	Entête d'un bloc de mémoire dans le tas	25
17	Algorithme utilisé pour l'allocation dynamique dans le <i>kernel</i>	26
18	Etat initial de la chaîne d'entêtes	26
19	Allocation d'une page	27
20	Allocation d'une page et d'une table des pages	27
21	Table des interruptions et exceptions sur IA-32	29
22	Table de correspondance des IRQs	30
23	Différents types de descripteur d'interruption	31
24	Relation entre le registre IDTR et l'IDT	32
25	Structure d'un caractère en mode texte VGA	33
26	Couleurs disponibles en mode texte VGA	33
27	Structure d'un <i>scan code</i>	36

Remerciements

Convention typographique

Lors de la rédaction de ce document, les conventions typographique ci-dessous ont été adoptées.

- Tous les mots empruntés à la langue anglaise ont été écrits en *italique*
- Toute référence à un nom de fichier (ou dossier), un chemin d'accès, une utilisation de paramètre, variable, ou commande utilisable par l'utilisateur, est écrite avec la police d'écriture **Courier New**.
- Tout extrait de fichier ou de code est écrit selon le format suivant :

```
1  fn main() {  
2      println!("Hello, world!");  
3  }
```

Acronymes

BIOS *Basic Input Output System*. 14, 15

CPU *Central Processing Unit*. 18, 28, 31, 35

CRTC *Cathode Ray Tube Controller*. 33

ELF *Executable and Linkable Format*. 10, 11, 14

GCC *GNU Compiler Collection*. 11, 13

GDT *Global Descriptor Table*. 17–21, 31, 32

GRUB *GRand Unified Bootloader*. 11, 15

IA-32 *Intel Architecture, 32-bit*. 16, 28, 29, 35

IDT *Interrupt Descriptor Table*. 29–32

IRQ *Interrupt Request*. 30, 35, 36

ISO *International Organization for Standardization*. 11, 15

ISR *Interrupt Service Routine*. 28, 30, 31

LDT *Local Descriptor Table*. 17–19, 21

MBR *Master Boot Record*. 14, 15

MMIO *Memory Mapped Input/Output*. 28

MMU *Memory Management Unit*. 16, 17, 19, 21

NMI *Non Maskable Interrupt*. 30

OS *Operating System*. 10, 13–17, 19, 21, 23, 25, 32–34

PC *Personal Computer*. 14, 33, 35

PIC *Programmable Interrupt Controller*. 30

PIO *Port Input/Output*. 28

PIT *Programmable Interval Timer*. 35

RAM *Random Access Memory*. 16, 23, 24, 26, 32

TSS *Task State Segment*. 17

VGA *Video Graphics Array*. 33–35

VRAM *Video Random Access Memory*. 14, 33

1 Introduction

1.1 Contexte

1.2 Objectif

2 Analyse

3 Conception

3.1 Environnement de développement

La machine utilisée pour le développement du projet est un MacBook Pro avec un processeur Intel à 3 GHz. Il a quand même fallu utiliser une machine virtuelle (VMware) utilisant Linux (Ubuntu 16.04.4 LTS) pour la compilation. Ce choix a été fait car il existe beaucoup plus de documentation sur l'implémentation de systèmes d'exploitation sur Linux que sur Mac. Bien que Mac OS soit un système UNIX, les exécutables générés sur cet environnement n'ont pas le même format que ceux générés sur Linux qui sont au format ELF. Ceci rend le développement d'OS légèrement différent sur Mac OS.

3.2 Technologies

3.2.1 Nasm

Bien que le système d'exploitation développé devait être sur Rust, certaines parties ont dû être faites en assembleur car étant trop bas niveau pour le Rust. Ces éléments seront décrits plus loin dans ce document. Nasm a été utilisé pour compiler le code assembleur x86 en ELF 32-bits. Nasm produit des fichiers objets pouvant être liés à d'autres fichiers objets afin de créer un exécutable.

3.2.2 Rustup

Rust sera décrit plus en détails dans un prochain chapitre. Ce qu'il faut savoir est que Rust est distribué sous trois versions différentes. La version *stable*, la version *beta* et la version *nightly*. La version *nightly* possède plus de fonctionnalités mais sa stabilité n'est pas garantie. Cette version a été utilisée pendant le développement du projet et l'utilitaire Rustup a été utilisé pour son installation. Cet utilitaire permet de simplifier l'installation de Rust quand on souhaite une version différente de la dernière version stable de Rust.

3.2.3 Cargo et Xargo

Lors du développement d'un système d'exploitation type *bare metal*, on souhaite s'affranchir de toute dépendance à une librairie externe. Tout doit être refait depuis le début. Le code est donc compilé sans la bibliothèque standard (std). Rust a tout de même besoin d'une base pour être compilé. Cette base est fournie par la librairie **core**. Cette librairie est minimale et permet de ne définir que les primitives de Rust. Pour gérer les dépendances d'un projet Rust, il est conseillé d'utiliser le gestionnaire de paquets cargo. Le problème est que cargo ne permet pas de lier la librairie **core** à un projet. Heureusement, un autre utilitaire basé sur cargo existe et permet d'installer par défaut la librairie **core** pour des projets sans bibliothèque standard. Cet utilitaire se nomme xargo et est utilisé pour compiler le code Rust en fichiers objets

3.2.4 QEMU

Le compilateur GCC a été utilisé pour *linker* les fichiers objet générés par nasm et xargo. GCC génère un fichier au format ELF. Pour utiliser ce fichier comme un système d'exploitation *bootable*, il faut en faire une image ISO *bootable*. Pour se faire, l'utilitaire **genisoimage** est utilisé, couplé au *bootloader* GRUB. L'image ISO est finalement exécutée par la machine virtuelle QEMU. QEMU est une machine virtuelle pouvant émuler une architecture. Pour ce projet, l'architecture i386 a été choisie afin d'émuler un processeur Intel 32-bits.

3.3 Architecture

4 Rust

5 Exécution du *kernel*

5.1 Compilation

Quand on veut compiler un simple code C en utilisant GCC par exemple, le compilateur passe par plusieurs étapes. Le préprocesseur génère d'abord un fichier C en fonction des directives de préprocesseur. Ce fichier C est ensuite compilé en code assembleur qui est lui même compilé en code objet. Le *linker* permet ensuite de lier les différents fichiers objets et générer un exécutable. Nous avons déjà eu un aperçu des différentes étapes de la compilation d'un OS de type *bare metal* dans la partie 3.2. A la différence de la compilation d'un code C, nous avons d'un côté du code assembleur et de l'autre du code Rust. Nasm et cargo permettent tous deux de générer des fichiers objets. Il n'y a donc que la dernière étape à effectuer ce que GCC permet de faire avec la commande suivante.

```
gcc $(OBJS) -T $(LINKER) -static -m32 -ffreestanding -nostdlib -o $@ $(RUST)
```

Ici, `$(OBJS)` représente les fichiers objets générés par `nasm`, `$(LINKER)` est un fichier permettant de faire l'édition des liens et `$(RUST)` représente les fichiers objets générés par Rust.[1]

5.2 Linking

Nous avons vu dans la partie précédente que GCC a besoin d'un fichier pour faire l'édition des liens. Si ce fichier n'est pas donné, il en utilise un par défaut. Le *linker* permet de structurer le code par sections. Prenons pour exemple le *script* utilisé pour ce projet.

```
1 ENTRY(entrypoint)
2 SECTIONS {
3     . = 1M;
4     .boot ALIGN(4) :
5     {
6         *(.multiboot)
7     }
8     .stack ALIGN(4) :
9     {
10        *(.stack)
11    }
12    .text ALIGN(4K) :
13    {
14        *(.text*)
15    }
16    .rodata ALIGN(4K) :
17    {
18        *(.rodata*)
19    }
20    .data ALIGN(4K) :
21    {
22        *(.data*)
23    }
24    .bss ALIGN(4K) :
25    {
26        *(COMMON)
27        *(.bss*)
28    }
29 }
```

L'appel à `ENTRY` permet de spécifier l'entrée du *kernel*. Pour un simple programme en C l'entrée serait le *main*. Ici, ce sera l'entrée de notre *kernel* donc la première fonction exécutée au *boot*. `SECTION` va dire au linker où placer les parties du code. Par exemple, la section `.text` contiendra le code et la section `.data` contiendra les variables initialisées [1, 2, 3, 4]. Voici donc la structure du fichier ELF qui serait généré à l'aide de ce *script*.

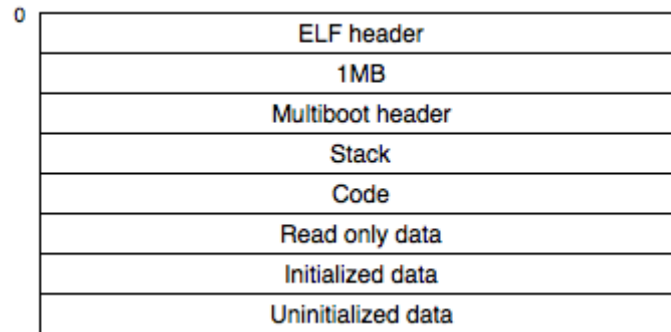


FIGURE 1 – Strucutre du fichier ELF

A noter que les sections commencent avec un *offset* de 1MB. Nous avons eu besoin de faire ça car les premiers 1MB dans un OS sont réservés [1, 5]. La mémoire vidéo (VRAM) se situe par exemple dans cette zone.

5.3 Boot

5.3.1 Principe général

Quand un ordinateur est allumé, un signal est envoyé à la carte mère qui démarre l'alimentation. Le processeur démarre alors en mode 16-bits. Le signal "Power Ok" est envoyé au BIOS qui est le *firmware* du PC (localisé en mémoire flash de la carte mère). Le BIOS initialise alors la séquence POST (*Power On Self Test*) qui vérifie que chaque périphérique est alimenté et que la mémoire est ok puis initialise chaque périphérique et enfin redonne la main au BIOS qui continue le *boot*. Le BIOS charge ensuite les 512 premiers bytes (MBR) du premier disque qui doit charger le *kernel* en mémoire et l'exécuter. Pour résumer, le *boot* d'une machine à base de BIOS se déroule de la manière ci-dessous.[1]

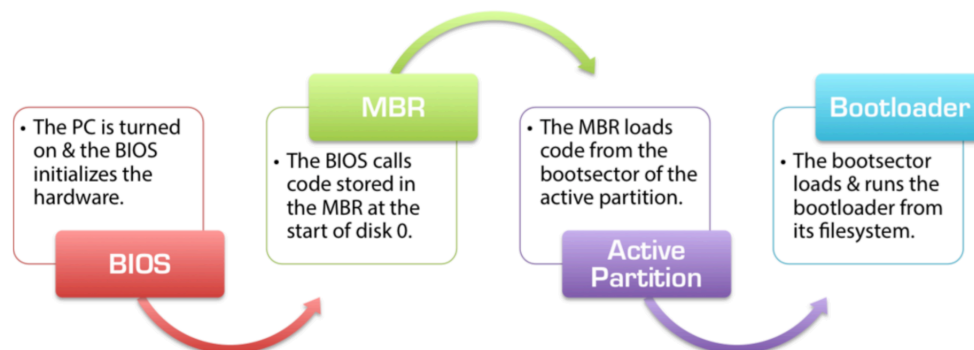


FIGURE 2 – Boot d'une machine à base de BIOS

5.3.2 GRUB

Le MBR contient ce qui est appelé le *bootloader*. Le *bootloader* est le morceau de code qui va charger le *kernel* en mémoire et l'exécuter. C'est ici qu'entre en scène GRUB. GRUB est un *bootloader* puissant et versatile permettant de charger n'importe quel type de système d'exploitation. Son initialisation se fait par étapes.

- *Stage 1* : Chargé en mémoire par le BIOS depuis le MBR, il contient le code pour charger le *Stage 1.5*
- *Stage 1.5* : Chargé en mémoire par le *Stage 1*, il contient les drivers nécessaires à l'accès au système de fichiers par le *Stage 2*
- *Stage 2* : Chargé en mémoire par le *Stage 1.5*, il affiche le menu de GRUB. Il permet de sélectionner et charger un OS

GRUB permet de charger n'importe quel type de système d'exploitation grâce au standard *Multiboot*. Ce standard permet à tout *bootloader* de charger tout OS compatible [1, 6].

5.3.3 Image ISO

Nous avons déjà pu voir que le *boot* du *kernel* se faisait à partir d'une image ISO dans la partie 3.2.4. Pour qu'une image ISO soit *bootable*, il est nécessaire que GRUB soit installé dans les huit premiers KB du disque. Prenons l'arborescence suivante :

```
isofiles
  boot
    grub
```

Les fichiers `kernel.elf` (kernel sur lequel nous voulons *booter*), `menu.lst` (fichier de configuration de GRUB) et `stage2_eltorito` doivent être copiés de manière à obtenir l'arborescence suivante :

```
isofiles
  boot
    grub
      menu.lst
      stage2_eltorito
    kernel.elf
```

Pour finir, il faut exécuter la commande :

```
genisoimage -R -b boot/grub/stage2_eltorito -input-charset utf8 -no-emul-boot \
-boot-info-table -o os.iso isofiles
```

Cette commande générera une image ISO *bootable* nommée `os.iso`[1].

6 Gestion mémoire

6.1 Introduction

Le système d'exploitation développé est exécuté sur une architecture IA-32 (Intel 32-bits) aussi appelée i386. La mémoire est donc adressée sur 32 bits. $2^{32} = 4Go$, on peut en déduire que la mémoire physique (RAM) a une taille totale de 4Go dans notre système d'exploitation. En réalité, la taille de la RAM dépend du *hardware*. Dans notre cas le matériel est émulé par QEMU. La taille de la mémoire physique de notre OS dépend de la configuration de l'émulateur. Ces 4Go sont donc virtuels. Lorsqu'une tâche est exécutée, elle est chargée en mémoire et est définie par la paire base et limite. La base est son adresse physique dans la RAM et la limite est sa taille. La figure 3 donne un exemple d'adressage de plusieurs processus [1].

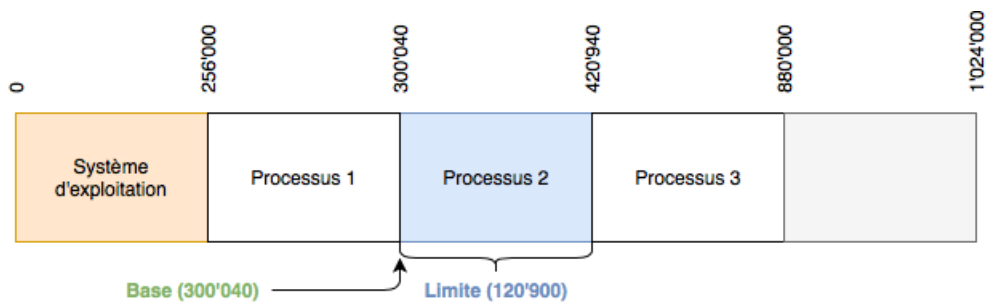


FIGURE 3 – Exemple d'adressage mémoire

Une tâche possède son propre espace d'adressage dit virtuel. Pour le processus 2 de la figure 3, l'adresse 0 est en fait à l'adresse physique 300040. Il y a donc besoin de traduire l'adresse virtuelle en adresse physique. C'est là qu'entre en jeu le MMU (Memory Management Unit). Le MMU est un dispositif matériel permettant de faire cette translation d'adresses. A chaque référencement mémoire, il va convertir l'adresse virtuelle en adresse physique et regarder si elle ne dépasse pas la limite du processus. Le MMU permet donc aussi de protéger la mémoire car il va empêcher toute référence à une zone extérieure au processus (voir figure 4) [1].

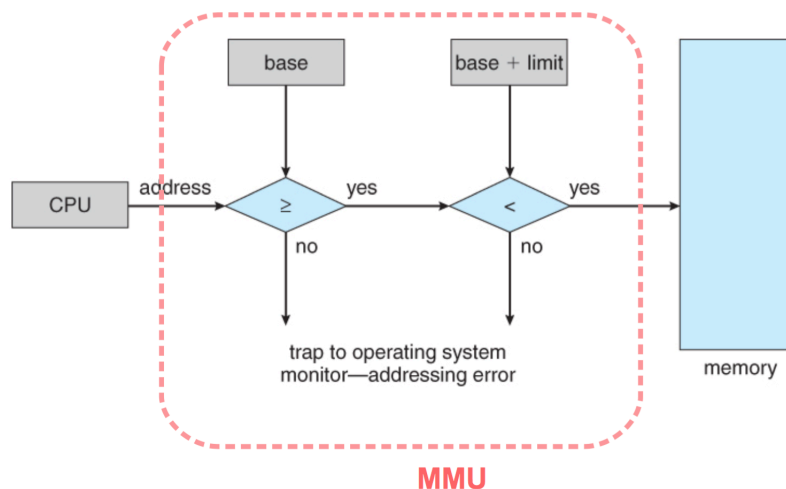


FIGURE 4 – Protection mémoire avec un MMU

Pour convertir une adresse virtuelle en adresse physique, le MMU passe par plusieurs étapes. Quand le *kernel* veut lire une donnée dans la mémoire, l'adresse de cette donnée est appelée adresse logique. Le MMU va commencer par convertir cette adresse en adresse linéaire. Une deuxième conversion est ensuite effectuée afin d'obtenir une adresse physique. Le MMU peut alors renvoyer la bonne donnée au *kernel*. Toutes ces étapes ne sont pas automatiques, le MMU utilise différentes techniques ayant besoin de certaines structures implémentées par le *kernel*. Ces techniques sont la segmentation et la pagination.

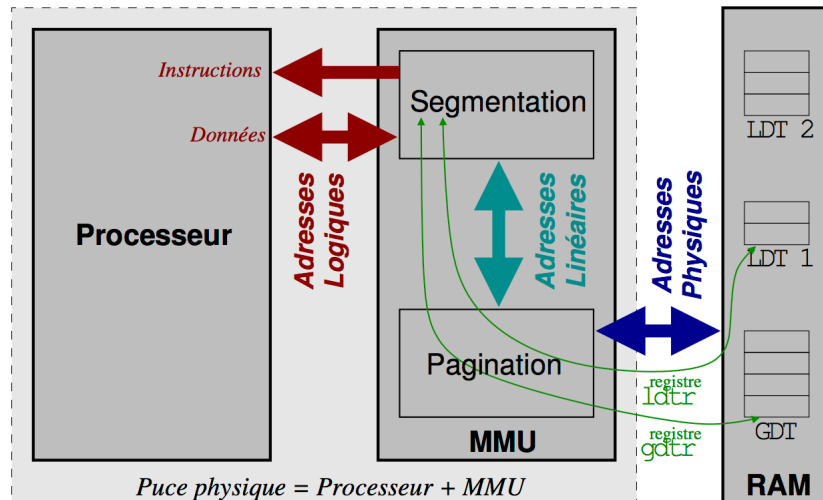


FIGURE 5 – Translation d'adresse

La segmentation est une technique permettant de découper la mémoire en segments de mémoire logique. Une adresse logique est convertie par le MMU en adresse linéaire en utilisant une table de descripteurs globale (GDT) ou locale (LDT). Si la pagination est activée, l'adresse linéaire est convertie en adresse physique. Toute cette mécanique est décrite dans la figure 5. A noter que la pagination n'est pas obligatoire et un OS pourrait s'en passer contrairement à la segmentation qui est indispensable en mode protégé (32-bits) [7].

6.2 Segmentation

6.2.1 Principe général

Comme expliqué précédemment, la segmentation est un mécanisme divisant l'espace d'adressage du processeur en espaces d'adressage plus petits appelés des segments. Un segment peut être utilisé pour contenir le code, les données ou la pile d'un processus étant exécuté par le processeur. Un segment peut aussi être utilisé pour contenir des structures de données tel qu'une LDT ou un TSS (structure contenant des informations à propos d'une tâche). Les segments d'un système d'exploitation sont contenus dans l'espace d'adressage linéaire. Pour lire l'octet d'un segment se trouvant dans l'espace d'adressage linéaire, le MMU utilise une adresse logique. L'adresse logique est composée d'un sélecteur de segment et d'un *offset*. Le sélecteur permet de trouver le bon segment dans la mémoire linéaire et l'*offset* permet de trouver l'octet dans ce segment. Dans le cas où la segmentation est utilisée seule (sans pagination), la mémoire linéaire est *mappée* directement dans la mémoire physique [8].

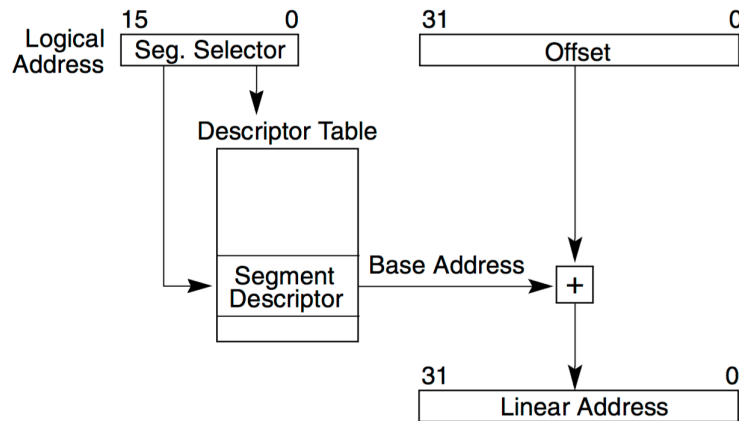


FIGURE 6 – Conversion d’une adresse logique en adresse linéaire

La figure 6 résume bien la conversion d’adresse logique en en adresse linéaire. On peut remarquer de plus que le sélecteur de segment passe par une table de descripteurs (GDT ou LDT) afin de trouver le bon segment dans la mémoire linéaire. En effet, Un sélecteur a une taille de 16 bits et contient l’index d’un descripteur dans une table, un bit indiquant si le descripteur est dans la GDT ou dans une LDT et enfin son niveau de privilège allant de 0 à 3 (figure 7) [1].

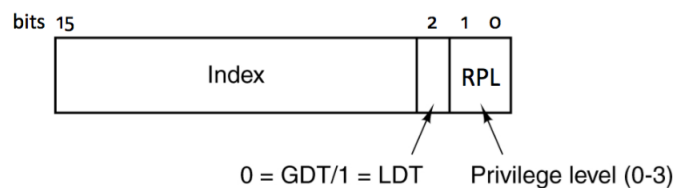


FIGURE 7 – Structure d’un sélecteur de segment

La gestion de la segmentation par le CPU se fait à l’aide de registres spéciaux nommés registres de segment. Ces registres sont au nombre de 6 et ont chacun une taille de 16 bits (même taille qu’un sélecteur de segment) [1, 9].

Registre	Segment
CS	<i>Code Segment</i>
DS	<i>Data Segment</i>
SS	<i>Stack Segment</i>
ES	<i>Extra Segment</i>
FS GS	<i>General Purpose Segments</i>

En mode protégé (32-bits), ces registres doivent pointer sur des descripteurs de segment de la GDT. Au minimum les trois premiers registres décrits doivent être utilisé en mode protégé (CS, DS, et SS). Les opérations adressant le code (décodage des instructions en mémoire, sauts, etc...) référencent le descripteur de segment sur lequel pointe le registre CS. Les opérations adressant les données (adressage de variables ou d’adresses mémoires) référencent le descripteur de segment sur lequel pointe le registre DS. Les opérations adressant la pile (**push** et **pop**) référencent le descripteur de segment sur lequel pointe le registre SS. Ces registres pointent sur des descripteurs de segments par l’intermédiaire de sélecteurs de segment.

6.2.2 GDT et LDT

Nous avons pu voir que pour traduire une adresse logique, des sélecteurs de segment sont utilisés. Ces sélecteurs pointent sur des entrées dans des tables de descripteurs. La GDT et la LDT sont deux types de table de descripteurs différents. La GDT est unique il ne peut y en avoir qu'une seule dans le système. Elle contient toutes les données utilisables en mode superviseur (*ring* 0 ou niveau de privilège 0). Une LDT est contenue dans la GDT. Elle peut être utilisée pour contenir le code et les données d'une tâche utilisateur par exemple. Dans le cas où plusieurs tâches sont exécutées en même temps, une LDT peut être créée par tâche ce qui permet en plus d'isoler le code de chaque processus. Une table de descripteurs est composée, comme son nom l'indique, de descripteurs. Chaque descripteur décrit une zone mémoire qui est défini par sa base (son adresse physique), sa limite (sa taille) et un niveau de privilèges (allant de 0 à 3, le niveau 0 ayant le plus de privilèges et le niveau 3 le moins). Ci dessous, la figure 8 montre un exemple d'une GDT [1]. Etant donné que les descripteurs de la GDT ont la même structure que les descripteurs de la LDT nous allons nous concentrer sur la GDT. De plus, aucune LDT n'est utilisé dans la version actuelle de l'OS.

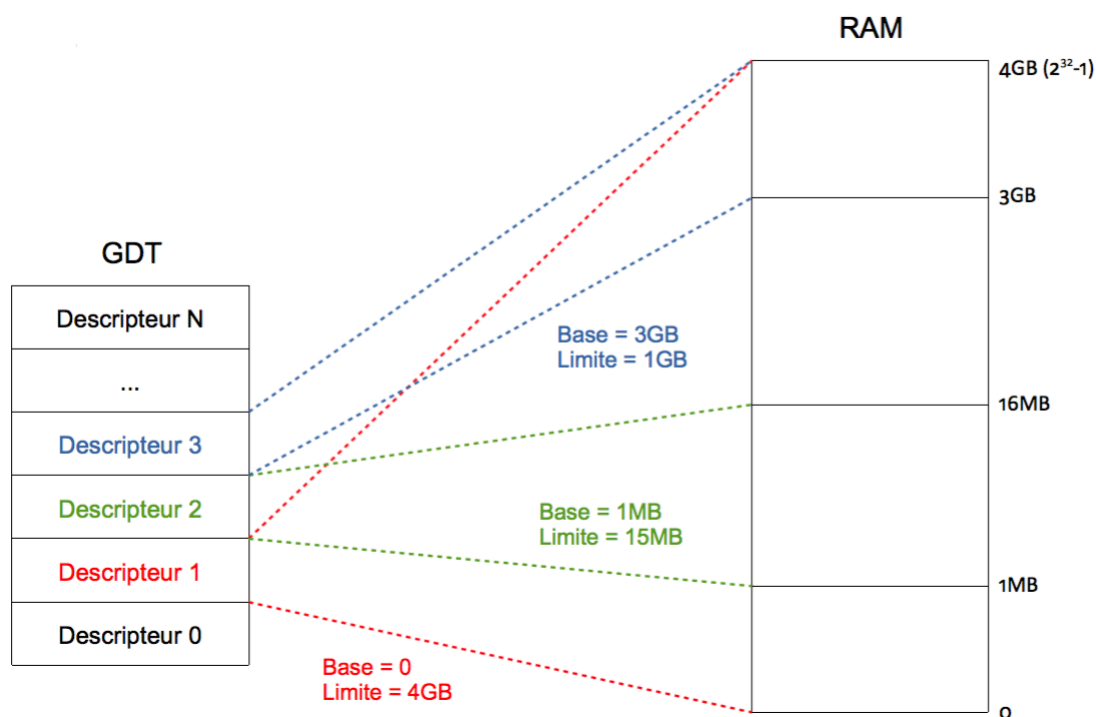


FIGURE 8 – Exemple d'une GDT

La GDT est contenue en mémoire. Cette dernière doit être initialisée par le *kernel* avant d'être chargée et utilisée par le MMU. Pour initialiser la GDT il faut construire ses entrées. Chaque entrée (ou descripteur) de la table de descripteurs est sur 64 bit et décrit une zone mémoire. L'adresse de cette zone mémoire est sur 32 bits et sa taille est sur 20 bits. Les bits restant sont des bits de contrôle pour l'accès aux données par exemple (niveau de privilèges, droit d'écriture/lecture, ...). Voir la figure 9 pour plus de détails [1, 10].

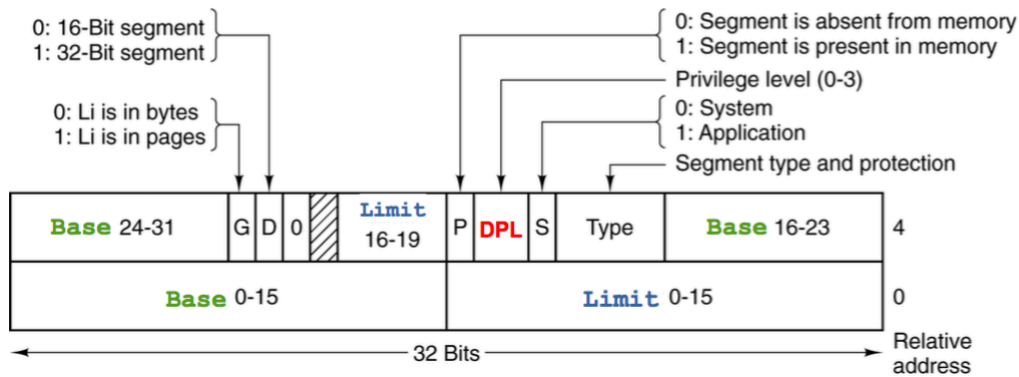


FIGURE 9 – Structure d’une entrée dans la GDT

L’OS développé a un adressage segmenté de type *flat*, c’est-à-dire que toute la mémoire est accédée de manière linéaire. Ce modèle de segmentation est le plus simple car il permet d’ignorer le mécanisme de segmentation car l’intégralité de la zone mémoire devient disponible. Dans un modèle de type *flat*, les segments de code et de données se chevauchent sur l’intégralité de la mémoire disponible. Ceci se fait en initialisant trois descripteurs dans la GDT. Un descripteur nul à l’index 0 (obligatoire dans tous les modèles de segmentation), un segment de code couvrant toute la mémoire et un segment de données couvrant aussi toute la mémoire. Les segments de code et de données adressent ainsi les mêmes zones mémoire. On verra par la suite que d’autres entrées ont été ajoutées à la GDT pour la gestion des tâches.

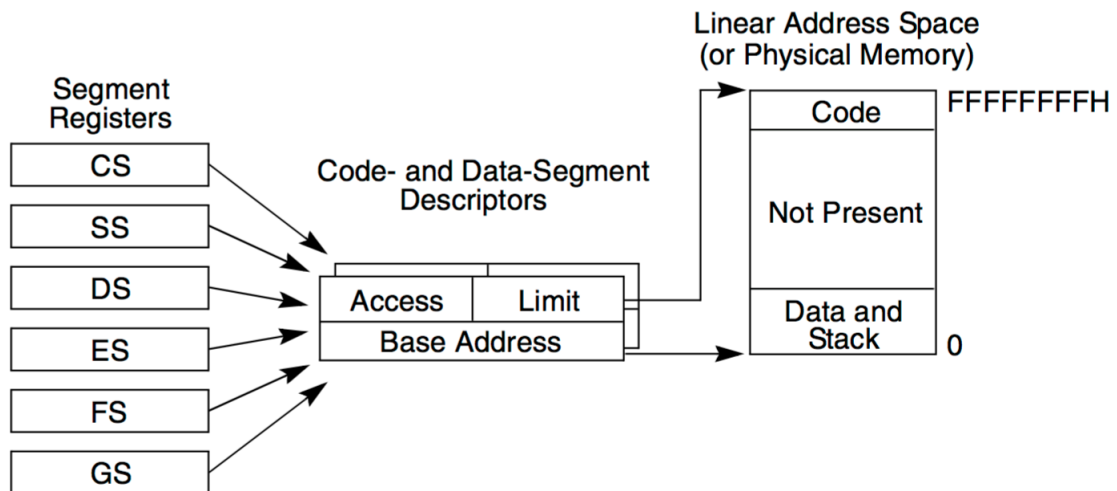


FIGURE 10 – Modèle de segmentation de type *flat*

Pour obtenir un segment sur toute la mémoire disponible, il faut mettre le bit de granularité à 1 pour avoir une limite en blocs de 4Ko. Ensuite, la limite doit être mise à la valeur 0xFFFFF ce qui donne une limite réelle de $0x100000 \times 0x1000$ soit 4Go. Le niveau de privilège doit être laissé à 0 (*ring 0*) si on veut créer un segment pour le *kernel* ou bien être mis à 3 (*ring 3*) si on veut créer un segment pour le mode utilisateur. Ci-dessous un exemple de code permettant de construire un segment de code et un segment de données.

```

fn new(base: u32, limit: u32, type: u8, s: u8, d: u8, g: u8, dpl: u8) -> GdtEntry;

pub fn make_code_segment(base: u32, limit: u32, dpl: u8) -> GdtEntry {
    GdtEntry::new(base, limit, 0xB, 0x1, 0x1, 0x1, dpl)
}

pub fn make_data_segment(base: u32, limit: u32, dpl: u8) -> GdtEntry {
    GdtEntry::new(base, limit, 0x3, 0x1, 0x1, 0x1, dpl)
}

```

Ici, `new` est le prototype d'une méthode pour une structure 64 bits représentant une entrée dans la GDT. le bit `s` est mis à 1 car on construit des segments de code et de données [8]. Le bit `d` est mis à 1 car on veut un segment de 32 bits. Le bit `g` est mis à 1 pour avoir une granularité de 4Ko. L'octet `0xB` correspond au type `CODE_EXEC_READ` et `0x3` correspond au type `DATA_READ_WRITE` [1]. Une fois la GDT construite, il faut dans un premier temps utiliser l'instruction `lgdt` pour la charger dans le registre GDTR. Ce registre est utilisé par le processeur pour faire le lien entre le MMU et la GDT créée [8]. L'adresse du descripteur de la GDT doit donc être donnée en argument à l'instruction `lgdt`. Le descripteur de GDT est défini par la structure 48-bits décrite dans la figure 11 [10].

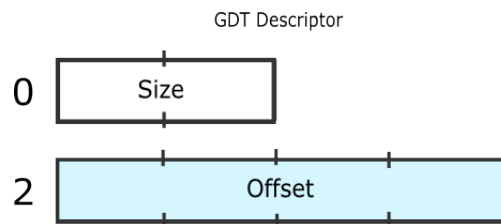


FIGURE 11 – Descripteur de GDT

Dans un descripteur de GDT *Size* est la limite sur 16 bits (c'est à dire la taille de la GDT - 1) et *Offset* est l'adresse physique de la GDT sur 32 bits. Dans le cas de notre OS, la GDT est déclarée statiquement dans le *kernel*. L'adresse de cette variable statique est utilisée pour construire un descripteur qui sera chargé dans le registre GTDR. Après avoir chargé la GDT avec l'instruction `lgdt`, il faut faire pointer les registres de segment sur les segments de la table de descripteurs à l'aide de sélecteurs. Pour rappel, un sélecteur est sur 16 bits et contient l'index d'un segment dans la GDT précédemment chargée. Pour récupérer l'index d'un segment dans la GDT à partir de l'index de son descripteur, il faut faire un décalage à gauche de 3 bits (voir figure 7). Prenons un descripteur se situant à l'index 2 de la GDT. Si on veut initialiser le segment de code (registre CS) avec ce descripteur, il faut mettre la valeur 16 dans le registre CS ($2 \ll 3 = 16$). Dans le cas où on veut définir ce segment avec un autre niveau de privilèges ou bien pour une LDT, il suffit de mettre les bons bits à 1 après avoir fait le décalage.

6.3 Pagination

6.3.1 Principe général

La pagination est une autres technique de gestion de mémoire qui diffère de la segmentation. Alors que la segmentation permet d'allouer des morceaux de mémoire de taille variable, la pagination divise la mémoire en blocs de taille fixe appelés pages (de 4Ko, 2Mo ou 4Mo). De plus, la segmentation est obligatoire dans une architecture i386 alors que la pagination ne l'est pas [11]. Quand une tâche fait référence à une adresse logique en mémoire, cette adresse est convertie en adresse linéaire grâce au mécanisme de segmentation et c'est le mécanisme de pagination qui permet de traduire cette adresse linéaire en adresse physique (comme vu précédemment dans la partie sur la segmentation). Quand la pagination est activée, l'adresse linéaire est divisée en deux parties lorsque des pages de 4Mo sont utilisées et en trois parties lorsque des pages de 4Ko sont utilisées. Le *kernel* développé utilise des pages de 4Ko, une adresse linéaire est donc sous la forme suivante :

- 10 bits pour le *directory index*
- 10 bits pour le *page index*
- 12 bits pour l'*offset*

On dit que cette pagination est une pagination à trois niveaux. En général, une pagination à trois niveaux est utilisée mais il peut exister des systèmes utilisant plus ou moins de niveaux. Le système d'exploitation doit créer un répertoire de pages (*Page Directory*) et au moins une table des pages (*Page Table*) pour chaque tâche. Les répertoires et les tables des pages ont la taille d'une page et sont composés d'entrées sur 32 bits (4 octets). Une entrée dans un répertoire permet d'adresser une table de pages et une entrée dans une table permet d'adresser une page. Dans notre cas, un répertoire permet donc d'adresser 1024 tables et une table 1024 pages ce qui permet bien d'adresser au total 4Go ($1024 \times 1024 \times 4096$). Une entrée est sur 32 bits mais seulement les 20 bits de poids fort sont utilisés pour l'adressage car les adresses sont alignées avec 4096 ce qui laisse les 12 bits de poids faible pour la configuration [12].

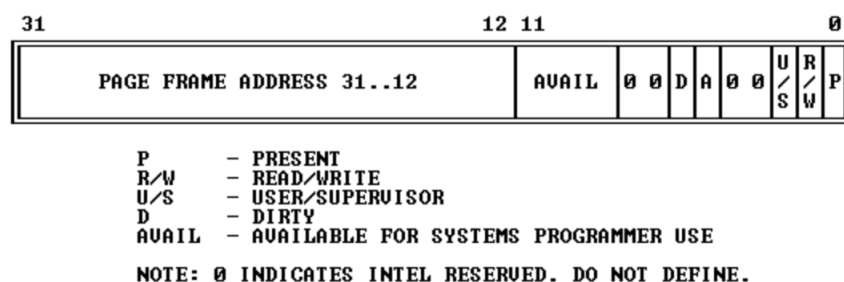


FIGURE 12 – Structure d'une *Page Entry*

Quand une adresse linéaire est lue, le *directory index* permet de lire la bonne entrée dans le *Page Directory*. Il faut ensuite utiliser le *page index* pour récupérer la bonne entrée dans la table des pages. De la même manière que l'entrée dans le répertoire de pages pointait sur une table des pages, l'entrée dans une table des pages pointe sur une *Page Frame*. Cette page contient finalement la donnée pointée par l'adresse linéaire, il faut utiliser l'*offset* pour trouver cette donnée dans la page. La figure 13 résume bien ce mécanisme [8]. A noter que le *Page Directory* est pointé par le registre CR3. A chaque fois qu'un changement de tâche a lieu, le registre CR3 doit être mis à jour avec le *Page Directory* de la nouvelle tâche [13].

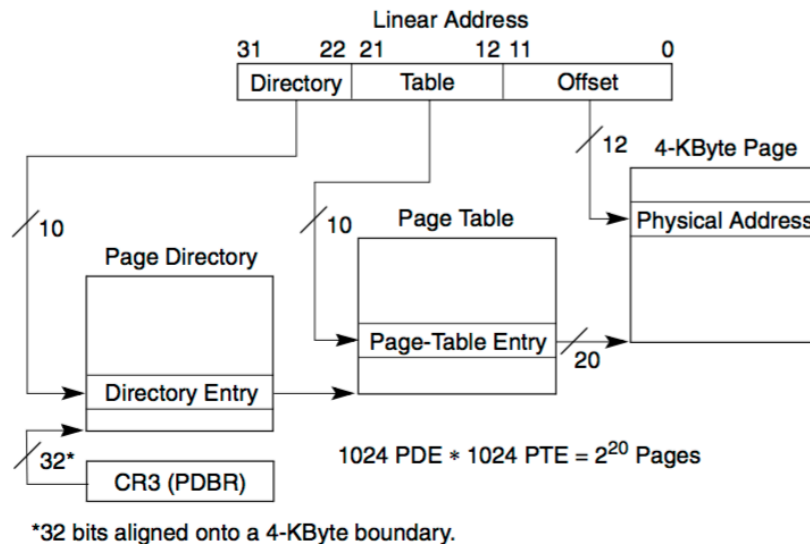


FIGURE 13 – Exemple de pagination à 3 niveaux

6.3.2 Activer la Pagination

Pour initialiser la pagination sur architecture x86, il faut d'abord construire un répertoire de pages valide contenant les entrées vers les pages du *kernel*. Il est obligatoire de commencer par cela car si la pagination est activée et que le *kernel* n'est pas *mappé* dans le répertoire chargé, une exception sera levée (*Page Fault*). Par soucis de simplicité pour la suite du développement de l'OS, le *kernel* va être déplacé au dernier Go de la RAM. Grâce à la pagination, ceci peut se faire assez simplement, il suffit de compléter le répertoire de pages ainsi que ses tables de pages correctement. Pour rappel, le *kernel* commence à l'adresse 0x100000 (1Mo) mais il faut aussi rendre accessible le premier Mo de RAM. Il faut donc déplacer les adresses physiques allant de 0x0 à la fin du *kernel* (qui n'est pas fixe). Dans un premier temps, le *linker* doit être modifié de cette manière :

```

1  SECTIONS {
2      /* Low memory Kernel */
3      . = 0x00100000;
4      .boot ALIGN(4) : { *(.multiboot) }
5      .low_text ALIGN(4K) : { *(.low_text) }
6      .low_data ALIGN(4K) : { *(.low_data) }
7      .low_bss ALIGN(4K) : { *(.low_bss) }
8      /* Higher-half Kernel */
9      . += 0xC0000000;
10     .stack ALIGN(4) : AT(ADDR(.stack) - 0xC0000000) { *(.stack) }
11     .text ALIGN(4K) : AT(ADDR(.text) - 0xC0000000) { *(.text*) }
12     .rodata ALIGN(4K) : AT(ADDR(.rodata) - 0xC0000000) { *(.rodata*) }
13     .data ALIGN(4K) : AT(ADDR(.data) - 0xC0000000) { *(.data*) }
14     .bss ALIGN(4K) : AT(ADDR(.bss) - 0xC0000000) { *(COMMON) *(.bss*) }
15 }

```

Ici, le *kernel* est divisé en deux parties. La première est celle qui va être appelée au démarrage du système et qui va initialiser la pagination. Une fois la pagination active, le *kernel* va continuer son exécution dans la deuxième partie qui est située dans le dernier Go de RAM. Nous sommes obligés de démarrer le *kernel* au début de la mémoire physique car toutes les adresses sont virtuelles. En réalité, le *kernel* dispose de beaucoup moins (variable

selon la configuration de l'émulateur, ici QEMU). Il n'existe donc pas d'adresse physique située à 3Go dans la mémoire physique du *kernel* et il est donc impossible de démarrer le système à cette adresse. Regardons plus en détail de quelle manière la première partie du *kernel* initialise la pagination. Comme dit précédemment, un répertoire de pages initial doit être construit. Etant donné que nous allons exécuter du code dans le premier Go et aussi dans le dernier, le *kernel* doit être *mappé* dans ces deux zones mémoire en même temps. La première partie va être adressée linéairement, ce qui veut dire que l'adresse physique 0x0 correspondra à l'adresse virtuelle 0x0 et ainsi de suite jusqu'à la fin du *kernel*. Cet adressage donne le répertoire de pages schématisé dans la figure 14.

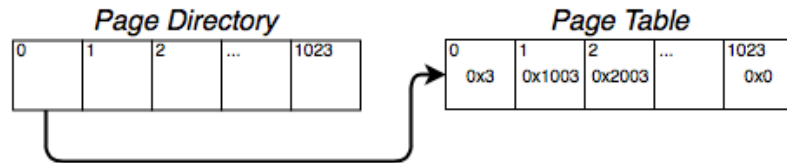


FIGURE 14 – Répertoire de pages adressant le *kernel* au début de la RAM

On peut voir ici que la première entrée du répertoire de pages pointe sur une table de pages adressant le début de la RAM. Chaque entrée est incrémentée de 4096 (0x1000 en hexadécimal) car une page fait 4096 octets. De plus, les deux premiers bits de poids faible de chaque page sont à 1 pour indiquer que la page est active et que l'on peut écrire et lire dedans (voir figure 12). L'entrée dans le répertoire de page correspondant au dernier Go (soit 0xC0000000 en hexadécimal) doit pointer sur une table des pages identique. Pour trouver une entrée dans le répertoire de pages depuis une adresse il faut faire un décalage à droite de 22 bits sur cette adresse (ce qui est équivalent à diviser par 4096, soit la taille d'une page, puis de nouveau diviser par 1024, soit le nombre de pages adressées par une table). Ici, $0xC0000000 \gg 22 = 0x300$ (768 en décimal). Il faut donc faire pointer l'entrée 768 du répertoire de pages à une table des pages identique à celle pointée par l'entrée 0 ce qui donne finalement le répertoire suivant.

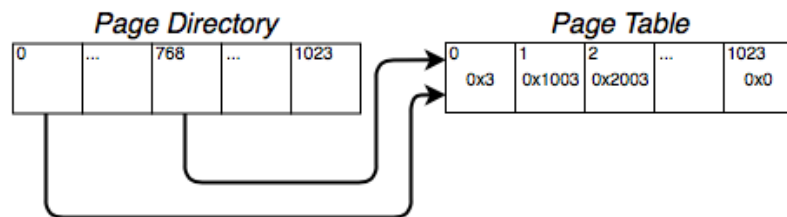


FIGURE 15 – Répertoire de pages adressant le *kernel* à la fin de la RAM

Une fois le répertoire de pages initialisé de cette manière, il ne reste plus qu'à faire pointer le registre CR3 dessus et activer la pagination en mettant le bit 31 du registre CR0 à 1. Le code peut ensuite sauter à la partie haute de la RAM où nous avons déplacé le *kernel*. A partir de là, tout le code qui sera exécuté sera dans le dernier Go de RAM, il n'y a donc plus besoin de faire pointer la première entrée du répertoire de pages sur le table des pages du *kernel* ce qui peut être fait en écrivant 0 dans cette entrée. Le code rust peut finalement être appelé avec la pagination active.

6.4 Allocation dynamique

Le dernier élément de gestion mémoire implémenté dans RustOS est l'allocation de mémoire dynamique. L'allocation dynamique consiste à réserver des blocs de mémoire pendant l'exécution du *kernel* ou bien d'un programme utilisateur. Jusqu'à maintenant, toutes les structures utilisées étaient déclarées statiquement se retrouvant donc dans la zone mémoire du *kernel*, plus précisément dans le segment bss (voir 5.2). Déclarer des variables statiquement est pratique mais fait augmenter la taille du *kernel* et ce n'est pas une solution viable pour allouer de plus grandes régions mémoire comme par exemple du code utilisateur qui peut faire plusieurs kilooctets. Pour implémenter l'allocation de mémoire dynamique, il faut dans un premier temps définir quelle zone mémoire peut être utilisée dans ce but. Actuellement, la totalité du code et des données est contenu dans le *kernel*. Des zones mémoire peuvent donc commencer à être allouées à la fin de ce dernier. Le *linker* a été légèrement modifié afin d'obtenir l'adresse de la fin du *kernel*. Ceci peut se faire en ajoutant une expression au *linker* et en la rendant accessible depuis le code assembleur.

Modification apportée au *linker* :

```
1 . += 0xC0000000;  
2 ...  
3 kernel_end = .;
```

Code assembleur permettant de rendre accessible l'expression `kernel_end` :

```
1 extern kernel_end  
2 get_kernel_end:  
3     mov eax, kernel_end  
4     ret
```

A partir de là, on peut définir la zone d'allocation mémoire aussi appelée tas (ou *heap* en anglais). Dans notre OS, le tas commence à la fin du *kernel* alignée avec la taille d'une page (4096 octets). Ce choix a été fait car le *kernel* n'aura besoin d'allouer que des nouvelles pages. La fin du tas dépend de la fin de la mémoire physique qui dépend de la configuration de QEMU. Quand le *kernel* aura besoin d'allouer une nouvelle page, il ira chercher le prochain bloc libre dans le tas situé donc entre la fin du *kernel* et la fin de la mémoire physique. La recherche de bloc libre se complexifie rapidement si des blocs sont libérés. De nombreuses méthodes sont possibles pour la recherche de bloc libre et la gestion des blocs libérés. Le *kernel* développé utilise une liste doublement chaînée pour gérer la mémoire dynamique. Chaque bloc mémoire alloué est précédé d'un entête contenant l'adresse du bloc précédent sur 32 bits, l'adresse du bloc suivant sur 32 bits, sa taille en octets sur 32 bits et un booléen indiquant si le bloc est libre ou non sur 8 bits. De plus, l'entête est aligné sur 16 octets.

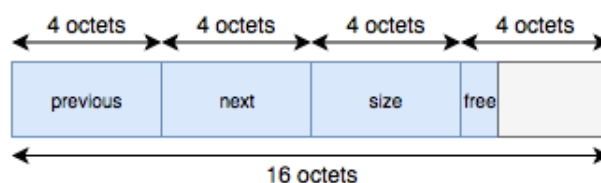


FIGURE 16 – Entête d'un bloc de mémoire dans le tas

Le tas ainsi construit, avec chaque entête lié à ses voisins par des pointeurs, permet de rechercher aisément un bloc libre. Un nouvel entête est créé quand le dernier bloc de la liste chaînée est alloué. Un algorithme a aussi été implémenter pour la gestion de mémoire libérée. Si un bloc est libéré au milieu de blocs alloués ce blocs est simplement marqué comme libre (en utilisant le booléen *free* de l'entête). Si deux blocs libres sont contiguës, ils sont fusionnés pour n'en former qu'un seul. Dans le *kernel*, la fonction pour l'allocation est **kmalloc** et la fonction pour la libération est **kfree**. La fonction **kmalloc** va allouer de nouvelles pages et tables de pages automatiquement s'il y a besoin (comme le montre la figure 17). De la même manière, la fonction **kfree** va libérer les pages et les tables de pages automatiquement. Ces deux fonctions permettent donc beaucoup d'abstraction au niveau de la pagination mais aussi de l'allocation car l'utilisation des entêtes est complètement transparent. La fonction **kmalloc** va simplement renvoyer une adresse sur 32 bits et **kfree** prend comme argument cette adresse pour libérer la mémoire.

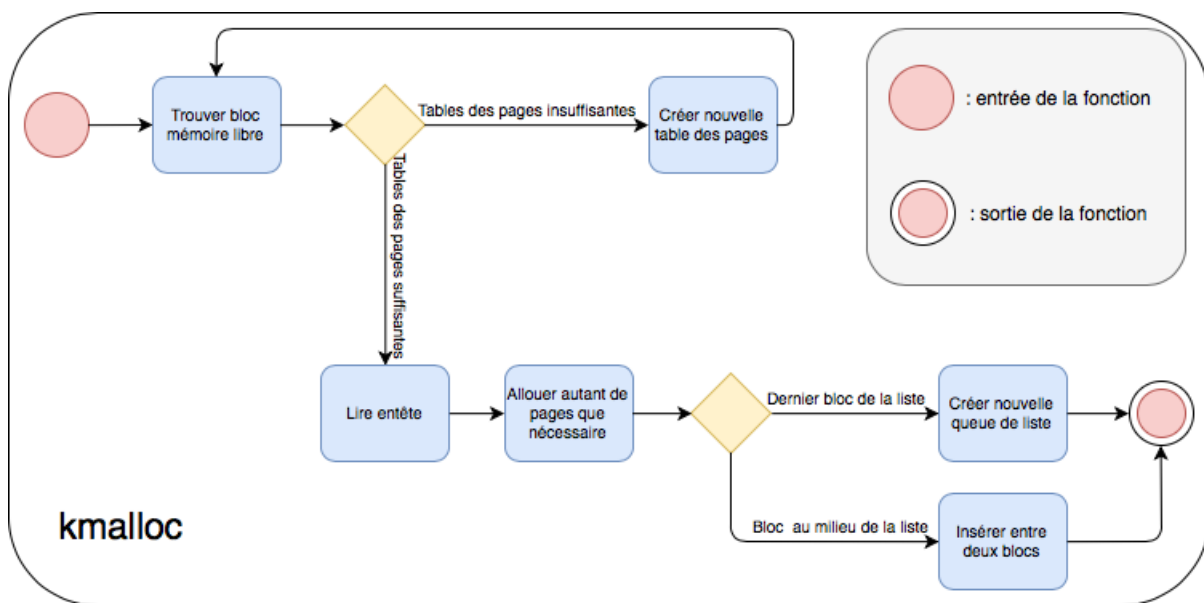


FIGURE 17 – Algorithme utilisé pour l'allocation dynamique dans le *kernel*

Voyons maintenant un exemple d'allocation et de libération mémoire utilisant les deux algorithmes décrits ci-dessus. Dans cet exemple, nous allons allouer plusieurs blocs mémoire puis les libérer afin de voir plus en détail le comportement de ces fonctions. Nous allons supposé que la fin du *kernel* se situe à l'adresse 0x3FC000 et que la fin de la RAM se situe à l'adresse 0x1000000. A l'initialisation du *kernel*, un premier entête est créé au début de la zone d'allocation dynamique. Ce premier entête est donc situé à l'adresse 0x3FC000 et a pour taille $0x1000000 - 0x3FC000 = 0xC04000$. C'est cette information qui permet de trouver la premier entrée de la liste et par la suite de la parcourir.

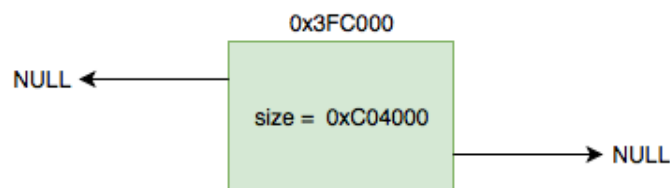


FIGURE 18 – Etat initial de la chaîne d'entêtes

Si le *kernel* a besoin d'allouer une nouvelle page maintenant, la fonction 17 va être appelée qui va parcourir les blocs libre. Dans ce cas précis, l'intégralité du tas est disponible donc la fonction va simplement créer un nouveau bloc. Dans la figure 19, les blocs libre sont en vert et les blocs alloués sont en rouge. A noter que que les blocs aloués sont alignés avec la taille d'une page (4096Ko ou 0x1000 en hexadécimal). Etant donné qu'il faut compté l'entête de 16 octets, il faudra toujours allouer une page de plus.

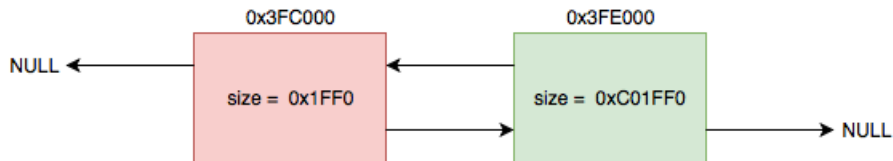


FIGURE 19 – Allocation d'une page

Si le *kernel* a de nouveau besoin d'une page, on va se retrouver dans la situation où il n'y a plus assez de place dans la table des pages. Pour rappel, une table des pages a une taille de 4Mo (0x400000 en hexadécimal). Comme expliqué avec la figure 17, la fonction d'allocation va se charge toute seule de créer une nouvelle table des pages et de l'ajouter au répertoire de pages. La page demandée par le *kernel* sera ensuite allouée. On obtient donc la liste de la figure 20, avec deux blocs alloués au lieu d'un seul. Le bloc à l'adresse 0x3FE000 contient alors la table des pages sur laquelle pointe la deuxième entrée du répertoire des pages. La page demandée par le *kernel* sera finalement à l'adresse 0x400000.

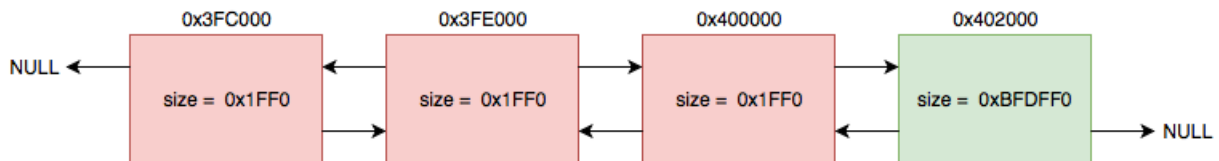


FIGURE 20 – Allocation d'une page et d'une table des pages

Libérer la page à l'adresse 0x400000 aura pour conséquence directe de libérer aussi la table des pages à nouvellement créée. La fonction ?? libère automatiquement les tables des pages vides.

7 Périphériques

7.1 Ports

Un processeur IA-32 a la possibilité de transférer des données en utilisant les ports d'entrée/sortie. Ces ports sont utilisés par le processeurs pour communiquer avec des périphériques. Ils peuvent être utilisés pour envoyer et recevoir des données (par exemple un *timer* va utiliser les ports d'entrée/sortie pour envoyer son état). Les ports peuvent aussi être utilisés pour contrôler un périphérique à partir de registres de contrôle (par exemple avec un contrôleur de disque) [14]. Étant donné que nous ne sommes pas sur du vrai *hardware*, QEMU va se charger d'émuler les différents périphériques utilisés par un processeur Intel 32-bits.

Les ports d'entrées/sorties sur architecture x86 se situent dans un espace d'adresses séparé de la mémoire physique. Cet espace permet d'adresser 64000 (soit 2^{16}) ports de 8 bits. Les ports sont donc adressés sur 16 bits mais il n'est pas possible d'écrire dans un PIO de la même manière que l'on écrirait dans la mémoire (avec une instruction `MOV`) car nous sommes dans deux espaces d'adresses différents. Ainsi, le CPU utilise des instructions spéciales pour accéder aux PIO. Ces instructions sont les instructions `IN` et `OUT`. `IN` permet de lire tandis que `OUT` permet d'écrire. À noter que l'adresse du port doit toujours être spécifiée dans le registre `dx` et la lecture et l'écriture se font toujours avec les registres `ax/al` [1].

Exemple de lecture et d'écriture dans un port d'entrée/sortie :

Écrire 4 dans le port 0x2A :

```
mov dx, 0x2A
mov al, 4
out dx, al
```

Lire un octet depuis le port 0x2A :

```
mov dx, 0x2A
in byte al, dx
```

Il existe une autre méthode pour écrire dans les ports utilisant le même bus d'adresse pour la mémoire physique et pour les périphériques. Cette méthode consiste à *mapper* les ports d'entrées/sorties dans la mémoire physique (MMIO). En écrivant dans la zone réservée aux ports, on écrirait alors directement dans les ports et pas dans la mémoire physique. Le *kernel* développé utilise la première méthode (PIO).

7.2 Interruptions et Exceptions

7.2.1 Principe général

Les interruptions et les exceptions sont des événements qui indiquent que l'attention du processeur est demandée quelque part soit dans le code, soit par un périphérique. Il existe deux types d'interruptions, les interruptions logicielles et les interruptions matérielles. Les exceptions sont générées par le processeur mais diffèrent des interruptions logicielles. Une interruption peut arriver à n'importe quel moment en réponse au signal d'un périphérique ou bien si le processeur le demande avec l'instruction `INT` (interruption logicielle). Une exception est levée lorsque le processeur détecte une erreur à l'exécution d'une instruction (par exemple une division par 0). Quand une interruption ou une exception a lieu, une routine logicielle est appelée (ISR). Les processeurs IA-32 supportent jusqu'à 256 interruptions dont les 32 premières sont réservées aux exceptions (voir figure 21) [1, 8].

Vector No.	Mnemonic	Description	Type	Error Code	Source
0	#DE	Divide Error	Fault	No	DIV and IDIV instructions.
1	#DB	RESERVED	Fault/Trap	No	For Intel use only.
2	—	NMI Interrupt	Interrupt	No	Nonmaskable external interrupt.
3	#BP	Breakpoint	Trap	No	INT 3 instruction.
4	#OF	Overflow	Trap	No	INTO instruction.
5	#BR	BOUND Range Exceeded	Fault	No	BOUND instruction.
6	#UD	Invalid Opcode (Undefined Opcode)	Fault	No	UD2 instruction or reserved opcode. ¹
7	#NM	Device Not Available (No Math Coprocessor)	Fault	No	Floating-point or WAIT/FWAIT instruction.
8	#DF	Double Fault	Abort	Yes (Zero)	Any instruction that can generate an exception, an NMI, or an INTR.
9		Coprocessor Segment Overrun (reserved)	Fault	No	Floating-point instruction. ²
10	#TS	Invalid TSS	Fault	Yes	Task switch or TSS access.
11	#NP	Segment Not Present	Fault	Yes	Loading segment registers or accessing system segments.
12	#SS	Stack-Segment Fault	Fault	Yes	Stack operations and SS register loads.
13	#GP	General Protection	Fault	Yes	Any memory reference and other protection checks.
14	#PF	Page Fault	Fault	Yes	Any memory reference.
15	—	(Intel reserved. Do not use.)		No	
16	#MF	x87 FPU Floating-Point Error (Math Fault)	Fault	No	x87 FPU floating-point or WAIT/FWAIT instruction.
17	#AC	Alignment Check	Fault	Yes (Zero)	Any data reference in memory. ³
18	#MC	Machine Check	Abort	No	Error codes (if any) and source are model dependent. ⁴
19	#XF	SIMD Floating-Point Exception	Fault	No	SSE and SSE2 floating-point instructions ⁵
20-31	—	Intel reserved. Do not use.			
32-255	—	User Defined (Non-reserved) Interrupts	Interrupt		External interrupt or INT <i>n</i> instruction.

NOTES:

1. The UD2 instruction was introduced in the Pentium Pro processor.
2. IA-32 processors after the Intel386 processor do not generate this exception.
3. This exception was introduced in the Intel486 processor.
4. This exception was introduced in the Pentium processor and enhanced in the P6 family processors.
5. This exception was introduced in the Pentium III processor.

FIGURE 21 – Table des interruptions et exceptions sur IA-32

Comme vu précédemment, une interruption logicielle peut être exécutée par le processeur avec l'instruction INT. L'instruction INT suivie du numéro d'interruption sur 8 bits déclenchera l'interruption en question. Par exemple, l'instruction INT 0x30 déclenchera l'interruption 48. Au moment de l'appel à l'instruction INT, le pointeur d'instruction va sauter à l'adresse du code contenant la routine d'interruption correspondant au numéro d'interruption logicielle spécifiée. C'est la table des descripteurs d'interruption (IDT) qui permet de définir l'adresse du code à exécuter pour chaque numéro d'interruption (que ce soit une interruption logicielle, matérielle ou une exception). A noter aussi que les interruptions logicielles sont synchrone étant donné qu'elles sont exécutées par le processeur, contrairement aux interruptions matérielles qui sont asynchrones (exécutées par les périphériques, elle peuvent arriver à n'importe quel moment).

Nous avons vu que les interruptions matérielles étaient générées par le *hardware*. Il existe deux types d'interruptions matérielles, les NMI (*Non Maskable Interrupt*) et les IRQ (Interrupt Request). Une NMI indique qu'un problème est survenu au niveau matériel (mémoire défectueuse, erreur de bus, ...). Comme son nom l'indique, une NMI ne peut pas être ignorée (ou masquée), l'interruption doit donc dans tous les cas être servie. Le but ici est d'arrêter le processeur afin d'éviter toute perte de données [1]. Une IRQ quant à elle peut être masquée. L'instruction CLI permet de masquer les interruptions et l'instruction STI permet de les démasquer. En général, un périphérique génère une IRQ lorsque des données sont prêtes à être lues, qu'une commande est terminée ou qu'un événement particulier a lieu (par exemple la pression d'une touche du clavier ou l'écriture de données sur le disque). Quand une interruption est générée, l'ISR correspondant à l'IRQ doit être appelée. C'est là qu'entre en jeu le contrôleur d'interruption (PIC). Le PIC va faire correspondre une IRQ à un numéro d'interruption (voir figure 22). A la manière des interruptions logicielles l'IDT va être utilisée pour appeler la bonne routine d'interruption. Le PIC permet donc de faire le lien entre le matériel et le logiciel.

IRQ	Description	Interruption
0	System timer (PIT)	0x08
1	Keyboard	0x09
2	Redirected to slave PIC	0x0A
3	Serial port (COM2/COM4)	0x0B
4	Serial port (COM1/COM3)	0x0C
5	Sound card	0x0D
6	Floppy disk controller	0x0E
7	Parallel port	0x0F
8	Real time clock	0x70
9	Redirected to IRQ2	0x71
10	Reserved	0x72
11	Reserved	0x73
12	PS/2 mouse	0x74
13	Math coprocessor	0x75
14	Hard disk controller	0x76
15	Reserved	0x77

FIGURE 22 – Table de correspondance des IRQs

En comparant la figure 21 avec la figure 22, on constate que certaines IRQs partagent le même numéro d'interruption que des exceptions. L'interruption du *timer* par exemple a le même numéro d'interruption que l'exception *Double Fault* (0x8). Si on laisse le *mapping* par défaut, une interruption du *timer* va déclencher une *Double Fault* ce qui n'est pas souhaitable. Il a donc été nécessaire de changer cette table de correspondance. Les IRQs 0 à 7 ont été associées aux interruptions 32 à 39 et les IRQs 8 à 15 ont été associées aux interruptions 40 à 47. Ce changement de *mapping* peut se faire assez simplement en assembleur en utilisant les ports des deux PICs utilisés par les IRQs. Un code d'exemple est donné sur le site OSDev [15].

7.2.2 IDT

La table des descripteurs d'interruption (ou IDT) est similaire à la GDT (la table des descripteurs globaux). Elle est aussi composée de descripteurs de 64-bits permettant chacun de référencer une interruption. Un descripteur est composé d'un offset indiquant l'adresse de l'ISR (la routine d'interruption), un selecteur de segment indiquant le segment où se trouve le code de l'ISR et un niveau de privilège indiquant le niveau de privilège requis pour exécuter l'ISR. Dans le cas d'un adressage de type *FLAT* comme celui utilisé, le selecteur de segment sera forcément le selecteur de segment de code. Il existe aussi plusieurs types de descripteurs d'interruptions [8] décrits dans la figure 23. Dans le cas de notre *kernel* seulement deux types ont été utilisés, le type *Interrupt Gate* et le type *Trap Gate*. La différence entre un *Interrupt Gate* et un *Trap Gate* est uniquement le comportement du CPU lors de l'exécution de l'ISR[1]. Dans le cas du *Interrupt Gate*, le CPU masquera les interruptions lors de l'exécution de l'ISR alors que dans un *Trap Gate* ce ne sera pas le cas.

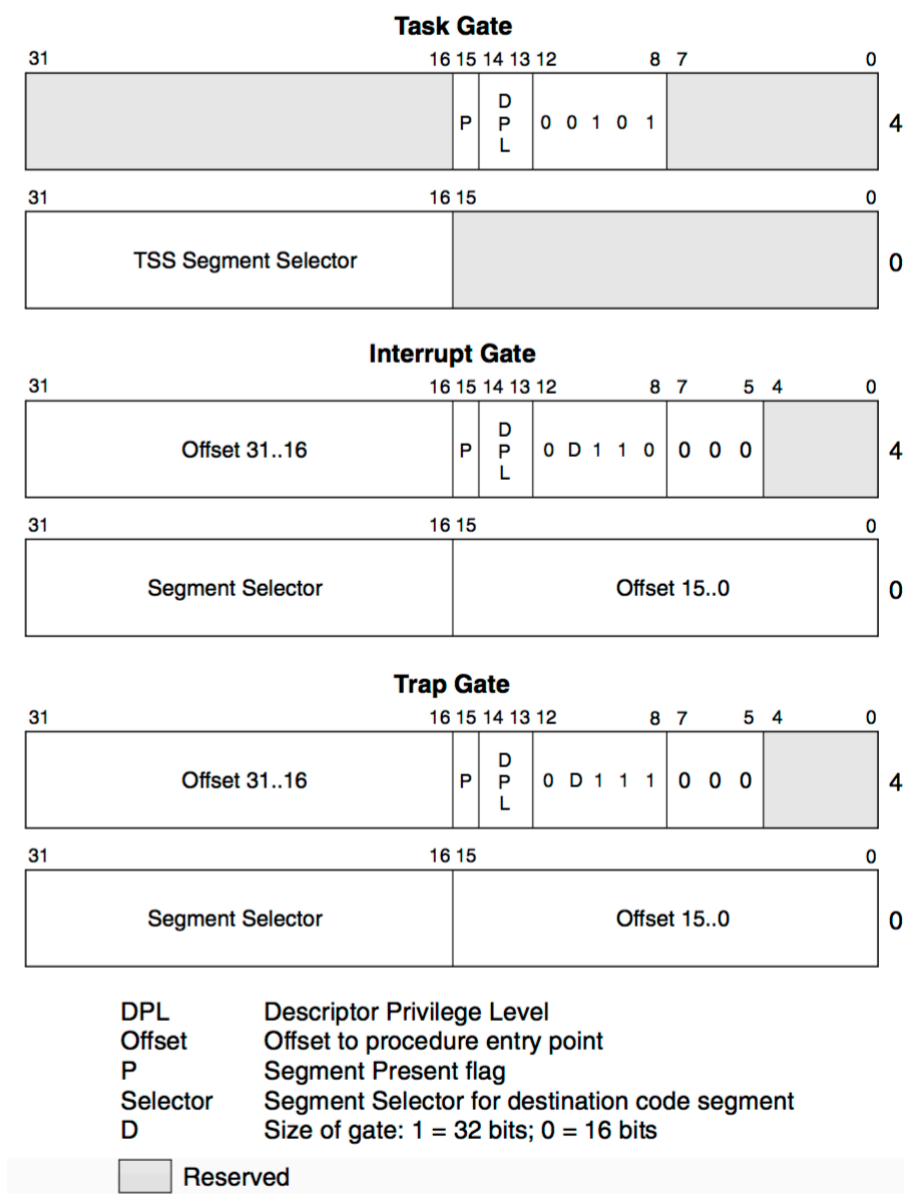


FIGURE 23 – Différents types de descripteur d'interruption

Comme pour la GDT, l'IDT est stockée en RAM et doit donc être initialisée et gérée par l'OS. De la même manière que l'instruction LGDT permet de charger la GDT, l'instruction LIDT permet de charger l'IDT dans le registre IDTR. Pour se faire il faut donner comme argument à l'instruction LIDT l'adresse du descripteur d'IDT sur 48 bits. Ce descripteur est composé de l'adresse de l'IDT sur 32 bits et de sa limite (sa taille en bytes - 1) sur 16 bits. Une fois la table des descripteurs d'interruption chargée avec l'instruction LIDT, les interruptions peuvent être activées en utilisant l'instruction STI. La figure 24 permet de résumer la relation entre le registre IDTR et l'IDT.

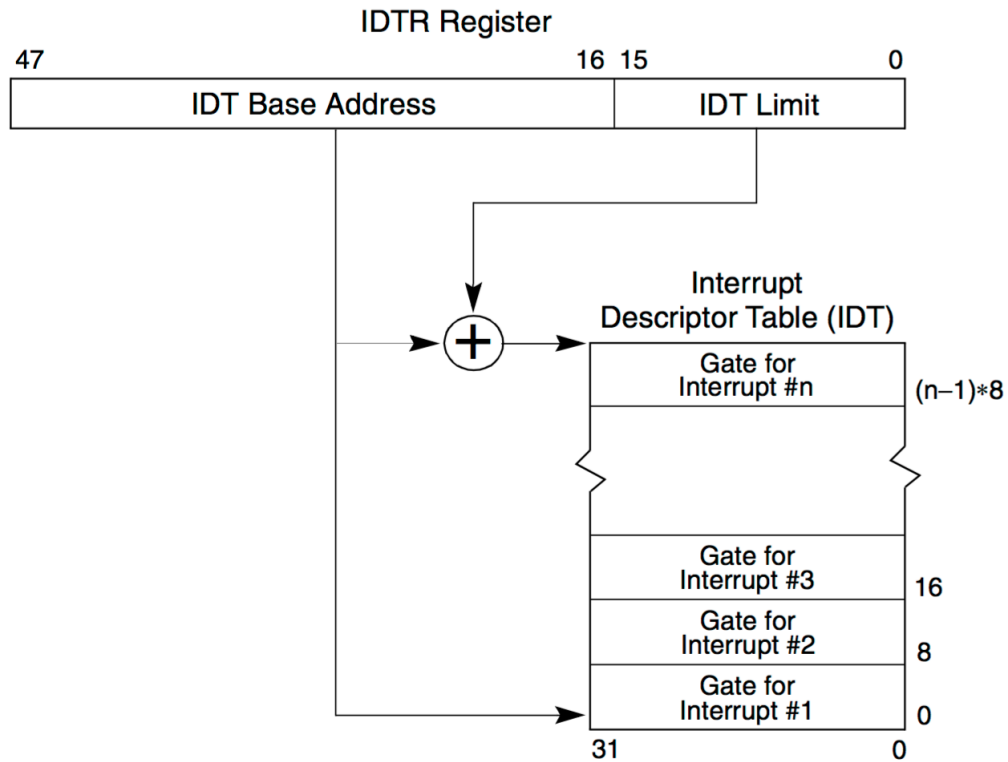


FIGURE 24 – Relation entre le registre IDTR et l'IDT

Dans le *kernel* développé, l'IDT est une structure statique en mémoire. Une fonction assembleur est donc appelée afin de charger cette structure dans le registre IDTR. En plus du chargement de l'IDT, une partie des routines d'interruptions est faite en assembleur. En effet, il a été nécessaire de passer par du code bas niveau car avant de rentrer dans une routine d'interruption, il faut sauvegarder le contexte. Il est obligatoire de sauvegarder le contexte car, comme déjà dit plus haut, une interruption peut avoir lieu à n'importe quel moment. La partie bas niveau de la routine d'interruption s'occupe donc de sauvegarder le contexte puis d'appeler un gestionnaire d'interruption haut niveau en rust. Ce gestionnaire prend comme argument le numéro d'interruption et appelle la routine d'interruption liée à cette interruption. Par exemple, la routine d'interruption du *timer* va simplement incrémenter un compteur. Lorsque une exception est levée le même mécanisme est employé sauf qu'ici le *kernel* va afficher un message d'erreur en fonction du numéro de l'exception.

7.3 VGA

Un PC possède généralement une carte graphique permettant de gérer l’affichage. Une grande majorité des carte graphiques, même modernes sont compatibles avec le standard d’affichage VGA. Dans notre cas, nous utilisons un émulateur (QEMU) qui va émuler l’affichage VGA. Pour écrire sur l’écran il faut écrire dans la mémoire vidéo (VRAM) qui commence à l’adresse 0xA0000 et finit à l’adresse 0xBFFFF. Différents modes d’écriture existent pour l’affichage mais nous allons nous concentrer sur un seul en particulier.

Le mode texte VGA a été utilisé pour l’affichage dans l’OS développé. En mode texte, l’écran est divisé en caractères plutôt qu’en pixels ce qui permet d’afficher simplement et rapidement quelque chose sur l’écran. La mémoire vidéo réservée au mode texte commence à l’adresse 0xB8000 et a une taille de 80×25 caractères. Un caractère est représenté par 2 octets (16 bits) ce qui fait une taille de 4000 octets ($80 \times 25 \times 2$). L’octet de poids faible d’un caractère représente la valeur ASCII de ce caractère et l’octet de poids fort représente l’attribut qui contient lui même la couleur du caractère et la couleur du fond (voir figure 25) [1]. La couleur en mode texte est donc codée sur 4 bits ce qui fait 16 couleurs différentes. Ces 16 couleurs sont décrites dans la figure 26 [16].

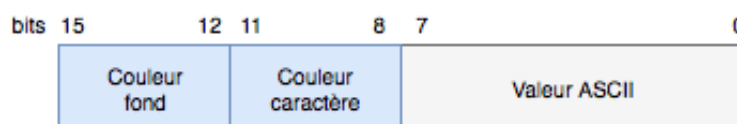


FIGURE 25 – Structure d’un caractère en mode texte VGA

Number	Colour	Name	Number + bright bit	bright Colour	Name
0		Black	0+8=8		Dark Gray
1		Blue	1+8=9		Light Blue
2		Green	2+8=A		Light Green
3		Cyan	3+8=B		Light Cyan
4		Red	4+8=C		Light Red
5		Magenta	5+8=D		Light Magenta
6		Brown	6+8=E		Yellow
7		Light Gray	7+8=F		White

FIGURE 26 – Couleurs disponibles en mode texte VGA

Le mode texte VGA permet aussi d’afficher un curseur. Le curseur ne se déplace pas automatiquement quand un caractère est écrit à l’écran, c’est simplement une zone de l’écran mise en évidence par un clignotement et dont la taille, la position et la visibilité peuvent être modifiés [17]. L’accès au curseur se fait en utilisant les registres du CRTC (*Cathode Ray Tube Controller*). Les registres du CRTC peuvent être accédés avec la paire registre d’adresse et registre de données. Ces registres se trouvent respectivement aux ports 0x3D4 et 0x3D5. L’écriture dans un registre du CRTC se fait donc en deux temps. Tout d’abord, l’adresse du registre doit être spécifiée en écrivant dans le port 0x3D4 puis la donnée doit être écrite dans le port 0x3D5 [1].

Pour l'implémentation du support VGA dans le *kernel* plusieurs structure ont été créées. Tout d'abord, la couleur est représentée par un simple **enum**. Une structure pour l'attribut a ensuite été écrite dont le constructeur prend comme argument la couleur de fond et la couleur du caractère. Après cela, une structure représentant un caractère sur 16 bits a été implémentée. Cette structure est composée du caractère sur 8 bits et de l'attribut sur 8 bits aussi. Ces structures permettent de simplifier la construction d'un caractère pour l'écrire dans le *frame buffer* (mémoire vidéo dédiée au mode texte VGA). Une dernière structure a finalement été créée comportant un *raw pointer* vers le *frame buffer* (qui est un tableau de 80 × 25 caractères) ainsi que les informations sur la couleur et la position du curseur. Si la pagination est active comme c'est le cas dans notre OS, il faut bien penser à mettre l'adresse virtuelle du *frame buffer* (0xB8000 étant l'adresse physique).

```
static mut SCREEN: Screen = Screen {
    buffer: 0xC00B8000 as *mut _,
    attribute: ColorAttribute::new(Color::Black, Color::White),
    cursor_x: 0,
    cursor_y: 0
};
```

La structure **Screen** est déclarée statiquement dans le *kernel* ce qui rend tout appel à une méthode **unsafe** pour rust. En plus des méthodes pour manipuler cette structure, des fonctions ont donc été implémentées afin d'interfacer l'écriture sur l'écran. Ces fonctions permettent en plus d'éviter de mettre le code en **unsafe** à chaque fois que l'on veut afficher quelque chose. En rust, les macros **print** et **println** écrivent sur la sortie standard. Etant donné que nous sommes dans une configuration *bare-metal*, nous n'avons pas de sortie standard et ces deux macros n'existent pas. Si nous étions en C, l'équivalent de ces macros aurait été la fonction **printf** et nous aurions eu à la coder entièrement. Heureusement, rust facilite grandement les choses avec le **trait** **Write**. Pour implémenter ce **trait** dans une structure, il faut simplement lui indiquer comment écrire une chaîne de caractères ce qui a donc été fait pour la structure **Screen**. L'implémentation de ce **trait** par une structure donne accès à cette dernière à de nombreuses méthodes mais celle qui nous intéresse ici est la méthode **write_fmt**. Cette méthode prend en paramètre une structure **Arguments**. La structure **Arguments** permet la gestion des arguments en ligne de commande ou les macros à arguments variable. La macro **print** possède un nombre d'arguments variable, nous pouvons donc implémenter cette macro ainsi que la macro **println** comme suit :

```
macro_rules! print {
    ($($arg:tt)*) => (vga_write_fmt(format_args!($($arg)*)));
}

macro_rules! println {
    () => (print!("{}", "\n"));
    ($fmt:expr) => (print!(concat!($fmt, "\n")));
    ($fmt:expr, $($arg:tt)*) => (print!(concat!($fmt, "\n"), $($arg)*));
}
```

A noter que la fonction **vga_write_fmt** est simplement une fonction interfaçant la méthode **write_fmt** appliquée à la structure statique représentant l'écran et la macro **format_args** convertie les arguments de la macro **print** en structure **Arguments**.

7.4 Timer

Dans chaque PC se trouve une puce pour mesure le temps et implémenter des *timers* [1]. Cette puce est le *Programmable Interval Timer* (PIT). Sur architecture IA-32, le PIT est généralement un Intel 8253 et dans notre cas, c'est celui émulé par QEMU. Le PIT génère une interruption matérielle (IRQ0) à une fréquence sélectionnable entre 18.2065 Hz et 1.19318 MHz. L'horloge du PIT oscille à 1.19318 MHz. La fréquence de sortie est modulée à l'aide d'un diviseur configurable par le *kernel*. Ce diviseur est une valeur sur 16 bits. Sa valeur maximale est donc 65536 ce qui explique la valeur minimale de la fréquence du PIT ($\frac{1193180}{65536} \simeq 18.2065$). Le PIT possède aussi plusieurs canaux avec chacun un diviseur propre mais seulement le premier canal (canal 0) est utilisé dans le *kernel* est c'est sur celui-ci que nous allons nous concentrer. Le PIT est programmable à l'aide de différents ports. Le port 0x43 contient le registre de commande du PIT et le port 0x40 permet de configurer le canal 0. Ci-dessous, le code permettant de programmer le *timer* à une fréquence `freq`.

```
let div = 1193180 / freq;
outb(0x43, 0x36);
outb(0x40, (div & 0xFF) as u8);
outb(0x40, (div >> 0x8) as u8);
```

Ecrire 0x36 dans le registre de commande du PIT indique la sélection du diviseur et le mode répétition (réinitialisation du compteur une fois celui-ci arrivé à 0) [1]. L'octet de poids faible du diviseur est ensuite écrit dans le port du canal 0 suivi par l'octet de poids fort. Dans le *kernel* développé, le *timer* est représenté par une structure statique contenant la fréquence du timer sur 32 bits et l'état actuel de son compteur sur 32 bits aussi. Ce compteur est incrémenté à chaque fois qu'une interruption à lieu sur l'IRQ0.

```
static mut TIMER: Timer = Timer { freq: 0, ticks: 0 };
```

Comme pour la structure statique permettant d'interfacer l'affichage texte VGA, toute modification de la variable `TIMER` sera **unsafe** au niveau du compilateur rust. Des fonctions ont donc été implémentées pour lire ou écrire le *timer*. Actuellement, le PIT est utilisé seulement pour arrêter le *kernel* pendant un temps donné avec la fonction `sleep`. Cette fonction prend une durée en millisecondes et rentre dans une boucle (attente active).

```
pub fn sleep(ms: u32) {
    let duration = get_ticks() + (ms * get_freq() / 1000);
    loop {
        if get_ticks() >= duration {
            break;
        }
        halt();
    }
}
```

Ici, `get_ticks` et `get_freq` sont des fonctions qui renvoient les attributs `ticks` et `freq` de la structure statique `TIMER`. La fonction `halt` appelle l'instruction assembleur `HLT` qui arrête le processeur jusqu'à ce qu'une interruption matérielle soit déclenchée. Elle est appelée après chaque comparaison pour éviter d'utiliser 100% du CPU lors de l'attente sur un `sleep`. On pourrait imaginer d'autres utilisations du *timer* comme par exemple la génération de nombres pseudo-aléatoires.

7.5 Clavier

Afin d'avoir un système d'exploitation un minimum complet, il est impératif de faire un support pour le clavier. Sur architecture Intel 32-bits, chaque pression ou relâchement d'une touche du clavier déclenche une interruption matérielle (IRQ1) [1]. Le périphérique clavier possède un *buffer* interne qui va stocker les données à chaque interruption. Ces données sont appelées *scan codes*. Chaque *scan code* correspond à une touche du clavier et le bit de poids fort du *scan code* indique si la touche a été pressée (*make code*) ou relâchée (*break code*). La valeur de la touche est stockée dans les sept bits de poids faible mais n'est pas à confondre avec un code ASCII (figure 27). Il faut donc faire correspondre la valeur chaque touche avec une valeur ASCII ce qui a été fait avec une table de correspondance dans le *kernel*.



FIGURE 27 – Structure d'un *scan code*

Le clavier comporte deux registres. Un registre de données et un registre d'état. Ces registres sont accessibles par les ports 0x60 et 0x64. Lors d'une interruption, le *scan code* de la touche est stockée dans le registre de données. Avant de lire une donnée venant du clavier, il est nécessaire de s'assurer que le *buffer* du clavier ne soit pas vide en venant lire le registre d'état. Normalement si une interruption a eu lieu, cela veut dire que le *buffer* est plein mais il est quand même important de vérifier pour éviter toute source d'erreur. Une donnée est prête à être lue si le bit 0 (bit de poids faible) du registre d'état est à 1. Ci-dessous, un exemple de routine d'interruption stockant la valeur ascii du *scan code* dans un *buffer*.

```
pub fn keyboard_handler() {  
    let state = inb(0x64) & 1;  
    if state == 1 {  
        let key = inb(0x60);  
        if key >> 7 == 0 {  
            buffer_write(KEY_MAP[key as usize])  
        }  
    }  
}
```

Au niveau de l'implémentation logicielle, un *buffer* circulaire a été utilisé pour la gestion du clavier. Un *buffer* circulaire est un *buffer* de taille fixe dont la fin est connectée au début. Ce type de structure de donnée est bien adapté aux flux constants de données comme la gestion d'un clavier. A chaque fois qu'une touche va être lue par la routine d'interruption du clavier, cette dernière va être stockée dans le *buffer*. La case du *buffer* contenant cette donnée est considérée comme libre si elle a été lue. Le *buffer* circulaire va donc permettre de gérer un nombre d'écritures supérieur au nombre de lectures.

8 Système de fichiers

8.1 Introduction

8.2 Structure

9 Tâches utilisateur

10 Résultats

11 Discussions

11.1 Problèmes rencontrés

11.2 Améliorations possibles

12 Conclusion

13 Références

- [1] Florent Glück. Programmation système avancée, 2017.
- [2] Linker scripts (osdev). https://wiki.osdev.org/Linker_Scripts.
- [3] Linker scripts (scoberlin). http://www.scoberlin.de/content/media/http/informatik/gcc_docs/ld_3.html.
- [4] Linker scripts (math.utah.edu). https://www.math.utah.edu/docs/info/ld_3.html.
- [5] Memory map (x86). [https://wiki.osdev.org/Memory_Map_\(x86\)](https://wiki.osdev.org/Memory_Map_(x86)).
- [6] Multiboot specifications. <https://www.gnu.org/software/grub/manual/multiboot/multiboot.html>.
- [7] David Decotigny et Thomas Petazzoni. Segmentation et interruptions. <http://sos.enix.org/wiki-fr/upload/SOSDownload/sos-texte-art2.pdf>.
- [8] Intel. Ia-32 intel architecture - software developer's manual - volume 3 : System programming guide. <http://flint.cs.yale.edu/cs422/doc/24547212.pdf>, 2003.
- [9] Segmentation. <https://wiki.osdev.org/Segmentation>.
- [10] Gdt. <https://wiki.osdev.org/GDT>.
- [11] David Decotigny et Thomas Petazzoni. Mise en place de la pagination. <http://sos.enix.org/wiki-fr/upload/SOSDownload/sos-texte-art4.pdf>.
- [12] Page translation. https://pdos.csail.mit.edu/6.828/2011/readings/i386/s05_02.htm.
- [13] Jean Gareau. Advanced embedded x86 programming : Paging, june 1998.
- [14] Intel. Ia-32 intel architecture - software developer's manual - volume 1 : System programming guide. <https://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-vol-1-manual.pdf>, 2016.
- [15] 8259 pic. https://wiki.osdev.org/8259_PIC.
- [16] Text ui. https://wiki.osdev.org/Text_UI.
- [17] Crt controller registers. <https://web.stanford.edu/class/cs140/projects/pintos/specs/freevga/vga/crtcreg.htm>.
- [18] Rust book first edition. <https://doc.rust-lang.org/book/first-edition>.
- [19] Rust book second edition. <https://doc.rust-lang.org/book/second-edition>.
- [20] Cargo book. <https://doc.rust-lang.org/cargo>.
- [21] Target option. https://doc.rust-lang.org/1.1.0/rustc_back/target/struct.Target.html.
- [22] Target i386 example. <https://github.com/rust-lang/rust/issues/33497>.
- [23] ___floatundisf issue. <https://users.rust-lang.org/t/kernel-modules-made-from-rust/9191>.
- [24] Writing an os in rust. <https://os.phil-opp.com>.
- [25] Writing an os in rust (second edition). <https://os.phil-opp.com/second-edition>.
- [26] Setting up paging. https://wiki.osdev.org/Setting_Up_Paging.