Syllabus LEPL1503 et LINFO1252

Version 2020

O. Bonaventure, E. Riviere, G. Detal, C. Paasch

10 December 2020

Table des matières

CHAPITRE 1

Introduction

1.1 Introduction

Les systèmes informatiques sont partout dans notre quotidien : de notre téléphone portable à notre montre connectée, de notre ordinateur de bureau aux très nombreux *serveurs* supportant les services internet que nous utilisons quotidiennement. De nombreux objets du quotidien sont eux-aussi des systèmes informatiques : on assiste ainsi à une explosion du nombre d'objets connectés (par exemple, des ampoules programmables ou des hauts-parleurs interactifs) et nombreux sont les objets qui intègrent désormais un ordinateur, comme les voitures ou les appareils électroménagers.

Malgré la diversité de leurs usages, tous ces appareils ont en commun des principes fondamentaux de fonctionnement et d'organisation. Un système informatique intègre toujours au moins un *processeur* (*CPU* en anglais), une mémoire, et un ou des dispositifs d'entrée/sortie lui permettant d'interagir avec son environnement.

Le *CPU* est un circuit électronique réalisant très rapidement des opérations simples :
– lire de l'information en mémoire ;
– écrire de l'information en mémoire ;
– effectuer des calculs.
Les opérations effectuées par le *processeur* sont mises en œuvre directement de manière électronique. On parle d'*instructions*. Le jeu d'instruction d'un *processeur* dépend de son modèle. Par exemple, les ordinateurs Apple et les PCs récents utilisent des processeurs Intel ou AMD permettant d'exécuter le jeu d'instruction `x86_64`. Dans un futur proche, les ordinateurs Apple utiliseront un CPU mettant en œuvre le jeu d'instruction `ARM A64`, le même que celui supporté par le processeur du nano-ordinateur RaspberryPI.

La très grande majorité des processeurs adoptent les principes de l'architecture dite de Von Neumann, du nom du pionnier de l'informatique Jon von Neumann qui l'a proposé. Suivant cette architecture, la mémoire principale est utilisée à la fois pour stocker les informations traitées et produites par le programme à exécuter, mais aussi les instructions composant ce programme.

1.1.1 Représentation de l'information

Un processeur manipule l'information sous forme binaire. L'élément de base pour stocker et représenter de l'information dans un système informatique est donc le *bit*. Un bit (*binary digit* en anglais) peut prendre deux valeurs qui par convention sont représentées par :
– `1`
– `0`
Physiquement, un bit est représenté sous la forme d'un signal électrique ou optique lorsqu'il est transmis et d'une charge électrique ou sous forme physique (par exemple, magnétique) lorsqu'il est stocké. Nous n'aborderons pas ces détails technologiques dans le cadre de ce cours. Ils font l'objet de nombreux cours d'électronique.

Le bit est l'unité de base de stockage et de transfert de l'information. En général, les systèmes informatiques ne traitent pas des bits individuellement, mais par blocs. On appelle un *nibble* un bloc de 4 bits consécutifs et un

octet (ou *byte* en anglais) un bloc de 8 bits consécutifs. On parle de mots (*word* en anglais) pour des groupes comprenant généralement 32 bits et de long mot pour des groupes de 64 bits.

Suivant l'architecture de Von Neumann, les données tout comme les instructions à exécuter par le processeur sont stockées sous forme binaire dans la mémoire. Le processeur va donc lire la prochaine instruction à exécuter sous forme d'un groupe de bits, la déchiffrer, et appliquer l'effet prévu pour l'instruction correspondante, avant de recommencer. L'identité de l'instruction à exécuter est déterminée en décodant les premiers bits de l'instruction ; le reste contient ses arguments ou *opérandes*.

Un jeu d'instruction peut contenir des centaines d'instructions différentes. Certaines instructions ont pour objectif de déterminer la prochaine instruction à lire depuis la mémoire, décoder et exécuter. Ces instructions de contrôle de flux permettent de mettre en œuvre les structures de contrôle : conditionnelles et boucles. D'autres instructions effectuent des calculs, ou permettent de lire ou d'écrire des donnés depuis et vers la mémoire.

1.1.2 Interaction avec le monde extérieur

Le processeur et la mémoire ne sont pas les deux seuls composants d'un système informatique. Celui-ci doit également pouvoir interagir avec le monde extérieur, ne fut-ce que pour pouvoir charger le programme à exécuter et les données à analyser. Cette interaction se réalise grâce à un grand nombre de dispositifs d'entrées/sorties et de stockage. Parmi ceux-ci, on peut citer :
- le clavier qui permet à l'utilisateur d'entrer des caractères ;
- l'écran qui permet à l'utilisateur de visualiser le fonctionnement des programmes et les résultats qu'ils produisent ;
- l'imprimante qui permet à l'ordinateur d'écrire sur papier les résultats de l'exécution de programmes ;
- le disque-dur, les clés USB, les CDs et DVDs qui permettent de stocker les données sous la forme de fichiers et de répertoires ;
- la souris ou la tablette graphique qui permettent à l'utilisateur de fournir à l'ordinateur des indications de positionnement ;
- le scanner qui permet à l'ordinateur de transformer un document en une image numérique ;
- le haut-parleur avec lequel l'ordinateur peut diffuser différentes sortes de son ;
- le microphone et la caméra qui permettent à l'ordinateur de capturer des informations sonores et visuelles pour les stocker ou les traiter.

Les dispositifs d'entrée/sortie et de stockage sont gérés par des contrôleurs de périphériques spécifiques. Par exemple, un contrôleur de périphérique pour le clavier peut être un micro-contrôleur qui interagit avec le dispositif matériel et détecte la frappe de touches. Ce contrôleur de périphérique dispose d'une mémoire propre, qui contient l'identifiant de la touche qui vient d'être frappée.

Il est nécessaire, lorsque l'utilisateur fait une saisie au clavier, que le système puisse récupérer l'information de la mémoire du contrôleur afin de la traiter. Les entrées/sorties se déroulent de manière concurrente (en même temps) que l'exécution par le processeur des instructions du programme principal. Il est donc nécessaire de signaler au processeur qu'un évènement externe est survenu. Ceci est possible grâce au mécanisme d'*interruption*.

Une *interruption* est un signal extérieur au processeur qui force celui-ci à arrêter l'exécution du programme en cours, et à passer le contrôle du processeur à une routine de traitement qui va pouvoir la prendre en compte. Cette routine va tout d'abord déterminer la cause de l'interruption, soit en interrogeant un à un les contrôleurs de périphériques soit en utilisant un vecteur d'interruptions qui indique directement le contrôleur à la source de l'interruption. Un code de traitement spécifique est ensuite appelé qui peut, dans notre exemple, récupérer l'information de la mémoire interne du contrôleur du clavier pour la placer en mémoire principale.

Outre les contrôleurs de périphériques externes comme le clavier, la souris ou une manette de jeu, il existe de nombreuses sources d'*interruption* possibles : une horloge générant une interruption de manière périodique (par exemple, toutes les 10 millisecondes), un dispositif de stockage annonçant la complétion d'une opération d'écriture ou de lecture, ou encore un périphérique réseau annonçant la réception de données.

L'accès direct à la mémoire ou DMA

Nous avons vu qu'une interruption peut permettre le transfert par le processeur d'une information (la touche pressée) entre la mémoire du contrôleur et la mémoire principale. Cette méthode est adéquate pour les périphériques comme le clavier ou un manette de jeux qui génèrent un nombre très limité d'information par seconde. Elle n'est toutefois pas viable pour les périphériques générant ou recevant de grandes quantités de données.
Si chaque réception d'une donnée (par exemple, un mot de 32 bits) génère une interruption, l'écriture de données sur un périphérique de stockage, ou la réception d'une informations sur le réseau, va simplement saturer le processeur d'interruptions et empêcher l'exécution du programme principal. Le système est alors inutilisable pour autre chose que le traitement de l'entrée/sortie.
Les systèmes informatiques modernes supportent le principe de *DMA* pour Direct Memory Access ou accès direct à la mémoire. Avec DMA, un contrôleur de périphérique est autorisé à accéder directement à la mémoire principale pour y lire et écrire des données. Il n'est alors plus nécessaire de générer une interruption pour chaque mot lu ou écrit, mais seulement lorsqu'une qu'un bloc (ensemble) de données est disponible ou a été consommé. Cela permet au processeur de continuer d'exécuter le programme principal en *parallèle* de l'opération d'entrée/sortie.

1.1.3 Rôle(s) du système d'exploitation

L'utilisation *directe* d'un système informatique par un programme unique est en théorie possible : c'est d'ailleurs ainsi que les premiers ordinateurs des années 50 étaient utilisés. Le programme devait alors prévoir les instructions spécifiques pour utiliser les ressources matérielles de l'ordinateur cible, et prendre en compte ses caractéristiques matérielles. Très rapidement (dès la fin des années 50), la nécessité d'un logiciel intermédiaire simplifiant et systématisant l'utilisation du matériel, comme par exemple la gestion des interruptions et des entrées/sorties que nous venons de décrire, s'est imposé. Tout système informatique comprend ainsi depuis un *système d'exploitation*.

Un système d'exploitation remplit trois rôles principaux :
– Le premier rôle du système d'exploitation est de rendre l'exécution et l'utilisation de programmes "utiles" pour l'utilisateur plus aisée et systématique, en simplifiant l'utilisation de ressources matérielles de nature pourtant hétérogènes.
– Son deuxième rôle est de rentre l'utilisation de ces ressources plus efficace, en permettant par exemple le recouvrement entre les opérations d'entrée/sortie et l'exécution des programmes, ou l'utilisation du système par plusieurs programmes et/ou plusieurs utilisateurs *à la fois*.
– Son troisième rôle, enfin, est d'assurer la sécurité et l'intégrité du système informatique lui même et des données qui lui sont confiées. Par exemple, un programme qui rencontre une erreur (e.g., qui essaie d'exécuter une instruction qui n'existe pas) ne doit pas remettre en cause ou stopper l'exécution des autres programmes, et les données d'un utilisateur doivent être protégé de l'accès par d'autres utilisateurs du même système.

Le système d'exploitation remplit ces trois rôles grâce à la **virtualisation** des ressources matérielles. À partir de ressources matérielles de natures variées, le système d'exploitation construit des représentations virtuelles. Ces représentations sont plus faciles à utiliser pour les programmeurs d'applications, et disponibles au travers d'interfaces programmatiques (*Application Programming Interfaces* en anglais - *API*). Par ailleurs, ces représentations sont généralement à visée universelle, c'est à dire qu'elles ne diffèrent pas (ou très peu) d'un système d'exploitation à l'autre, même lorsque les systèmes informatiques et le matériel les composant diffèrent fortement.

Exemples de virtualisation

Nous illustrons ci-dessous le principe de virtualisation en utilisant trois exemples. Bien entendu, l'objectif dans cette introduction n'est pas de comprendre *en détail* les mécanismes et algorithmes permettant leur mise en œuvre, que nous couvrirons dans les chapitres dédiés de ce cours, mais seulement d'illustrer le principe général.

On peut tout d'abord illustrer le principe de virtualisation avec l'utilisation des dispositifs de stockage. Il existe de nombreux dispositifs de stockage (disque dur, clé USB, CD, DVD, mémoire flash, ...). Chacun de ces dispositifs a des caractéristiques électriques et mécaniques propres. Ils permettent en général la lecture et/ou l'écriture de blocs de données de quelques centaines d'octets. Nous reviendrons sur leur fonctionnement ultérieurement. Peu d'applications sont capables de piloter directement de tels dispositifs pour y lire ou y écrire des blocs de données directement, et même si c'était le cas la prise en compte de tous les types de dispositifs disponibles sur le

marché serait impossible. Par contre, la majorité des applications sont capables d'utiliser ces systèmes de stockage par l'intermédiaire du *système de fichiers*, un des composants d'un système d'exploitation. La représentation virtualisée qu'est le système de fichiers (arborescence des fichiers, de répertoires, etc.) et l'API associée (open(2), close(2), read(2), write(2)) sont un exemple des services fournis par le système d'exploitation aux applications.

Un deuxième exemple de virtualisation est la notion de *processus* mise en œuvre par tous les systèmes d'exploitation modernes. Celle-ci est une représentation virtuelle de la notion de programme principal s'exécutant sur le processeur, comme nous l'avons décrit précédemment. Elle permet le partage d'un processeur unique entre plusieurs programmes appartenant à un ou plusieurs utilisateurs. Un processus est l'exécution d'une suite d'instructions contenue dans un fichier programme. Le système d'exploitation donne l'illusion à chaque processus qu'il s'exécute de façon totalement isolée sur un processeur qui lui est dédiée, mais en réalité plusieurs processus alternent leur exécution sur un (ou quelques) processeur(s) partagé(s). Le système d'exploitation met en œuvre la notion de processus en alternant rapidement l'exécution de ces processus sur le ou les processeur(s). Ce principe permet de répondre au deuxième rôle du système d'exploitation, celui de l'efficacité. Lorsqu'un processus doit, par exemple, attendre la complétion d'une entrée/sortie (par exemple, si celui-ci attend qu'une touche du clavier soit pressée, que l'interruption correspondante arrive, qu'elle soit traitée, avant de pouvoir reprendre son exécution), le processeur peut être utilisé par un autre processus.

Un troisième et dernier exemple est la notion de *mémoire virtuelle*. Elle répond à deux problématiques :
– La mémoire physique est une ressource limitée, dont le volume varie selon les systèmes. Un partage *explicite* de la mémoire physique entre processus est complexe à mettre en œuvre : dans les systèmes d'exploitation plus anciens ayant fait ce choix, chaque processus devait prendre en compte, pour accéder à ses instructions et à ses données, les limites de l'espace en mémoire physique qui lui était alloué dynamiquement lors de son initialisation. L'espace mémoire disponible pour chaque processus était fixé une fois pour toute lors de cette initialisation, même si seulement une partie était utilisée en réalité.
– Un deuxième problème est celui de l'isolation entre processus. Idéalement, les données utilisées par un processus ne doivent pas être accessibles par les autres processus s'exécutant sur le système.

La mémoire virtuelle répond élégamment à ces deux problématiques en offrant à chaque processus une vision virtuelle d'un espace mémoire de taille fixe (très grande), dédié, dans lesquelles les adresses déterminées lors de la compilation du programme sont directement valides. Un programme est libre d'allouer et d'utiliser une quantité arbitraire de mémoire (dans les limites de quotas fixée par le système d'exploitation, mais pas nécessairement dans les limites de la mémoire physique disponible), et les données stockées en mémoire physique pour un processus P_A ne sont pas accessibles *via* la mémoire virtuelle d'un processus P_B, sauf si celui-ci l'a explicitement demandé. La mémoire virtuelle participe ainsi des trois rôles du système d'exploitation.

Nous verrons en détails dans ce cours comment tirer parti de ces abstractions. Nous allons maintenant aborder de façon introductive la question de leur mise en œuvre au sein d'un système d'exploitation moderne.

1.1.4 Mise en œuvre du système d'exploitation

La mise en œuvre d'un système d'exploitation est une tâche complexe, qui doit prendre en compte plusieurs facteurs possiblement contradictoires :

1. la nécessité de fournir des abstractions et virtualisations des ressources les plus simples possibles à utiliser pour les programmeurs (et donc de plus haut niveau) ;
2. l'universalité des fonctionnalités, permettant de supporter des applications et usages variés avec un même système d'exploitation ;
3. la performance et le surcoût de ces couches d'abstraction et de virtualisation ;
4. leur complexité de mise en œuvre, et ce faisant, la complexité de leur mise en œuvre *correcte* (sans bug).

La conception d'un système d'exploitation est donc souvent une affaire de compromis entre ces différents aspects. Les coûts de mise en œuvre d'une abstraction dépendent par ailleurs fortement des capacités du matériel utilisé. Nous avons vu plus haut l'exemple de la DMA, permettant le transfert de données par blocs, directement entre un contrôleur de périphérique et la mémoire. Sans le support matériel de la DMA, un système d'exploitation ne peut pas mettre en œuvre efficacement le recouvrement entre les phases d'entrée/sortie d'un processus et les phases de traitement d'un autre processus. Les fonctionnalités des processeurs ont évolué, en réalité, conjointement à celle des systèmes d'exploitation, afin de permettre la mise en œuvre d'abstraction et de virtualisation plus poussées à un coût raisonnable.

Nous verrons plusieurs exemples de support matériel à la virtualisation des ressources et aux fonctions des systèmes d'exploitation dans ce cours. À titre d'illustration, nous allons utiliser le cas de la mémoire virtuelle dans cette introduction.

Le compromis entre abstraction et performance : exemple de la mémoire virtuelle

Comme expliqué précédemment, la mémoire virtuelle a de grands avantages : elle offre à chaque processus l'illusion d'un espace mémoire de grande taille, dont la structure est connue à l'avance (par exemple, la première instruction à exécuter est toujours au même emplacement, la *pile* commence toujours au même endroit, etc.). Le principe de mémoire virtuelle est connu depuis la fin des années 1950, et a été mis en œuvre dans des super-ordinateurs dès les années 1960. On peut donc s'interroger : pourquoi des systèmes d'exploitation pour PC jusqu'aux années 1990 (comme MS-DOS), et des systèmes d'exploitations actuels pour systèmes embarqués (comme uCLinux) ne supportent-ils pas le concept de mémoire virtuelle, et gèrent le partage de la mémoire physique de façon explicite, en indiquant aux processus la plage d'adresses *physiques* qu'ils sont en droit d'utiliser ?
Pour comprendre cela, décrivons de façon simplifiée le fonctionnement de la mémoire virtuelle. Nous le reverrons en détail lors du cours dédié. Un processus P_A est composé d'instructions utilisant des adresses en mémoire virtuelle. Une adresse virtuelle correspond a une adresse en mémoire physique qui est déterminée lors de l'exécution du programme. Il est nécessaire de faire la traduction dynamique entre des adresses virtuelles et des adresses physiques pour chaque instruction accédant à la mémoire en lecture ou écriture. Par exemple, l'adresse virtuelle `0x0000FF00` pour le processus P_A peut correspondre à l'adresse `0x5FD6FF00` en mémoire physique. Cette traduction est effectuée en consultant une structure de donnée stockée elle aussi en mémoire, appelée la *table des pages*. Sans support matériel spécifique, il est nécessaire de transformer toute lecture ou écriture dans la mémoire en deux opérations :

1. Lire la page des tables du processus en cours pour déterminer la correspondance entre adresse virtuelle et adresse physique ;
2. Traduire l'adresse et effectuer l'opération de lecture ou écriture.

L'opération (1) demande systématiquement un accès mémoire supplémentaire pour lire la page des tables. Chaque accès mémoire dans le programme original est ainsi transformé en *deux* accès mémoire. La mémoire étant typiquement un facteur limitant la performance d'exécution des processus, le temps d'exécution peut être simplement doublé ! Le compromis entre utilité et coût n'est alors clairement pas favorable à la mise en œuvre de la mémoire virtuelle.
Pour cette raison, quasiment tous les processeurs modernes intègrent un circuit dédié à la gestion de la virtualisation de la mémoire, appelé la *MMU* (Memory Management Unit). La MMU conserve dans une mémoire très rapide des informations sur les associations entre adresses virtuelles et adresses physiques les plus récemment utilisées, et peut assurer la traduction *en ligne* des adresses. Cela permet, dans la grande majorité des cas, que l'accès mémoire soit aussi rapide qu'un accès direct. Lorsque l'information n'est pas disponible, en revanche, le coût est important : le système d'exploitation doit reprendre la main pour fournir l'information nécessaire à la MMU, ce qui peut prendre un temps équivalent à des centaines voire des milliers d'opérations en mémoire. Comme ce cas de figure n'arrive pas souvent (nous verrons pourquoi), le support matériel qu'est la MMU permet de fournir l'abstraction de haut niveau qu'est la mémoire virtuelle à un coût qui est désormais considéré comme acceptable en regarde de la plus-value qu'elle apporte.

Mécanisme vs. politique

Un aspect important de la mise en œuvre des systèmes d'exploitation, et dont nous discuterons régulièrement dans ce cours, est la séparation entre les mécanismes permettant d'abstraire une ressource matérielle, et les politiques arbitrant le partage de cette ressource (entre les différent programmes, les différents utilisateurs, etc.).

Illustrons ce principe avec l'abstraction de la ressource processeur *via* la notion de processus. Comme nous l'avons expliqué précédemment, chaque processus a l'illusion de s'exécuter sur un processeur unique. En réalité, le système d'exploitation partage le temps de chaque processeur entre l'ensemble des processus disponibles. Bien entendu, un seul processus peut s'exécuter sur un processeur à un moment donné. Régulièrement, le système d'exploitation va donc alterner les processus s'exécutant sur chaque processeur, afin que chaque processus ait régulièrement l'occasion d'exécuter des instructions. L'abstraction processus nécessite donc :

1. Un **mécanisme** permettant d'alterner un processus pour un autre sur un processeur. Ce mécanisme est appelé le *changement de contexte*. Il consiste en deux phases : (1) la sauvegarde de l'état complet du processeur (valeurs des registres, prochaine instruction à exécuter, etc.) dans la mémoire et (2) la restauration de l'état tel que sauvegardé en mémoire de l'autre processus, afin de remettre le processeur dans l'état exact où celui-ci se trouvait lors de sa précédente interruption.
2. Une **politique** qui décide lesquels des processus disponibles pour l'exécution doivent se voir allouer un processeur ou quand un processus en cours d'exécution doit être interrompu. On appelle cette politique une *politique d'ordonnancement* (scheduling en anglais).

Le mécanisme de *changement de contexte* doit avoir un coût le plus faible possible, car son utilisation est un pur surcoût pour le système. La définition de la politique adéquate, en revanche, est plus subtile car elle dépend des objectifs du système informatique considéré. Par exemple, on peut vouloir un partage équitable du temps processeur entre les différents utilisateurs, ou au contraire privilégier des tâches par rapport à d'autres. Pour certaines tâches, comme des simulations de modèles mathématiques, on cherchera à maximiser le *débit applicatif*, c'est à dire le nombre d'instructions utiles effectuées par seconde : on préfèrera alors le moins de changements de contexte possible. Pour d'autres processus dits interactifs on cherchera à minimiser le temps d'attente entre la disponibilité du processus pour être exécuté et la mise à disposition d'un processeur : ici, au contraire, on voudra alterner les processus rapidement pour minimiser le temps d'attente. Ce dernier cas est par exemple celui d'un jeu vidéo. Sur la base de notre exemple de l'entrée/sortie clavier au début de cette introduction, on peut souhaiter minimiser le temps entre la réception de l'interruption depuis le contrôleur de périphérique clavier et le temps auquel le processus jeu peut prendre en compte la commande.

1.1.5 Interaction entre les applications et le système d'exploitation

Le système d'exploitation fournit des services aux applications permettant de simplifier et d'uniformiser l'utilisation des ressources du système informatique. Certains de ces services sont accessibles sous la forme de programmes utilitaires permettant la gestion du système ou son utilisation (comme, par exemple, un gestionnaire de login, un serveur ssh(1), ou un interpréteur de commande). D'autres sont accessibles sous la forme d'*appels système* ; ceux-ci sont appelés directement par le programme ou, le plus souvent, utilisés par une librairie (comme la *libc*) qui facilite leur utilisation.

La mise en œuvre des appels systèmes participe du troisième rôle d'un système d'exploitation, qui est celui de l'isolation des programmes (et des utilisateurs) entre eux. Les processeurs modernes fournissent (au minimum) deux modes d'exécution :

– Le *mode utilisateur* est celui utilisé par les processus courants des utilisateurs, ainsi que par les utilitaires fournis par le système d'exploitation. En mode utilisateur, les programmes utilisent les abstractions fournies par le système d'exploitation comme la mémoire virtuelle, les processus, ou le système de fichiers. Certaines opérations, et l'accès à certaines parties de la mémoire, ne sont pas autorisés en mode utilisateur. Elles seront refusées par le processeur. Par exemple, il n'est pas permis d'utiliser les instructions configurant le traitement des interruptions en mode utilisateur (c'est raisonnable : bloquer le traitement des interruptions permettrait de bloquer le système informatique sans possibilité de corriger le tir).
– Le *mode protégé* est utilisé par le noyau du système d'exploitation. Le noyau est le composant de plus bas niveau, qui interagit directement avec le matériel. Les instructions qui ne sont pas autorisées en mode utilisateur, comme la configuration des interruptions, ou l'envoi d'instructions aux gestionnaires de périphériques, etc. sont autorisées en mode protégé.

L'utilisation de ces deux modes permet d'assurer l'isolation entre les processus s'exécutant en mode utilisateur, tout en permettant au noyau de gérer les ressources de manière globale. L'ensemble des ressources du système informatique, et donc l'ensemble des données stockées par les processus utilisateurs en mémoire, sont accessibles au noyau : la sécurité de ce composant est donc critique !

Un appel système permet d'utiliser une fonctionnalité (par exemple, créer un nouveau processus) qui requiert une action du noyau. Afin d'utiliser un appel système, un processus utilisateur va tout d'abord préparer les arguments de cet appel (comme on le ferait pour un appel de fonction classique). Le passage en mode noyau afin de demander le service de cet appel ne revient pas, en revanche, à un appel de fonction classique (où l'on branche vers la première instruction de la fonction à exécuter). À la place, le processus appelant doit générer une interruption logicielle, ou *trap*. Cette interruption fonctionne d'une manière similaire aux interruption matérielles que nous avons évoqué précédemment. Le processeur va automatiquement passer en mode protégé et exécuter le code point d'entrée du noyau, qui va déterminer l'appel système à traiter et effectuer l'opération. À la fin de ce traitement, le contrôle revient au processus appelant en mode utilisateur.

On notera que les interruption logicielles sont aussi utiles pour permettre au système d'exploitation de traiter les erreurs d'un processus en particulier sans que les autres processus soient affectés. Par exemple, si un processus tente d'effectuer une division par zéro, ou bien tente d'accéder à une partie de la mémoire non autorisée (le fameux `segmentation fault`) alors le processeur génère une interruption logicielle, redonnant ainsi le contrôle au noyau et permettant de traiter le problème (par exemple, en affichant un message d'erreur et en terminant le processus et libérant les ressources qu'il utilisait).

1.1.6 Unix

Il existe de nombreux systèmes d'exploitation, visant des usages variés, des objets connectés (comme RIOT) aux super-calculateurs, en passant par les ordinateurs personnels (comme Windows ou MaxOS). Nous nous intéresserons dans ce cours principalement au système *GNU/Linux*, qui est aujourd'hui le système d'exploitation de loin le plus répandu dans le monde. En effet, des déclinaisons de systèmes utilisant *Linux* équipent des systèmes informatiques de nature très variées, depuis les smartphones (Android) jusqu'aux supercalculateurs.

GNU/Linux est un représentant de la famille de systèmes d'exploitation Unix. Unix est aujourd'hui un nom générique [1]. La première version de Unix a été développée pour faciliter le traitement de documents sur mini-ordinateur.

Quelques variantes de Unix

De nombreuses variantes de Unix ont été produites durant les quarante dernières années. Il est impossible de les décrire toutes, mais en voici quelques unes.

- *Unix*. Initialement développé aux AT&T Bell Laboratories, Unix a été ensuite développé par d'autres entreprises. C'est aujourd'hui une marque déposée par `The Open group`, voir http ://www.unix.org/
- *BSD Unix*. Les premières versions de Unix étaient librement distribuées par Bell Labs. Avec le temps, des variantes d'Unix sont apparues. La variante développée par l'université de Berkeley en Californie a été historiquement importante car c'est dans cette variante que de nombreuses innovations ont été introduites dont notamment les piles de protocoles TCP/IP utilisées sur Internet. Aujourd'hui, *FreeBSD* et *OpenBSD* sont deux descendants de *BSD Unix*. Ils sont utilisés dans de nombreux serveurs et systèmes embarqués. *MacOS*, développé par Apple, s'appuie fortement sur un noyau et des utilitaires provenant de *FreeBSD*.
- *Minix* est un système d'exploitation développé initialement par *Andrew Tanenbaum* à l'université d'Amsterdam. *Minix* est fréquemment utilisé pour l'apprentissage du fonctionnement des systèmes d'exploitation.
- *Linux* est un noyau de système d'exploitation largement inspiré de *Unix* et *Minix*. Développé par *Linus Torvalds* durant ses études d'informatique, il est devenu la variante de Unix la plus utilisée à travers le monde. Il est maintenant développé par des centaines de développeurs qui collaborent via Internet.
- *Solaris* est le nom commercial de la variante Unix de Oracle.

Dans le cadre de ce cours, nous nous focaliserons sur le système *GNU/Linux*, c'est-à-dire un système qui intègre le noyau *Linux* et les librairies et utilitaires développés par le projet *GNU* de la *FSF*.

Un système Unix est composé de trois grands types de logiciels :

- Le noyau du système d'exploitation qui est chargé automatiquement au démarrage de la machine et qui prend en charge toutes les interactions entre les logiciels et le matériel.
- De nombreuses librairies qui facilitent l'écriture et le développement d'applications
- De nombreux programmes utilitaires simples qui permettent de résoudre un grand nombre de problèmes courants. Certains de ces utilitaires sont chargés automatiquement lors du démarrage de la machine. La plupart sont exécutés uniquement à la demande des utilisateurs.

Les systèmes de la famille Unix présentent généralement une arborescence de fichiers unique (contrairement, par exemple, à Windows qui utilise le principe de "volumes" séparés). La racine de cette arborescence est le répertoire `/` par convention. Ce répertoire contient généralement une dizaine de sous répertoires dont les noms varient d'une variante de Unix à l'autre. Généralement, on retrouve dans la racine les sous-répertoires suivants :

- `/usr` : sous-répertoire contenant la plupart des utilitaires et librairies installées sur le système
- `/bin` et `/sbin` : sous-répertoire contenant quelques utilitaires de base nécessaires à l'administrateur du système

1. Formellement, Unix est une marque déposée par l'Open Group, un ensemble d'entreprises qui développent des standards dans le monde de l'informatique. La première version de Unix écrite en C date de 1973, http ://www.unix.org/what_is_unix/history_timeline.html

– `/tmp` : sous-répertoire contenant des fichiers temporaires. Son contenu est généralement effacé au redémarrage du système.
– `/etc` : sous-répertoire contenant les fichiers de configuration du système
– `/home` : sous-répertoire contenant les répertoires personnels des utilisateurs du système
– `/dev` : sous-répertoire contenant des fichiers spéciaux
– `/root` : sous-répertoire contenant des fichiers propres à l'administrateur système. Dans certains variantes de Unix, ces fichiers sont stockés dans le répertoire racine.

Les systèmes Unix ont introduit une bonne partie des concepts et abstractions modernes fournies par les systèmes d'exploitation, comme le partage et l'isolation de la mémoire ou la notion de processus. L'exécution d'un processus est initiée par le système d'exploitation (généralement suite à une requête faite par un autre processus). Ce processus peut s'exécuter pendant une fraction de secondes, quelques secondes ou des journées entières. Pendant son exécution, le processus peut potentiellement accéder aux différentes ressources (processeurs, mémoire, dispositifs d'entrées/sorties et de stockage) du système en utilisant des appels systèmes (ou en utilisant des fonctions de la librairie standard utilisant ces appels système). À la fin de son exécution, le processus se termine [2] ce qui permet au système d'exploitation de libérer les ressources qui lui avaient été allouées. Sous un système d'exploitation Unix, tout processus retourne au processus qui l'avait initié le résultat de son exécution qui est résumée par un nombre entier. Cette valeur de retour est utilisée en général pour déterminer si l'exécution d'un processus s'est déroulée correctement (zéro comme valeur de retour) ou non (valeur de retour différente de zéro).

Dans le cadre de ce cours, nous aurons l'occasion de voir en détails de nombreuses librairies d'un système Unix et verrons le fonctionnement d'appels systèmes qui permettent aux logiciels d'interagir directement avec le noyau. Le système Unix étant majoritairement écrit en langage C, ce langage est le langage de choix pour de nombreuses applications. Nous le verrons donc en détails.

2. Certains processus sont lancés automatiquement au démarrage du système et ne se terminent qu'à son arrêt. Ces processus sont souvent appelés des *daemons*. Il peut s'agir de services qui fonctionnent en permanence sur la machine, comme par exemple un serveur web ou un *daemon* d'authentification.

CHAPITRE 2

Shell

2.1 Utilisation d'un système Unix

Dans cette section, nous allons décrire comment utiliser un système Unix tel que GNU/Linux en mettant l'accent sur l'utilisation de la ligne de commande et l'automatisation de tâches via des *scripts*.

2.1.1 Utilitaires

Unix a été conçu à l'époque des mini-ordinateurs. Un mini-ordinateur servait plusieurs utilisateurs en même temps. Ceux-ci y étaient connectés par l'intermédiaire d'un terminal équipé d'un écran et d'un clavier. Les programmes traitaient les données entrées par l'utilisateur via le clavier ou stockées sur le disque. Les résultats de l'exécution des ces programmes étaient affichés à l'écran, sauvegardés sur disque ou parfois imprimés sur papier.

Unix ayant été initialement développé pour manipuler des documents contenant du texte, il comprend de nombreux utilitaires facilitant ces traitements. Une description détaillée de l'ensemble de ces utilitaires sort du cadre de ce cours. De nombreux livres et ressources Internet fournissent une description détaillée. Voici cependant une brève présentation de quelques utilitaires de manipulation de texte qui peuvent s'avérer très utiles en pratique.

- cat(1) : utilitaire permettant notamment de lire et afficher le contenu d'un fichier.
- echo(1) : utilitaire permettant d'afficher une chaîne de caractères passée en argument.
- head(1) et tail(1) : utilitaires permettant respectivement d'extraire le début ou la fin d'un fichier.
- wc(1) : utilitaire permettant de compter le nombre de caractères et de lignes d'un fichier.
- grep(1) : utilitaire permettant notamment d'extraire d'un fichier texte les lignes qui contiennent ou ne contiennent pas une chaîne de caractères passée en argument.
- sort(1) : utilitaire permettant de trier les lignes d'un fichier texte.
- uniq(1) : utilitaire permettant de filtrer le contenu d'un fichier texte afin d'en extraire les lignes qui sont uniques ou dupliquées (cela requiert que le fichier d'entrée soit trié, car ne compare que les lignes consécutives).
- more(1) : utilitaire permettant d'afficher page par page un fichier texte (less(1) est une variante courante de more(1)).
- gzip(1) et gunzip(1) : utilitaires permettant respectivement de compresser et de décompresser des fichiers. Les fichiers compressés prennent moins de place sur le disque que les fichiers standard et ont par convention un nom qui se termine par `.gz`.
- tar(1) : utilitaire permettant de regrouper plusieurs fichiers dans une archive. Souvent utilisé en combinaison avec gzip(1) pour réaliser des backups ou distribuer des logiciels.
- sed(1) : utilitaire permettant d'éditer, c'est-à-dire de modifier les caractères présents dans un flux de données.
- awk(1) : utilitaire incluant un petit langage de programmation et qui permet d'écrire rapidement de nombreux programmes de manipulation de fichiers de texte.

Pages de manuel

Les systèmes d'exploitation de la famille Unix contiennent un grand nombre de librairies, d'appels systèmes et d'utilitaires. Toutes ces fonctions et tous ces programmes sont documentés dans des pages de manuel qui sont accessibles via la commande `man`. Les pages de manuel sont organisées en 8 sections.
– Section 1 : Utilitaires disponibles pour tous les utilisateurs
– Section 2 : Appels systèmes en C
– Section 3 : Fonctions de la librairie
– Section 4 : Fichiers spéciaux
– Section 5 : Formats de fichiers et conventions pour certains types de fichiers
– Section 6 : Jeux
– Section 7 : Utilitaires de manipulation de fichiers textes
– Section 8 : Commandes et procédure de gestion du système
Dans le cadre de ce cours, nous aborderons principalement les fonctionnalités décrites dans les trois premières sections des pages de manuel. L'accès à une page de manuel se fait via la commande `man` avec comme argument le nom de la commande concernée. Vous pouvez par exemple obtenir la page de manuel de `gcc` en tapant `man gcc`. `man` supporte plusieurs paramètres qui sont décrits dans sa page de manuel accessible via `man man`. Dans certains cas, il est nécessaire de spécifier la section du manuel demandée. C'est le cas par exemple pour `printf` qui existe comme utilitaire (section 1) et comme fonction de la librairie (section 3 - accessible via `man 3 printf`).
Outre ces pages de manuel locales, il existe également de nombreux sites web où l'on peut accéder aux pages de manuels de différentes versions de Unix dont notamment :
– les pages de manuel de Debian GNU/Linux
– les pages de manuel de FreeBSD
– les pages de manuel de MacOS
Dans la version en-ligne de ces notes, toutes les références vers un programme Unix, un appel système ou une fonction de la librairie pointent vers la page de manuel Linux correspondante.

2.1.2 Shell

Avant le développement des interfaces graphiques telles que *X11*, *Gnome* ou *Aqua*, l'utilisateur interagissait exclusivement avec l'ordinateur par l'intermédiaire d'un interpréteur de commandes. Dans le monde Unix, le terme anglais *shell* est le plus souvent utilisé pour désigner cet interpréteur et nous ferons de même.

Un *shell* est un programme qui a été spécialement conçu pour faciliter l'utilisation d'un système Unix via le clavier. De nombreux shells Unix existent. Les plus simples permettent à l'utilisateur de taper une série de commandes à exécuter en les combinant. Les plus avancés sont des interpréteurs de commandes qui supportent un langage complet permettant le développement de scripts plus ou moins ambitieux. Dans le cadre de ce cours, nous utiliserons bash(1) qui est un des shells les plus populaires et les plus complets. La plupart des commandes bash(1) que nous utiliserons sont cependant compatibles avec de nombreux autres shells tels que zsh ou csh.

Bien que les interfaces graphiques se soient désormais généralisées, le shell reste un moyen d'interaction avec le système parfaitement complémentaire, particulièrement utile pour les informaticiens, ou toute personne devant automatiser des traitements et opérations sur un système informatique. Avec les interfaces graphiques actuelles, le shell est accessible par l'intermédiaire d'une application qui est généralement appelée `terminal` ou `console`.

Lorsqu'un utilisateur se connecte à un système Unix, en direct ou à travers une connexion réseau, le système vérifie son mot de passe puis exécute automatiquement le shell qui est associé à cet utilisateur depuis son répertoire par défaut. Ce shell permet à l'utilisateur d'exécuter et de combiner des commandes. Un shell supporte deux types de commande : les commandes internes qu'il implémente directement et les commandes externes qui font appel à un utilitaire stocké sur disque. Un exemple de commande interne est cd(1posix) qui prend comme argument un chemin (relatif ou absolu) et y positionne le répertoire courant. Les utilitaires présentés dans la section précédente sont des exemples de commandes externes.

Voici quelques exemples d'utilisation de commandes externes.

```
$ cat exemple.txt
Un simple fichier de textes
aaaaaaaaaa bbbbbb
```

```
bbbbb ccccccccccc
eeeee fffffffff
aaaaaaaaaa bbbbbb
$ grep fichier exemple.txt
Un simple fichier de textes
$ wc exemple.txt
       5      13      98 exemple.txt
```

2.1.3 Combiner des commandes

La plupart des utilitaires fournis avec un système Unix ont été conçus pour être utilisés en combinaison avec d'autres. Cette combinaison efficace de plusieurs petits utilitaires est un des points forts des systèmes Unix par rapport à d'autres systèmes d'exploitation. On peut imaginer par exemple associer sort(1) et head(1) pour n'afficher que les premiers noms en ordre alphabétique d'une liste d'étudiants disponible initialement sous forme non triée. Afin de permettre cette combinaison, chaque programme Unix en cours d'exécution (appelé un *processus*) est associé à trois *flux* standards :

– une entrée standard (*stdin* en anglais) qui est un flux d'informations par lequel le processus reçoit les données à traiter. Par défaut, l'entrée standard est associée au clavier.
– une sortie standard (*stdout* en anglais) qui est un flux d'informations sur lequel le processus écrit le résultat de son traitement. Par défaut, la sortie standard est associée au terminal.
– une sortie d'erreur standard (*stderr* en anglais) qui est un flux de données sur lequel le processus écrira les messages d'erreur éventuels. Par défaut, la sortie d'erreur standard est associée au même terminal que *stdout*.

La puissance du *shell* vient de la possibilité de combiner des commandes en redirigeant les entrées et sorties standards. Les shells Unix supportent différentes formes de redirection. Tout d'abord, il est possible de forcer un programme à lire son entrée standard depuis un fichier plutôt que depuis le clavier. Cela se fait en ajoutant à la fin de la ligne de commande le caractère < suivi du nom du fichier à lire. Ensuite, il est possible de rediriger la sortie standard vers un fichier. Cela se fait en utilisant > ou >>. Lorsqu'une commande est suivie de `> file`, le fichier `file` est créé si il n'existait pas et remis à zéro si il existait, et la sortie standard de cette commande est redirigée vers le fichier `file`. Lorsqu'un commande est suivie de `>> file`, la sortie standard est sauvegardée à la fin du fichier `file` (si `file` n'existait pas, il est créé).

Voici un exemple d'utilisation des redirections :

```
$ echo "Un petit fichier de textes" > file.txt
$ echo "aaaaa bbbbb" >> file.txt
$ echo "bbbb ccc" >> file.txt
$ grep -v bbbb < file.txt > file.out
$ cat file.out
Un petit fichier de textes
```

Note : Rediriger la sortie d'erreur standard

La redirection > *file* redirige par défaut la sortie standard vers le fichier *file*. La sortie d'erreur standard reste dirigé, quand à elle, vers le terminal de l'utilisateur. Il arrive toutefois que l'on souhaite diriger les messages d'erreur vers un fichier différent. On peut pour cela utiliser la notation *2> file_errors* (le flux *stdout* est numéroté 1 et le flux *stderr* est numéroté 2 ; la notation > *file* est implicitement équivalente à *1> file*).

Si l'on souhaite rediriger à la fois *stdout* et *stderr* vers le même fichier on ne peux pas utiliser > *file 2> file* ! Il faut d'abord rediriger la sortie *stderr* vers *stdout*, puis diriger ce dernier vers le fichier. Le flux *stdout* est noté *&1*, on utilise donc *2>&1 > file*.

Des informations plus complètes sur les mécanismes de redirection de bash(1) peuvent être obtenues dans le chapitre 20 de [ABS].

Les shells Unix supportent un second mécanisme qui est encore plus intéressant pour combiner plusieurs programmes. Il s'agit de la redirection de la sortie standard d'un programme vers l'entrée standard d'un autre sans passer par un fichier intermédiaire. Cela se réalise avec le symbole | (*pipe* en anglais). L'exemple suivant illustre quelques combinaisons d'utilitaires de manipulation de texte.

```
$ echo "Un petit texte" | wc -c
      15
$ echo "bbbb ccc" >> file.txt
$ echo "aaaaa bbbbb" >> file.txt
$ echo "bbbb ccc" >> file.txt
$ cat file.txt
bbbb ccc
aaaaa bbbbb
bbbb ccc
$ cat file.txt | sort | uniq
aaaaa bbbbb
bbbb ccc
```

Le premier exemple utilise echo(1) pour générer du texte et le passer directement à wc(1) qui compte le nombre de caractères. Le deuxième exemple utilise cat(1) pour afficher sur la sortie standard le contenu d'un fichier. Cette sortie est reliée à sort(1) qui trie le texte reçu sur son entrée standard en ordre alphabétique croissant. Cette sortie en ordre alphabétique est reliée à uniq(1) qui la filtre pour en retirer les lignes dupliquées.

2.1.4 Scripts : les bases

Tout shell Unix peut également s'utiliser comme un interpréteur de commande qui permet d'interpréter des scripts. Un système Unix peut exécuter deux types de programmes :
– des programmes exécutables en langage machine. C'est le cas de la plupart des utilitaires dont nous avons parlé jusqu'ici.
– des programmes écrits dans un langage interprété. C'est le cas des programmes écrits pour le shell, mais également pour d'autres langages interprétés comme python ou perl.

Lors de l'exécution d'un programme, le système d'exploitation reconnaît [1] si il s'agit d'un programme directement exécutable ou d'un programme interprété en analysant les premiers octets du fichier. Par convention, sous Unix, les deux premiers caractères d'un programme écrit dans un langage qui doit être interprété sont `#!`. Ils sont suivis par le nom complet de l'interpréteur qui doit être utilisé pour interpréter le programme.

Le programme bash(1) le plus simple est le suivant :

```
#!/bin/bash
echo "Hello, world"
```

L'exécution de ce script shell retourne la sortie suivante :

```
Hello, world
```

Par convention en bash(1), le caractère `#` marque le début d'un commentaire en début ou en cours de ligne. Comme tout langage, bash(1) permet à l'utilisateur de définir des variables. Celles-ci peuvent contenir des chaînes de caractères ou des nombres. Le script ci-dessous utilise deux variables, `PROG` et `COURS` et les utilise pour afficher un texte avec la commande `echo`.

```
#!/bin/bash
PROG="LINFO"
COURS=1252
echo $PROG$COURS
```

On note dans l'exemple ci-dessus l'utilisation du symbole `$` pour référer à la valeur de la variable. Dans la majorité des cas, cette notation suffit. Il y a une subtilité auxquelles ont doit faire attention : si il y a une ambiguïté possible sur le nom de la variable pour l'interpréteur il convient d'entourer son nom d'accolades `{ }`. Par exemple, `milieu = "mi"; echo do$milieuno` affichera `do` seulement car l'interpréteur considère la seconde partie comme la variable `$milieuno` non définie et donc égale à la chaîne vide (et cela sans générer de message d'erreur). Avec `echo do${milieu}no`, par contre, le résultat est celui attendu.

1. Sous Unix et contrairement à d'autres systèmes d'exploitation, le suffixe d'un nom de fichier ne joue pas de rôle particulier pour indiquer si un fichier contient un programme exécutable ou non. Comme nous le verrons ultérieurement, le système de fichiers Unix contient des bits de permission qui indiquent notamment si un fichier est exécutable ou non.

Un script bash(1) peut également prendre des arguments passés en ligne de commande. Par convention, ceux-ci ont comme noms `$1`, `$2`, `$3`, ... Le nombre d'arguments s'obtient avec `$#` et la liste complète avec `$@`. L'exemple ci-dessous illustre l'utilisation de ces arguments.

```
#!/bin/bash
# $# nombre d'arguments
# $1 $2 $3 ... arguments
echo "Vous avez passe" $# "arguments"
echo "Le premier argument est :" $1
echo "Liste des arguments :" $@
```

L'exécution de ce script produit la sortie suivante :

```
Vous avez passe 2 arguments
Le premier argument est : LINFO
Liste des arguments : LINFO 1252
```

Concernant le traitement des arguments par un script bash, il est utile de noter que lorsque l'on appelle un script en redirigeant son entrée ou sa sortie standard, le script n'est pas informé de cette redirection. Ainsi, si l'on exécute le script précédent en faisant `args.sh arg1 > args.out`, le fichier `args.out` contient les lignes suivantes :

```
Vous avez passe 2 arguments
Le premier argument est : LINFO
Liste des arguments : LINFO 1252
```

2.1.5 Scripts : conditionnelles

Un script permet d'utiliser des structures de contrôle comme dans tout langage de programmation.

bash(1) supporte tout d'abord la construction conditionnelle `if [ condition ]; then ... fi` qui permet notamment de comparer les valeurs de variables. bash(1) définit de nombreuses conditions différentes, dont notamment :

- `$i -eq $j` est vraie lorsque les deux variables `$i` et `$j` contiennent le même nombre.
- `$i -lt $j` est vraie lorsque la valeur de la variable `$i` est numériquement strictement inférieure à celle de la variable `$j`
- `$i -ge $j` est vraie lorsque la valeur de la variable `$i` est numériquement supérieure ou égale à celle de la variable `$j`
- `$s = $t` est vraie lorsque la chaîne de caractères contenue dans la variable `$s` est égale à celle qui est contenue dans la variable `$t`
- `-z $s` est vraie lorsque la chaîne de caractères contenue dans la variable `$s` est vide

D'autres types de test sont définis dans la page de manuel : bash(1). Le script ci-dessous fournit un premier exemple d'utilisation de tests avec bash(1).

```
#!/bin/bash
# Vérifie si les deux nombres passés en arguments sont égaux
if [ $# -ne 2 ]; then
    echo "Erreur, deux arguments sont nécessaires" > /dev/stderr
    exit 2
fi
if [ $1 -eq $2 ]; then
        echo "Nombres égaux"
else
        echo "Nombres différents"
fi
exit 0
```

Tout d'abord, ce script vérifie qu'il a bien été appelé avec deux arguments. Vérifier qu'un programme reçoit bien les arguments qu'il attend est une règle de bonne pratique qu'il est bon de respecter dès le début. Si le script n'est pas appelé avec le bon nombre d'arguments, un message d'erreur est affiché sur la sortie d'erreur standard et le script se termine avec un code de retour. Ces codes de retour sont importants car ils permettent à un autre programme, par exemple un autre script bash(1) de vérifier le bon déroulement d'un programme appelé. Le script

`src/eq.sh` utilise des appels explicites à exit(1posix) même si par défaut, un script bash(1) qui n'en contient pas retourne un code de retour nul à la fin de son exécution.

Un autre exemple d'utilisation des codes de retour est le script `src/wordin.sh` repris ci-dessous qui utilise grep(1) pour déterminer si un mot passé en argument est présent dans un fichier texte. Pour cela, il exploite la variable spéciale `$?` dans laquelle bash(1) sauve le code de retour du dernier programme exécuté par le script.

```
#!/bin/bash
# wordin.sh
# Vérifie si le mot passé en premier argument est présent
# dans le fichier passé comme second argument
if [ $# -ne 2 ]; then
    echo "Erreur, deux arguments sont nécessaires" > /dev/stderr
    exit 2
fi
grep $1 $2 >/dev/null
# $? contient la valeur de retour de grep
if [ $? -eq 0 ]; then
        echo "Présent"
        exit 0
else
        echo "Absent"
        exit 1
fi
```

Ce programme utilise le fichier spécial `/dev/null`. Celui-ci est en pratique l'équivalent d'un trou noir. Il accepte toutes les données en écriture mais celles-ci ne peuvent jamais être relues. `/dev/null` est très utile lorsque l'on veut ignorer la sortie d'un programme et éviter qu'elle ne s'affiche sur le terminal. bash(1) supporte également `/dev/stdin` pour représenter l'entrée standard, `/dev/stdout` pour la sortie standard et `/dev/stderr` pour l'erreur standard.

Les scripts servent souvent à réaliser des opérations sur des fichiers, et il est parfois nécessaire de pouvoir tester si un fichier est présent, n'est pas vide, etc. On peut alors utiliser les conditions suivantes :
– `-f file` est vraie si `file` existe et est un fichier ;
– `-s file` est vraie si `file` n'est pas vide ;
– `-r file`, `-w file`, `-x file` est vraie si `file` peut, respectivement, être *lu*, *écrit* ou *exécuté* par l'utilisateur lançant le script ;
– `-s file` est vraie si `file` est le nom d'un répertoire.
L'exemple ci-dessous illustre l'utilisation des conditions sur les fichiers :

```
#!/bin/bash
# vérifie si le fichier fourni en entrée contient des données
if [ $# -ne 1 ]; then
    echo "Erreur, un seul argument est nécessaire" > /dev/stderr
    exit 2
fi
if [ -d $1 ]; then
    echo "Erreur, $1 est un répertoire, pas un fichier" > /dev/stderr
    exit 2
fi
if [ ! -f $1 ]; then
    echo "Erreur, $1 n'existe pas" > /dev/stderr
    exit 2
fi
if [ -s $1 ]; then
    echo "Le fichier $1 contient des données" > /dev/stdout
else
        echo "Le fichier $1 ne contient pas de données" > /dev/stdout
fi
```

On note l'utilisation du combinateur logique de négation `!` pour la troisième condition. Deux autres opérateurs logiques sont disponibles : `-a` est le ET logique (AND) et le `-o` est le OU logique (OR) :

– `-a` renvoie une valeur positive (faux) si au moins une des deux conditions renvoie une valeur positive, et 0 sinon (vrai) ;
– `-o` renvoie une valeur positive (faux) si les deux conditions renvoient une valeur positive, et 0 sinon (vrai).
La deuxième et la troisième condition de l'exemple ci-dessus peuvent ainsi être combinées de la manière suivante :

```
if [ -d $1 -o ! -f $1 ]; then
    echo "Erreur, $1 n'existe pas ou est un répertoire" > /dev/stderr
    exit 2
fi
```

La structure `case` permet de vérifier une entrée contre une série de motifs. Cela est souvent utile, par exemple, pour l'analyse des paramètres fournis à un script (en utilisant - et --). La description de `case` dépasse cependant le cadre de ce cours.

2.1.6 Scripts : boucles

On utilise régulièrement des boucles pour répéter une opération pour plusieurs argument. Voici un exemple d'utilisation de la boucle `for` :

```
#!/bin/bash
# exemple_for.sh
students="Julie Maxime Hakim"
for s in $students; do
        l=`wc -l TP1-$s.txt | cut -d' ' -f1`
        echo "Bonjour $s, ton compte rendu de TP comporte $l lignes."
done
```

```
Bonjour Julie, ton compte rendu de TP comporte 26 lignes.
Bonjour Maxime, ton compte rendu de TP comporte 32 lignes.
Bonjour Hakim, ton compte rendu de TP comporte 28 lignes.
```

Une utilisation courante de la boucle `for` est pour répéter un même traitement sur tous les fichiers présents dans une liste. La boucle est alors un itérateur : pour chaque itération la variable *s* prend une des valeurs des éléments de la liste séparés par des espaces.

Cet exemple utilise, par ailleurs, une autre construction utile de bash(1) : les symboles d'apostrophe inversée \` \`. Ceux-ci permettent d'obtenir le résultat (envoyé sur *stdout*) de l'exécution de la commande qu'il englobent. Examinons en détail cette commande. Celle-ci combine en utilisant un *pipe* les commandes wc(1) et cut(1). La première permet d'obtenir le nombre de ligne du fichier fourni en paramètre (utilisation de l'option `-l`). Toutefois, la sortie de la commande contient trop d'information (par exemple, `wc -l TP1-Hakim.txt` renvoie " 28 TP1-Hakim.txt"). Afin d'isoler l'information qui nous intéressent, l'utilitaire cut(1) est utilisé. Celui-ci permet ici de sélectionner quelle colonne (la première ici) conserver, lorsque les colonnes sont séparées par des espaces (comme spécifié en utilisant l'option "-d' ").

La boucle `for` peut aussi prendre comme entrée (liste sur laquelle itérer) une expression utilisant des caractères *joker* `*` ou `?`. bash(1) transforme alors l'expression en la liste des noms de fichiers et répertoires dans le répertoire courant correspondant à l'expression :
– `*` représente n'importe quelle suite de caractères (y compris la chaîne vide). `ab*xyz` correspond à, par exemple, `abcdxyz` ou `abcxyz` mais pas à `abcdyz`.
– `?` représente un caractère unique inconnu (mais pas la chaîne vide). `ab?de` correspond, ainsi, à `abcde`, `abXde` mais pas à `abde`.
Voici un exemple de l'utilisation du caractère `*`, qui calcule une signature de chaque fichier sous la forme d'un *hash* SHA-1 :

```
#!/bin/bash
for f in TP1-*.txt; do
        echo "Génération de la signature du fichier $f ..."
        sha1sum $f > $f.sha1
done
```

```
Génération de la signature du fichier TP1-Hakim.txt ...
Génération de la signature du fichier TP1-Julie.txt ...
Génération de la signature du fichier TP1-Maxime.txt ...
```

bash(1) permet aussi l'utilisation de boucles `while` et `until` sur un principe similaire, mais nous ne les couvrirons pas dans ce cours.

De nombreuses références sont disponibles pour aller plus loin dans la maîtrise de bash(1) [Cooper2011].

CHAPITRE 3

Langage C

3.1 Le langage C

Différents langages permettent au programmeur de construire des programmes qui seront exécutés par le processeur. En réalité, le processeur ne comprend qu'un langage : le langage machine. Ce langage est un langage binaire dans lequel toutes les commandes et toutes les données sont représentés sous la forme de séquences de bits.

Le langage machine est peu adapté aux humains et il est extrêmement rare qu'un informaticien doive manipuler des programmes directement en langage machine. Par contre, pour certaines tâches bien spécifiques, comme par exemple le développement de routines spéciales qui doivent être les plus rapides possibles ou qui doivent interagir directement avec le matériel, il est important de pouvoir efficacement générer du langage machine. Cela peut se faire en utilisant un langage d'assemblage. Chaque famille de processeurs a un langage d'assemblage qui lui est propre. Le langage d'assemblage permet d'exprimer de façon symbolique les différentes instructions qu'un processeur doit exécuter. Nous aurons l'occasion de traiter à plusieurs reprises des exemples en langage d'assemblage dans le cadre de ce cours. Cela nous permettra de mieux comprendre la façon dont le processeur fonctionne et exécute les programmes. Le langage d'assemblage est converti en langage machine grâce à un *assembleur*.

Le langage d'assemblage est le plus proche du processeur. Il permet d'écrire des programmes compacts et efficaces. C'est aussi souvent la seule façon d'utiliser des instructions spéciales du processeur qui permettent d'interagir directement avec le matériel pour par exemple commander les dispositifs d'entrée/sortie. C'est essentiellement dans les systèmes embarqués qui disposent de peu de mémoire et pour quelques fonctions spécifiques des systèmes d'exploitation que le langage d'assemblage est utilisé de nos jours. La plupart des programmes applicatifs et la grande majorité des systèmes d'exploitation sont écrits dans des langages de plus haut niveau.

Le langage *C* [KernighanRitchie1998], développé dans les années 70 pour écrire les premières versions du système d'exploitation *Unix*, est aujourd'hui l'un des langages de programmation les plus utilisés pour développer des programmes qui doivent être rapides ou doivent interagir avec le matériel. La plupart des systèmes d'exploitation sont écrits en langage C.

Le langage C a été conçu à l'origine comme un langage proche du processeur qui peut être facilement compilé, c'est-à-dire traduit en langage machine, tout en conservant de bonnes performances.

La plupart des livres qui abordent la programmation en langage C commencent par présenter un programme très simple qui affiche à l'écran le message *Hello, world !*.

```
/************************************
 * Hello.c
 *
 * Programme affichant sur la sortie
 * standard le message "Hello, world!"
 *
 ************************************/
```

```c
#include <stdio.h>
#include <stdlib.h>

int main(int argc, char *argv[])
{
   // affiche sur la sortie standard
   printf("Hello, world!\n");

   return EXIT_SUCCESS;
}
```

Pour être exécuté, ce programme doit être compilé. Il existe de nombreux compilateurs permettant de transformer le langage C en langage machine. Dans le cadre de ce cours, nous utiliserons gcc(1). Dans certains cas, nous pourrons être amenés à utiliser d'autres compilateurs comme *llvm*.

La compilation du programme `src/hello.c` peut s'effectuer comme suit sur une machine de type Unix :

```
$ gcc -Wall -o hello hello.c
$ ls -l
total 80
-rwxr-xr-x  1 obo  obo  8704 15 jan 22:32 hello
-rw-r--r--  1 obo  obo   288 15 jan 22:32 hello.c
```

gcc(1) supporte de très nombreuses options et nous aurons l'occasion de discuter de plusieurs d'entre elles dans le cadre de ce cours. Pour cette première utilisation, nous avons choisi l'option *-Wall* qui force gcc(1) à afficher tous les messages de type *warning* (dans cet exemple il n'y en a pas) et l'option *-o* suivie du nom de fichier *hello* qui indique le nom du fichier dans lequel le programme exécutable doit être sauvegardé par le compilateur [1].

Lorsqu'il est exécuté, le programme *hello* affiche simplement le message suivant sur la sortie standard :

```
$ ./hello
Hello, world!
$
```

Même si ce programme est très simple, il illustre quelques concepts de base en langage C. Tout d'abord comme en Java, les compilateurs récents supportent deux façons d'indiquer des commentaires en C :
– un commentaire sur une ligne est précédé des caractères //
– un commentaire qui comprend plusieurs lignes débute par /* et se termine par */
Ensuite, un programme écrit en langage C comprend principalement des expressions en langage C mais également des expressions qui doivent être traduites par le *préprocesseur*. Lors de la compilation d'un fichier en langage C, le compilateur commence toujours par exécuter le préprocesseur. Celui-ci implémente différentes formes de macros qui permettent notamment d'inclure des fichiers (directives `#include`), de compiler de façon conditionnelle certaines lignes ou de définir des constantes. Nous verrons différentes utilisations du préprocesseur C dans le cadre de ce cours. À ce stade, les trois principales fonctions du préprocesseur sont :
– définir des substitutions via la macro `#define`. Cette macro est très fréquemment utilisée pour définir des constantes ou des substitutions qui sont valables dans l'ensemble du programme.

  ```c
  #define ZERO  0
  #define STRING "LEPL1503"
  ```

– importer (directive `#include`) un fichier. Ce fichier contient généralement des prototypes de fonctions et des constantes. En langage C, ces fichiers qui sont inclus dans un programme sont appelés des *header files* et ont par convention un nom se terminant par `.h`. Le programme `src/hello.c` ci-dessus importe deux fichiers *headers* standards :
  – `<stdio.h>` : contient la définition des principales fonctions de la librairie standard permettant l'interaction avec l'entrée et la sortie standard, et notamment printf(3)
  – `<stdlib.h>` : contient la définition de différentes fonctions et constantes de la librairie standard et notamment `EXIT_SUCCESS` et `EXIT_FAILURE`. Ces constantes sont définies en utilisant la macro `#define` du préprocesseur

1. Si cette option n'était pas spécifiée, le compilateur aurait placé le programme compilé dans le fichier baptisé *a.out*.

```
#define EXIT_FAILURE    1
#define EXIT_SUCCESS    0
```

– inclure du code sur base de la valeur d'une constante définie par un `#define`. Ce contrôle de l'inclusion de code sur base de la valeur de constantes est fréquemment utilisé pour ajouter des lignes qui ne doivent être exécutées que lorsque l'on veut déboguer un programme. C'est aussi souvent utilisé pour faciliter la portabilité d'un programme entre différentes variantes de Unix, mais cette utilisation sort du cadre de ce cours.

```
#define DEBUG
/* ... */
#ifdef DEBUG
printf("debug : ...");
#endif /* DEBUG */
```

Il est également possible de définir des macros qui prennent un ou plusieurs paramètres [CPP].

Les *headers* standards sont placés dans des répertoires bien connus du système. Sur la plupart des variantes de Unix ils se trouvent dans le répertoire `/usr/include/`. Nous aurons l'occasion d'utiliser régulièrement ces fichiers standards dans le cadre du cours.

Le langage Java a été largement inspiré du langage C et de nombreuses constructions syntaxiques sont similaires en Java et en C. Un grand nombre de mots clés en C ont le même rôle qu'en Java. Les principaux types de données primitifs supportés par le C sont :
– `int` et `long` : utilisés lors de la déclaration d'une variable de type entier
– `char` : utilisé lors de la déclaration d'une variable permettant de stocker un caractère
– `double` et `float` : utilisés lors de la déclaration d'une variable permettant de stocker un nombre représenté en virgule flottante.
Notez que dans les premières versions du langage C, contrairement à Java, il n'y avait pas de type spécifique permettant de représenter un booléen. Dans de nombreux programmes écrits en C, les booléens sont représentés par des entiers et les valeurs booléennes sont définies [2] comme suit.

```
#define false   0
#define true    1
```

Les compilateurs récents qui supportent le type booléen permettent de déclarer des variables de type `bool` et contiennent les définitions suivantes [2] dans le header standard stdbool.h de [C99].

```
#define false   (bool)0
#define true    (bool)1
```

Au-delà des types de données primitifs, Java et C diffèrent et nous aurons l'occasion d'y revenir dans un prochain chapitre. Le langage C n'est pas un langage orienté objet et il n'est donc pas possible de définir d'objet avec des méthodes spécifiques en C. C permet la définition de structures, d'unions et d'énumérations sur lesquelles nous reviendrons.

En Java, les chaînes de caractères sont représentées grâce à l'objet `String`. En C, une chaîne de caractères est représentée sous la forme d'un tableau de caractères dont le dernier élément contient la valeur `\0`. Alors que Java stocke les chaînes de caractères dans un objet avec une indication de leur longueur, en C il n'y a pas de longueur explicite pour les chaînes de caractères mais le caractère `\0` sert de marqueur de fin de chaîne de caractères. Lorsque le langage C a été développé, ce choix semblait pertinent, notamment pour des raisons de performance. Avec le recul, ce choix pose question [Kamp2011] et nous y reviendrons lorsque nous aborderons certains problèmes de sécurité.

```
char string[10];
string[0] = 'j';
string[1] = 'a';
string[2] = 'v';
string[3] = 'a';
string[4] = '\0';
printf("String : %s\n", string);
```

2. Formellement, le standard [C99] ne définit pas de type `bool` mais un type `_Bool` qui est en pratique renommé en type `bool` dans la plupart des compilateurs. La définition précise et complète se trouve dans stdbool.h

L'exemple ci-dessus illustre l'utilisation d'un tableau de caractères pour stocker une chaîne de caractères. Lors de son exécution, ce fragment de code affiche `String : java` sur la sortie standard. Le caractère spécial `\n` indique un passage à la ligne. printf(3) supporte d'autres caractères spéciaux qui sont décrits dans sa page de manuel.

Au niveau des constructions syntaxiques, on retrouve les mêmes boucles et tests en C et en Java :
– test `if (condition) { ... } else { ... }`
– boucle `while (condition) { ... }`
– boucle `do { ... } while (condition);`
– boucle `for (init; condition; incr) { ... }`
En Java, les conditions sont des expressions qui doivent retourner un résultat de type `boolean`. Le langage C est beaucoup plus permissif puisqu'une condition est une expression qui retourne un nombre entier.

La plupart des expressions et conditions en C s'écrivent de la même façon qu'en Java.

Après ce rapide survol du langage C, revenons à notre programme `src/hello.c`. Tout programme C doit contenir une fonction nommée `main` dont la signature [3] est :

```
int main(int argc, char *argv[])
```

Lorsque le système d'exploitation exécute un programme C compilé, il démarre son exécution par la fonction `main` et passe à cette fonction les arguments fournis en ligne de commande [4]. Comme l'utilisateur peut passer un nombre quelconque d'arguments, il faut que le programme puisse déterminer combien d'arguments sont utilisés. En Java, la méthode `main` a comme signature `public static void main(String args[])` et l'attribut `args.length` permet de connaître le nombres de paramètres passés en arguments d'un programme. En C, le nombre de paramètres est passé dans la variable entière `argc` et le tableau de chaînes de caractères `char *argv[]` contient tous les arguments. Le programme `src/cmdline.c` illustre comment un programme peut accéder à ses arguments.

```
/************************************
 * cmdline.c
 *
 * Programme affichant ses arguments
 * sur la sortie standard
 *
 ************************************/

#include <stdio.h>
#include <stdlib.h>

int main(int argc, char *argv[])
{
  int i;
  printf("Ce programme a %d argument(s)\n", argc);
  for (i = 0; i < argc; i++)
    printf("argument[%d] : %s\n", i, argv[i]);
  return EXIT_SUCCESS;
}
```

Par convention, en C le premier argument (se trouvant à l'indice `0` du tableau `argv`) est le nom du programme qui a été exécuté par l'utilisateur. Une exécution de ce programme est illustrée ci-dessous.

```
Ce programme a 5 argument(s)
argument[0] : ./cmdline
argument[1] : 1
argument[2] : -list
```

3. Il est également possible d'utiliser dans un programme C une fonction `main` qui ne prend pas d'argument. Sa signature sera alors `int main (void)`.

4. En pratique, le système d'exploitation passe également les variables d'environnement à la fonction `main`. Nous verrons plus tard comment ces variables d'environnement sont passées du système au programme et comment celui-ci peut y accéder. Sachez cependant que sous certaines variantes de Unix, et notamment Darwin/MacOS ainsi que sous certaines versions de Windows, le prototype de la fonction `main` inclut explicitement ces variables d'environnement (`int main(int argc, char *argv[], char *envp[])`)

```
argument[3] : abcdef
argument[4] : lepl1503
```

Outre le traitement des arguments, une autre différence importante entre Java et C est la valeur de retour de la fonction `main`. En C, la fonction `main` retourne un entier. Cette valeur de retour est passée par le système d'exploitation au programme (typiquement un *shell* ou interpréteur de commandes) qui a demandé l'exécution du programme. Grâce à cette valeur de retour il est possible à un programme d'indiquer s'il s'est exécuté correctement ou non. Par convention, un programme qui s'exécute sous Unix doit retourner `EXIT_SUCCESS` lorsqu'il se termine correctement et `EXIT_FAILURE` en cas d'échec. La plupart des programmes fournis avec un Unix standard respectent cette convention. Dans certains cas, d'autres valeurs de retour non nulles sont utilisées pour fournir plus d'informations sur la raison de l'échec. En pratique, l'échec d'un programme peut être dû aux arguments incorrects fournis par l'utilisateur ou à des fichiers qui sont inaccessibles.

À titre d'illustration, le programme `src/failure.c` est le programme le plus simple qui échoue lors de son exécution.

```c
/************************************
 * failure.c
 *
 * Programme minimal qui échoue toujours
 *
 ************************************/

#include <stdlib.h>

int main(int argc, char *argv[])
{
  return EXIT_FAILURE;
}
```

Enfin, le dernier point à mentionner concernant notre programme `src/hello.c` est la fonction `printf`. Cette fonction de la librairie standard se retrouve dans la plupart des programmes écrits en C. Elle permet l'affichage de différentes formes de textes sur la sortie standard. Comme toutes les fonctions de la librairie standard, elle est documentée dans sa page de manuel printf(3). printf(3) prend un nombre variable d'arguments. Le premier argument est une chaîne de caractères qui spécifie le format de la chaîne de caractères à afficher. Une présentation détaillée de printf(3) prendrait de nombreuses pages. À titre d'exemple, voici un petit programme utilisant printf(3)

```c
  char weekday[] = "Monday";
  char month[] = "April";
  int day = 1;
  int hour = 12;
  int min = 42;
  char str[] = "SINF1252";
  int i;

  // affichage de la date et l'heure
  printf("%s, %s %d, %d:%d\n", weekday, month, day, hour, min);

  // affichage de la valeur de PI
  printf("PI = %f\n", 4 * atan(1.0));

  // affichage d'un caractère par ligne
  for(i = 0; str[i] != '\0'; i++)
    printf("%c\n", str[i]);
```

Lors de son exécution, ce programme affiche :

```
Monday, April 1, 12:42
PI = 3.141593
L
E
P
```

```
L
1
5
0
3
```

Le langage C permet bien entendu la définition de fonctions. Outre la fonction `main` qui doit être présente dans tout programme, le langage C permet la définition de fonctions qui retournent ou non une valeur. En C, comme en Java, une fonction de type `void` ne retourne aucun résultat tandis qu'une fonction de type `int` retournera un entier. Le programme ci-dessous présente deux fonctions simples. La première, `usage` ne retourne aucun résultat. Elle affiche un message d'erreur sur la sortie d'erreur standard et termine le programme via exit(2) avec un code de retour indiquant un échec. La seconde, `digit` prend comme argument un caractère et retourne `1` si c'est un chiffre et `0` sinon. Le code de cette fonction peut paraître bizarre à un programmeur habitué à Java. En C, les *char* sont représentés par l'entier qui correspond au caractère dans la table des caractères utilisées (voir **RFC 20** pour une table ASCII simple). Toutes les tables de caractères placent les chiffres `0` à `9` à des positions consécutives. En outre, en C une expression a priori booléenne comme `a < b` est définie comme ayant la valeur `1` si elle est vraie et `0` sinon. Il en va de même pour les expressions qui sont combinées en utilisant `&&` ou `||`. Enfin, les fonctions getchar(3) et putchar(3) sont des fonctions de la librairie standard qui permettent respectivement de lire (écrire) un caractère sur l'entrée (la sortie) standard.

```c
/************************************
 * filterdigit.c
 *
 * Programme qui extrait de l'entrée
 * standard les caractères représentant
 * des chiffres
 ************************************/

#include <stdio.h>
#include <stdlib.h>

// retourne vrai si c est un chiffre, faux sinon
// exemple simplifié, voir isdigit dans la librarire standard
// pour une solution complète
int digit(char c)
{
  return ((c >= '0') && (c <= '9'));
}

// affiche un message d'erreur
void usage()
{
  fprintf(stderr, "Ce programme ne prend pas d'argument\n");
  exit(EXIT_FAILURE);
}

int main(int argc, char *argv[])
{
  char c;

  if (argc > 1)
    usage();

  while ((c = getchar()) != EOF) {
    if (digit(c))
            putchar(c);
  }

  return EXIT_SUCCESS;
}
```

Il existe de nombreux livres consacrés au langage C. La référence la plus classique est [KernighanRitchie1998],

mais certains éléments commencent à dater. Un tutoriel intéressant a été publié par Brian Kernighan [Kernighan]. [King2008] propose une présentation plus moderne du langage C.

3.2 Types de données

Dans les sections précédentes, nous avons abordé quelques types de données de base dont les `int` et les `char`. Pour utiliser ces types de données à bon escient, il est important de comprendre en détail la façon dont ils sont supportés par le compilateur et leurs limitations. Celles-ci dépendent souvent de leur représentation en mémoire et durant cette semaine nous allons commencer à analyser de façon plus détaillée comment la mémoire d'un ordinateur est structurée.

3.2.1 Nombres entiers

Toutes les données stockées sur un ordinateur sont représentées sous la forme de séquences de bits. Ces séquences de bits peuvent d'abord permettre de représenter des nombres entiers. Un système informatique peut travailler avec deux types de nombres entiers :
– les nombres entiers signés (`int` notamment en C)
– les nombres entiers non-signés (`unsigned int` notamment en C)
Une séquence de n bits $b_0...b_i...b_{n-1}$ peut représenter le nombre entier $\sum_{i=0}^{n-1} b_i \times 2^i$. Par convention, le bit b_{n-1}, associé au facteur du plus grand indice 2^{n-1}, est appelé le *bit de poids fort* tandis que le bit b_0, associé à 2^0, est appelé le *bit de poids faible*. Les suites de bits sont communément écrites dans l'ordre descendant des indices $b_{n-1}...b_i...b_0$. À titre d'exemple, la suite de bits `0101` correspond à l'entier non signé représentant la valeur cinq. Le bit de poids fort de cette séquence de quatre bits (ou *nibble*) est `0`. Le bit de poids faible est `1`. La table ci-dessous reprend les différentes valeurs décimales correspondant à toutes les séquences de quatre bits consécutifs.

binaire	octal	hexadécimal	décimal
0000	00	0	0
0001	01	1	1
0010	02	2	2
0011	03	3	3
0100	04	4	4
0101	05	5	5
0110	06	6	6
0111	07	7	7
1000	10	8	8
1001	11	9	9
1010	12	A	10
1011	13	B	11
1100	14	C	12
1101	15	D	13
1110	16	E	14
1111	17	F	15

Écrire une séquence de bits sous la forme d'une suite de `0` et de `1` peut s'avérer fastidieux. La représentation décimale traditionnelle n'est pas pratique (optimale) non plus car il faut un ou deux chiffres pour représenter une séquence de quatre bits (ou *nibble*) en fonction de la valeur de ces bits. En pratique, de nombreux systèmes informatiques utilisent une représentation hexadécimale pour afficher des séquences de bits. Cette notation hexadécimale est définie sur base de la table ci-dessus en utilisant des lettres pour représenter chaque séquence de quatre bits dont la valeur numérique est supérieure à `9`. Tout comme avec la représentation décimale habituelle, il est possible d'utiliser la représentation hexadécimale pour de longues séquences de bits. La notation octale est parfois utilisée et est supportée par les compilateurs C. Elle utilise un chiffre pour représenter trois bits consécutifs. A titre d'exemple, voici quelques conversions de nombres en notation décimale vers les notations hexadécimales et binaires.
– L'entier décimal `123` s'écrit `0x7b` en notation hexadécimale et `0b00000000000000000000000001111011` en notation binaire

– L'entier décimal `7654321` s'écrit `0x74cbb1` en notation hexadécimale et `0b00000000001110100110010111010001` en notation binaire

Dans ces exemples, nous avons pris la convention de représentation des nombres en langage C. En C, un nombre décimal s'écrit avec la représentation standard. Un nombre entier en notation hexadécimale est par convention préfixé par `0x` tandis qu'un nombre entier en notation binaire est préfixé par `0b`. Ainsi, les déclarations ci-dessous correspondent toutes à la même valeur.

```
int i;
i = 123;   // décimal
i = 0x7b;  // hexadécimal
i = 0173;  // octal !!
//  i = 0b1111011; // binaire, seulement certains compilateurs
```

Certains compilateurs permettent d'entrer des constantes en binaire directement en préfixant le nombre avec `0b` comme dans `i=0b1111011;`. Cependant, cette notation n'est pas portable sur tous les compilateurs. Elle doit donc être utilisée avec précaution, contrairement à la notation hexadécimale qui fait partie du langage.

Note : Notation octale

La notation octale peut poser des surprises désagréables au programmeur C débutant. En effet, pour des raisons historiques, les compilateurs C considèrent qu'un entier dont le premier chiffre est `0` est écrit en représentation octale et non en représentation décimale.

Ainsi, le fragment de code ci-dessous affichera à l'écran le message `65 et 53 sont différents` car le compilateur C interprète la ligne `j=065;` comme contenant un entier en notation octale et non décimale.

```
int i, j;
i = 65;   // décimal
j = 065;  // octal !!!
if (i == j)
  printf("%d et %d sont égaux\n", i, j);
else
  printf("%d et %d sont différents\n", i, j);
```

Le langage C supporte différents types de données qui permettent de représenter des nombres entiers non signés. Les principaux sont repris dans le tableau ci-dessous.

Type	Explication
`unsigned short`	Nombre entier non signé représenté sur au moins 16 bits
`unsigned int`	Nombre entier non signé représenté sur au moins 16 bits
`unsigned long`	Nombre entier non signé représenté sur au moins 32 bits
`unsigned long long`	Nombre entier non signé représenté sur au moins 64 bits

Le nombre de bits utilisés pour stocker chaque type d'entier non signé peut varier d'une implémentation à l'autre. Le langage C permet de facilement déterminer le nombre de bits utilisés pour stocker un type de données particulier en utilisant l'expression `sizeof`. Appliquée à un type de données, celle-ci retourne le nombre d'octets que ce type occupe. Ainsi, sur de nombreuses plateformes, `sizeof(int)` retournera la valeur `4`.

Les systèmes informatiques doivent également manipuler des nombres entiers négatifs. Cela se fait en utilisant des nombres dits signés. Au niveau binaire, il y a plusieurs approches possibles pour représenter des nombres signés. La première est de réserver le bit de poids fort dans la représentation du nombre pour stocker le signe et stocker la valeur absolue du nombre dans les bits de poids faible. Mathématiquement, un nombre de *n* bits utilisant cette notation pourrait se convertir via la formule $(-1)^{b_{n-1}} \times \sum_{i=0}^{n-2} b_i \times 2^i$.

En pratique, cette notation est rarement utilisée pour les nombres entiers car elle rend la réalisation des circuits électroniques de calcul plus compliquée. Un autre inconvénient de cette notation est qu'elle utilise deux séquences de bits différentes pour représenter la valeur zéro (`00...0` et `10...0`). La représentation la plus courante pour les nombres entiers signés est la notation en *complément à 2*. Avec cette notation, une séquence de *n* bits correspond au nombre entier $-(b_{n-1}) \times 2^{n-1} + \sum_{i=0}^{n-2} b_i \times 2^i$. Avec cette notation, le nombre négatif de 4 bits le plus petit correspond à la valeur `-8`. En notation en complément à deux, il n'y a qu'une seule représentation pour le nombre zéro, la séquence dont tous les bits valent `0`. Par contre, il existe toujours un nombre entier négatif qui n'a pas d'équivalent positif.

binaire	décimal signé
0000	0
0001	1
0010	2
0011	3
0100	4
0101	5
0110	6
0111	7
1000	-8
1001	-7
1010	-6
1011	-5
1100	-4
1101	-3
1110	-2
1111	-1

En C, les types de données utilisés pour représenter des entiers sont signés par défaut. Ils ont la même taille que leurs équivalents non signés et sont repris dans la table ci-dessous.

Type	Explication
`short`	Nombre entier signé représenté sur au moins 16 bits
`int`	Nombre entier signé représenté sur au moins 16 bits
`long`	Nombre entier signé représenté sur au moins 32 bits
`long long`	Nombre entier signé représenté sur au moins 64 bits

Dans de nombreux systèmes Unix, on retrouve dans le fichier stdint.h les définitions des types d'entiers supportés par le système avec le nombre de bits et les valeurs minimales et maximales pour chaque type. La table ci-dessous reprend à titre d'exemple l'information relative aux types `short` (16 bits) et `unsigned int` (32 bits).

Type	Bits	Minimum	Maximum
`short`	16	-32768	32767
`unsigned int`	32	0	4294967295

Le fichier stdint.h contient de nombreuses constantes qui doivent être utilisées lorsque l'on a besoin des valeurs minimales et maximales pour un type donné. Voici à titre d'exemple quelques unes de ces valeurs :

```
#define INT8_MAX         127
#define INT16_MAX        32767
#define INT32_MAX        2147483647
#define INT64_MAX        9223372036854775807LL

#define INT8_MIN          -128
#define INT16_MIN         -32768
#define INT32_MIN        (-INT32_MAX-1)
#define INT64_MIN        (-INT64_MAX-1)

#define UINT8_MAX         255
#define UINT16_MAX        65535
#define UINT32_MAX        4294967295U
#define UINT64_MAX        18446744073709551615ULL
```

L'utilisation d'un nombre fixe de bits pour représenter les entiers peut causer des erreurs dans certains calculs. Par exemple, voici un petit programme qui affiche les 10 premières puissances de cinq et dix.

```
short int i = 1;
unsigned short j = 1;
int n;

printf("\nPuissances de 5 en notation signée\n");
for (n = 1; n < 10; n++) {
```

```
    i = i * 5;
    printf("5^%d=%d\n", n, i);
  }

  printf("\nPuissances de 10 en notation non signée\n");
  for (n = 1; n < 10; n++) {
    j = j * 10;
    printf("10^%d=%d\n", n, j);
  }
```

Lorsqu'il est exécuté, ce programme affiche la sortie suivante.

```
Puissances de 5 en notation signée
5^1=5
5^2=25
5^3=125
5^4=625
5^5=3125
5^6=15625
5^7=12589
5^8=-2591
5^9=-12955

Puissances de 10 en notation non signée
10^1=10
10^2=100
10^3=1000
10^4=10000
10^5=34464
10^6=16960
10^7=38528
10^8=57600
10^9=51712
```

Il est important de noter que le langage C ne contient aucun mécanisme d'exception qui permettrait au programmeur de détecter ce problème à l'exécution. Lorsqu'un programmeur choisit une représentation pour stocker des nombres entiers, il est essentiel qu'il ait en tête l'utilisation qui sera faite de cet entier et les limitations qui découlent du nombre de bits utilisés pour représenter le nombre en mémoire. Si dans de nombreuses applications ces limitations ne sont pas pénalisantes, il existe des applications critiques où un calcul erroné peut avoir des conséquences énormes [Bashar1997].

3.2.2 Nombres réels

Outre les nombres entiers, les systèmes informatiques doivent aussi pouvoir manipuler des nombres réels. Ceux-ci sont également représentés sous la forme d'une séquence fixe de bits. Il existe deux formes de représentation pour les nombres réels :

– la représentation en *simple précision* dans laquelle le nombre réel est stocké sous la forme d'une séquence de 32 bits ;
– la représentation en *double précision* dans laquelle le nombre réel est stocké sous la forme d'une séquence de 64 bits.

La plupart des systèmes informatiques qui permettent de manipuler des nombres réels utilisent le standard IEEE-754. Un nombre réel est représenté en virgule flottante et la séquence de bits correspondante est décomposée en trois parties [5] :

– le bit de poids fort indique le signe du nombre. Par convention, `0` est utilisé pour les nombres positifs et `1` pour les nombres négatifs.
– e bits sont réservés pour stocker l'exposant [6].

5. Source : http ://en.wikipedia.org/wiki/Single-precision_floating-point_format
6. En pratique, le format binaire contient $127 + exp$ en simple précision et non l'exposant exp. Ce choix facilite certaines comparaisons entre nombres représentés en virgule flottante. Une discussion détaillée de la représentation binaire des nombres en virgule flottante sort du cadre de ce cours dédié aux systèmes informatiques. Une bonne référence à ce sujet est [Goldberg1991].

– Les f bits de poids faible servent à stocker la partie fractionnaire du nombre réel.

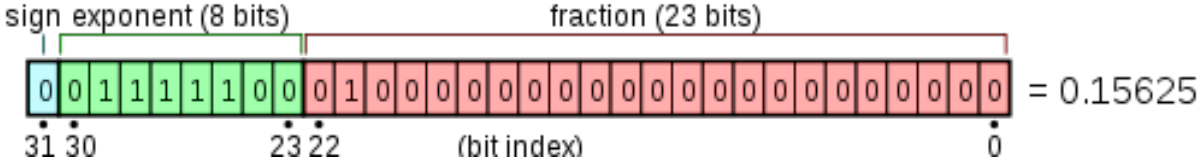


FIGURE 3.1 – Exemple de nombre en virgule flottante (simple précision). (source : Wikipedia)

En simple (resp. double) précision, *8* (resp. *11*) bits sont utilisés pour stocker l'exposant et *23* (resp. *52*) bits pour la partie fractionnaire. L'encodage des nombres réels en simple et double précision a un impact sur la précision de divers algorithmes numériques. Une analyse détaillée de ces problèmes sort du cadre de ce cours, mais il est important de noter deux propriétés importantes de la notation en virgule flottante utilisée actuellement. Ces deux propriétés s'appliquent de la même façon à la simple qu'à la double précision.

– une représentation en virgule flottante sur n bits ne permet jamais de représenter plus de 2^n nombres réels différents ;
– les représentations en virgule flottante privilégient les nombres réels compris dans l'intervalle $[-1, 1]$. On retrouve autant de nombres réels représentables dans cet intervalle que de nombres dont la valeur absolue est supérieure à *1*.

En C, ces nombres en virgule flottante sont représentés en utilisant les types `float` (simple précision) et `double` (double précision). Les fichiers float.h et math.h définissent de nombreuses constantes relatives à ces types. Voici, à titre d'exemple, les valeurs minimales et maximales pour les `float` et les `double` ainsi que les constantes associées. Pour qu'un programme soit portable, il faut utiliser les constantes définies dans float.h et math.h et non leurs valeurs numériques.

```
#define FLT_MIN 1.17549435e-38F
#define FLT_MAX 3.40282347e+38F

#define DBL_MIN 2.2250738585072014e-308
#define DBL_MAX 1.7976931348623157e+308
```

3.2.3 Les tableaux

En langage C, les tableaux permettent d'agréger des données d'un même type. Il est possible de définir des vecteurs et des matrices en utilisant la syntaxe ci-dessous.

```
#define N 10
int vecteur[N];
float matriceC[N][N];
float matriceR[N][2*N];
```

Les premières versions du langage C ne permettaient que la définition de tableaux dont la taille est connue à la compilation. Cette restriction était nécessaire pour permettre au compilateur de réserver la zone mémoire pour stocker le tableau. Face à cette limitation, de nombreux programmeurs définissaient la taille du tableau via une directive `#define` du pré-processeur comme dans l'exemple ci-dessus. Cette directive permet d'associer une chaîne de caractères quelconque à un symbole. Dans l'exemple ci-dessus, la chaîne `10` est associée au symbole `N`. Lors de chaque compilation, le préprocesseur remplace toutes les occurrences de `N` par `10`. Cela permet au compilateur de ne traiter que des tableaux de taille fixe.

Un tableau à une dimension peut s'utiliser avec une syntaxe similaire à celle utilisée par Java. Dans un tableau contenant `n` éléments, le premier se trouve à l'indice `0` et le dernier à l'indice `n-1`. L'exemple ci-dessous présente le calcul de la somme des éléments d'un vecteur.

```
int i;
int sum = 0;
for (i = 0; i < N; i++) {
  sum += v[i];
}
```

Le langage C permet aussi la manipulation de matrices carrées ou rectangulaires qui sont composées d'éléments d'un même type. L'exemple ci-dessous calcule l'élément minimum d'une matrice rectangulaire. Il utilise la constante `FLT_MAX` qui correspond au plus grand nombre réel représentable avec un `float` et qui est définie dans float.h.

```
#define L 2
#define C 3
float matriceR[L][C] = { {1.0,2.0,3.0},
                         {4.0,5.0,6.0} };
int i, j;
float min = FLT_MAX;
for (i = 0; i < L; i++)
  for (j = 0; j < C; j++)
    if (matriceR[i][j] < min)
      min=matriceR[i][j];
```

Les compilateurs récents qui supportent [C99] permettent l'utilisation de tableaux dont la taille n'est connue qu'à l'exécution. Nous en reparlerons ultérieurement.

3.2.4 Caractères et chaînes de caractères

Historiquement, les caractères ont été représentés avec des séquences de bits de différentes longueurs. Les premiers ordinateurs utilisaient des blocs de cinq bits. Ceux-ci permettaient de représenter 32 valeurs différentes. Cinq bits ne permettent pas facilement de représenter à la fois les chiffres et les lettres et les premiers ordinateurs utilisaient différentes astuces pour supporter ces caractères sur 5 bits. Ensuite, des représentations sur six puis sept et huit bits ont été utilisées. Au début des années septante, le code de caractères ASCII sur `7` et `8` bits s'est imposé sur un grand nombre d'ordinateurs et a été utilisé comme standard pour de nombreuses applications et notamment sur Internet **RFC 20**. La table de caractères ASCII définit une correspondance entre des séquences de bits et des caractères. **RFC 20** contient la table des caractères ASCII représentés sur 7 bits. À titre d'exemple, le chiffre *0* correspond à l'octet *0b00110000* et le chiffre *9* à l'octet *0b00111001*. La lettre *a* correspond à l'octet *0b01100001* et la lettre *A* à l'octet *0b01000001*. De nombreux détails sur la table ASCII sont disponibles sur la page Wikipedia : https ://en.wikipedia.org/wiki/ASCII

Les inventeurs du C se sont appuyés sur la table ASCII et ont choisi de représenter un caractère en utilisant un octet. Cela correspond au type `char` que nous avons déjà évoqué.

Concernant le type `char`, il est utile de noter qu'un `char` est considéré en C comme correspondant à un entier. Cela implique qu'il est possible de faire des manipulations numériques sur les caractères. À titre d'exemple, une fonction toupper(3) permettant de transformer un caractère représentant une minuscule dans le caractère représentant la majuscule correspondante peut s'écrire :

```
// conversion de minuscules en majuscules
int toUpper(char c) {
  if (c >= 'a' && c <= 'z')
    return c + ('A' - 'a');
  else
    return c;
}
```

En pratique, l'utilisation de la table ASCII pour représenter des caractères souffre d'une limitation majeure. Avec *7* ou *8* bits il n'est pas possible de représenter exactement tous les caractères écrits de toutes les langues. Une table des caractères sur *7* bits est suffisante pour les langues qui utilisent peu de caractères accentués comme l'anglais. Pour le français et de nombreuses langues en Europe occidentale, la table sur *8* bits est suffisante et la norme ISO-8859 contient des tables de caractères *8* bits pour de nombreuses langues. La norme *Unicode* va plus loin en permettant de représenter les caractères écrits de toutes les langues connues sur Terre. Une description détaillée du support de ces types de caractères sort du cadre de ce cours sur les systèmes informatiques. Il est cependant important que vous soyez conscient de cette problématique pour pouvoir la prendre en compte lorsque vous développerez des applications qui doivent traiter du texte dans différentes langues.

À titre d'exemple, la fonction toupper(3) qui est implémentée dans les versions récentes de Linux est nettement plus complexe que celle que nous avons vue ci-dessus. Tout d'abord, la fonction toupper(3) prend comme argu-

ment un `int` et non un `char`. Cela lui permet d'accepter des caractères dans n'importe quel encodage. Ensuite, le traitement qu'elle effectue dépend du type d'encodage qui a été défini via setlocale(3) (voir locale(7)).

Dans la suite de ce document, nous supposerons qu'un caractère est toujours représentable en utilisant le type `char` permettant de stocker un octet.

En C, les chaînes de caractères sont représentées sous la forme d'un tableau de caractères. Une chaîne de caractères peut être initialisée de différentes façons reprises ci-dessous.

```
char name1[] = { 'U', 'n', 'i', 'x' };
char name2[] = { "Unix" };
char name3[] = "Unix";
```

Lorsque la taille de la chaîne de caractères n'est pas indiquée à l'initialisation (c'est-à-dire dans les deux dernières lignes ci-dessus), le compilateur C la calcule et alloue un tableau permettant de stocker la chaîne de caractères suivie du caractère `\0` qui par convention termine *toujours* les chaînes de caractères en C. En mémoire, la chaîne de caractères correspondant à `name3` occupe donc cinq octets. Les quatre premiers contiennent les caractères *U*, *n*, *i* et *x* et le cinquième le caractère `\0`. Il est important de bien se rappeler cette particularité du langage C car comme nous le verrons plus tard ce choix a de nombreuses conséquences.

Cette particularité permet d'implémenter facilement des fonctions de manipulation de chaînes de caractères. À titre d'exemple, la fonction ci-dessous calcule la longueur d'une chaîne de caractères.

```
int length(char str[])
{
  int i = 0;
  while (str[i] != 0) {// '\0' et 0 sont égaux
    i++;
  }
  return i;
}
```

Contrairement à des langages comme Java, C ne fait aucune vérification sur la façon dont un programme manipule un tableau. En C, il est tout à fait légal d'écrire le programme suivant :

```
char name[5] = "Unix";
printf("%c", name[6]);
printf("%c", name[12345]);
printf("%c", name[-1]);
```

En Java, tous les accès au tableau `name` en dehors de la zone mémoire réservée provoqueraient une `ArrayIndexOutOfBoundsException`. En C, il n'y a pas de mécanisme d'exception et le langage présuppose que lorsqu'un programmeur écrit `name[i]`, il a la garantie que la valeur `i` sera telle qu'il accédera bien à un élément valide du tableau `name`. Ce choix de conception du C permet d'obtenir du code plus efficace qu'avec Java puisque l'interpréteur Java doit vérifier tous les accès à chaque tableau lorsqu'ils sont exécutés. Malheureusement, ce choix de conception du C est aussi à l'origine d'un très grand nombre de problèmes qui affectent la sécurité de nombreux logiciels. Ces problèmes sont connus sous la dénomination *buffer overflow*. Nous aurons l'occasion d'y revenir plus tard.

3.2.5 Les pointeurs

Une différence majeure entre le C et la plupart des langages de programmation actuels est que le C est proche de la machine (langage de bas niveau) et permet au programmeur d'interagir directement avec la mémoire où les données qu'un programme manipule sont stockées. En Java, un programme peut créer autant d'objets qu'il souhaite (ou presque[7]). Ceux-ci sont stockés en mémoire et le *garbage collector* retire de la mémoire les objets qui ne sont plus utilisés. En C, un programme peut aussi réserver des zones pour stocker de l'information en mémoire. Cependant, comme nous le verrons plus tard, c'est le programmeur qui doit explicitement allouer et libérer la mémoire.

7. En pratique, l'espace mémoire accessible à un programme Java est limité par différents facteurs. Voir notamment le paramètre `-Xm` de la machine virtuelle Java http ://docs.oracle.com/javase/6/docs/technotes/tools/solaris/java.html

Les *pointeurs* sont une des caractéristiques principales du langage C par rapport à de nombreux autres langages. Un *pointeur* est défini comme étant une variable contenant l'adresse d'une autre variable. Pour bien comprendre le fonctionnement des pointeurs, il est important d'avoir en tête la façon dont la mémoire est organisée sur un ordinateur. D'un point de vue abstrait, la mémoire d'un ordinateur peut être vue sous la forme d'une zone de stockage dans laquelle il est possible de lire ou d'écrire de l'information. Chaque zone permettant de stocker de l'information est identifiée par une *adresse*. La mémoire peut être vue comme une implémentation de deux fonctions C :
– `data read(addr)` est une fonction qui, sur base d'une adresse, retourne la valeur stockée à cette adresse.
– `void write(addr, data)` est une fonction qui écrit la donnée `data` à l'adresse `addr` en mémoire.
Ces adresses sont stockées sur un nombre fixe de bits qui dépend en général de l'architecture du microprocesseur. Les valeurs les plus courantes aujourd'hui sont *32* et *64*. Par convention, les adresses sont représentées sous la forme d'entiers non signés. Sur la plupart des architectures de processeurs, une adresse correspond à une zone mémoire permettant de stocker un octet. Lorsque nous utiliserons une représentation graphique de la mémoire, nous placerons toujours les adresses numériquement basses en bas de la figure et elles croîtront vers le haut.

Considérons l'initialisation ci-dessous et supposons qu'elle est stockée dans une mémoire où les adresses sont encodées sur *3* bits. Une telle mémoire dispose de huit zones permettant chacune de stocker un octet.

```
char name[] = "Unix";
char c = 'Z';
```

Après exécution de cette initialisation et en supposant que rien d'autre n'est stocké dans cette mémoire, celle-ci contiendra les informations reprises dans la table ci-dessous.

Adresse	Contenu
111	0
110	0
101	Z
100	0
011	x
010	i
001	n
000	U

En langage C, l'expression `&var` permet de récupérer l'adresse à laquelle une variable a été stockée. Appliquée à l'exemple ci-dessus, l'expression `&(name[0])` retournerait la valeur `0b000` tandis que `&c` retournerait la valeur `0b101`.

L'expression `&` peut s'utiliser avec n'importe quel type de donnée. Les adresses de données en mémoire sont rarement affichées, mais quand c'est le cas, on utilise la notation hexadécimale comme dans l'exemple ci-dessous.

```
int i = 1252;
char str[] = "sinf1252";
char c = 'c';

printf("i vaut %d, occupe %ld bytes et est stocké à l'adresse : %p\n",
       i, sizeof(i), &i);
printf("c vaut %c, occupe %ld bytes et est stocké à l'adresse : %p\n",
       c, sizeof(c), &c);
printf("str contient \"%s\" et est stocké à partir de l'adresse : %p\n",
       str, &str);
```

L'exécution de ce fragment de programme produit la sortie suivante.

```
i vaut 1252, occupe 4 bytes et est stocké à l'adresse : 0x7fff89f99cbc
c vaut c, occupe 1 bytes et est stocké à l'adresse : 0x7fff89f99caf
str contient "sinf1252" et est stocké à partir de l'adresse : 0x7fff89f99cb0
```

L'intérêt des pointeurs en C réside dans la possibilité de les utiliser pour accéder et manipuler des données se trouvant en mémoire de façon efficace. En C, chaque pointeur a un type et le type du pointeur indique le type de la donnée qui est stockée dans une zone mémoire particulière. Le type est associé au pointeur lors de la déclaration de celui-ci.

```
int i = 1;       // entier
int *ptr_i;      // pointeur vers un entier
char c = 'Z';    // caractère
char *ptr_c;     // pointeur vers un char
```

Grâce aux pointeurs, il est possible non seulement d'accéder à l'adresse où une donnée est stockée, mais aussi d'accéder à la valeur qui est stockée dans la zone mémoire pointée par le pointeur en utilisant l'expression `*ptr`. Il est également possible d'effectuer des calculs sur les pointeurs comme représenté dans l'exemple ci-dessous.

```
  int i = 1;            // entier
  int *ptr_i;           // pointeur vers un entier
  char str[] = "Unix";
  char *s;              // pointeur vers un char

  ptr_i = &i;
  printf("valeur de i : %d, valeur pointée par ptr_i : %d\n", i, *ptr_i);
  *ptr_i = *ptr_i + 1252;
  printf("valeur de i : %d, valeur pointée par ptr_i : %d\n", i, *ptr_i);
  s = str;
  for (i = 0; i < strlen(str); i++){
    printf("valeur de str[%d] : %c, valeur pointée par *(s+%d) : %c\n",
           i, str[i], i, *(s+i));
  }
```

L'exécution de ce fragment de programme produit la sortie suivante.

```
valeur de i : 1, valeur pointée par ptr_i : 1
valeur de i : 1253, valeur pointée par ptr_i : 1253
valeur de str[0] : U, valeur pointée par *(s+0) : U
valeur de str[1] : n, valeur pointée par *(s+1) : n
valeur de str[2] : i, valeur pointée par *(s+2) : i
valeur de str[3] : x, valeur pointée par *(s+3) : x
```

En pratique en C, les notations `char*` et `char[]` sont équivalentes et l'une peut s'utiliser à la place de l'autre. En utilisant les pointeurs, la fonction de calcul de la longueur d'une chaîne de caractères peut se réécrire comme suit.

```
int length(char *s)
{
  int i = 0;
  while (*(s+i) != '\0')
    i++;
  return i;
}
```

Les pointeurs sont fréquemment utilisés dans les programmes écrits en langage C et il est important de bien comprendre leur fonctionnement. Un point important à bien comprendre est ce que l'on appelle l'*arithmétique des pointeurs*, c'est-à-dire la façon dont les opérations sur les pointeurs sont exécutées en langage C. Pour cela, il est intéressant de considérer la manipulation d'un tableau d'entiers à travers des pointeurs.

```
#define SIZE 3
unsigned int tab[3];
tab[0] = 0x01020304;
tab[1] = 0x05060708;
tab[2] = 0x090A0B0C;
```

En mémoire, ce tableau est stocké en utilisant trois mots consécutifs de 32 bits comme le montre l'exécution du programme ci-dessous :

```
int i;
for (i = 0; i < SIZE; i++) {
  printf("%X est à l'adresse %p\n", tab[i], &(tab[i]));
}
```

```
1020304 est à l'adresse 0x7fff5fbff750
5060708 est à l'adresse 0x7fff5fbff754
90A0B0C est à l'adresse 0x7fff5fbff758
```

La même sortie est produite avec le fragment de programme suivant qui utilise un pointeur.

```
unsigned int* ptr = tab;
for (i = 0; i < SIZE; i++) {
  printf("%X est à l'adresse %p\n", *ptr, ptr);
  ptr++;
}
```

Ce fragment de programme est l'occasion de réfléchir sur la façon dont le C évalue les expressions qui contiennent des pointeurs. La première est l'assignation `ptr=tab`. Lorsque `tab` est déclaré par la ligne `unsigned int tab[3]`, le compilateur considère que `tab` est une constante qui contiendra toujours l'adresse du premier élément du tableau. Il faut noter que puisque `tab` est considéré comme une constante, il est interdit d'en modifier la valeur en utilisant une assignation comme `tab=tab+1`. Le pointeur `ptr`, par contre, correspond à une zone mémoire qui contient une adresse. Il est tout à fait possible d'en modifier la valeur. Ainsi, l'assignation `ptr=tab` (ou `ptr=&(tab[0])`) place dans `ptr` l'adresse du premier élément du tableau. Les pointeurs peuvent aussi être modifiés en utilisant des expressions arithmétiques.

```
ptr = ptr + 1;  // ligne 1
ptr++;          // ligne 2
ptr = ptr - 2;  // ligne 3
```

Après l'exécution de la première ligne, `ptr` va contenir l'adresse de l'élément `1` du tableau `tab` (c'est-à-dire `&(tab[1])`). Ce résultat peut surprendre car si l'élément `tab[0]` se trouve à l'adresse `0x7fff5fbff750` c'est cette adresse qui est stockée dans la zone mémoire correspondant au pointeur `ptr`. On pourrait donc s'attendre à ce que l'expression `ptr+1` retourne plutôt la valeur `0x7fff5fbff751`. Il n'en est rien. En C, lorsque l'on utilise des calculs qui font intervenir des pointeurs, le compilateur prend en compte le type du pointeur qui est utilisé. Comme `ptr` est de type `unsigned int*`, il pointe toujours vers une zone mémoire permettant de stocker un entier non signé sur 32 bits. L'expression `ptr+1` revient en fait à calculer la valeur `ptr+sizeof(unsigned int)` et donc `ptr+1` correspondra à l'adresse `0x7fff5fbff754`. Pour la même raison, l'exécution de la deuxième ligne placera l'adresse `0x7fff5fbff758` dans `ptr`. Enfin, la dernière ligne calculera `0x7fff5fbff758-2*sizeof(unsigned int)`, ce qui correspond à `0x7fff5fbff750`.

Il est intéressant pour terminer cette première discussion de l'arithmétique des pointeurs, de considérer l'exécution du fragment de code ci-dessous.

```
unsigned char* ptr_char = (unsigned char *) tab;
printf("ptr_char contient %p\n", ptr_char);
for (i = 0; i < SIZE + 1; i++) {
  printf("%X est à l'adresse %p\n", *ptr_char, ptr_char);
  ptr_char++;
}
```

L'exécution de ce fragment de code produit une sortie qu'il est intéressant d'analyser.

```
ptr_char contient 0x7fff5fbff750
4 est à l'adresse 0x7fff5fbff750
3 est à l'adresse 0x7fff5fbff751
2 est à l'adresse 0x7fff5fbff752
1 est à l'adresse 0x7fff5fbff753
```

Tout d'abord, l'initialisation du pointeur `ptr_char` a bien stocké dans ce pointeur l'adresse en mémoire du premier élément du tableau. Ensuite, comme `ptr_char` est un pointeur de type `unsigned char *`, l'expression `*ptr_char` a retourné la valeur de l'octet se trouvant à l'adresse `0x7fff5fbff750`. Le pointeur `ptr_char` a été incrémenté en respectant l'arithmétique des pointeurs. Comme `sizeof(unsigned char)` retourne `1`, la valeur stockée dans `ptr_char` a été incrémentée d'une seule unité par l'instruction `ptr_char++`. En analysant les quatre `unsigned char` se trouvant aux adresses `0x7fff5fbff750` à `0x7fff5fbff753`, on retrouve bien l'entier `0x01020304` qui avait été placé dans `tab[0]`.

3.2.6 Les structures

Outre les types de données décrits ci-dessus, les programmes informatiques doivent souvent pouvoir manipuler des données plus complexes. À titre d'exemples, un programme de calcul doit pouvoir traiter des nombres complexes, un programme de gestion des étudiants doit traiter des fiches d'étudiants avec nom, prénom, numéro de matricule,... Dans les langages orientés objet comme Java, cela se fait en encapsulant des données de différents types avec les méthodes permettant leur traitement. C n'étant pas un langage orienté objet, il ne permet pas la création d'objets et de méthodes directement associées. Par contre, C permet de construire des types de données potentiellement complexes.

C permet la définition de structures qui combinent différents types de données simples ou structurés. Contrairement aux langages orientés objet, il n'y a pas de méthode directement associée aux structures qui sont définies. Une structure est uniquement un type de données. Voici quelques exemples de structures simples en C.

```
// structure pour stocker une coordonnée 3D
struct coord {
  int x;
  int y;
  int z;
};

struct coord point = {1, 2, 3};
struct coord p;

// structure pour stocker une fraction
struct fraction {
  int numerator;
  int denominator;
};

struct fraction demi = {1, 2};
struct fraction f;

// structure pour représenter un étudiant
struct student {
  int matricule;
  char prenom[20];
  char nom[30];
};

struct student s = {1, "Linus", "Torvalds"};
```

Le premier bloc définit une structure dénommée `coord` qui contient trois entiers baptisés `x`, `y` et `z`. Dans une structure, chaque élément est identifié par son nom et il est possible d'y accéder directement. La variable `point` est de type `struct coord` et son élément `x` est initialisé à la valeur `1` tandis que son élément `z` est initialisé à la valeur `3`. La variable `p` est également de type `struct coord` mais elle n'est pas explicitement initialisée lors de sa déclaration.

La structure `struct fract` définit une fraction qui est composée de deux entiers qui sont respectivement le numérateur et le dénominateur. La structure `struct student`, elle, définit un type de données qui comprend un numéro de matricule et deux chaînes de caractères.

Les structures permettent de facilement regrouper des données qui sont logiquement reliées entre elles et doivent

être manipulées en même temps. C permet d'accéder facilement à un élément d'une structure en utilisant l'opérateur ' . '. Ainsi, la structure `point` dont nous avons parlé ci-dessus aurait pu être initialisée par les trois expressions ci-dessous :

```
point.x = 1;
point.y = 2;
point.z = 3;
```

Dans les premières versions du langage C, une structure devait nécessairement contenir uniquement des données qui ont une taille fixe, c'est-à-dire des nombres, des caractères, des pointeurs ou des tableaux de taille fixe. Il n'était pas possible de stocker des tableaux de taille variable comme une chaîne de caractères `char []`. Les compilateurs récents [C99] permettent de supporter des tableaux flexibles à l'intérieur de structures. Nous ne les utiliserons cependant pas dans le cadre de ce cours.

Les structures sont utilisées dans différentes librairies et appels système sous Unix et Linux. Un exemple classique est la gestion du temps sur un système Unix. Un système informatique contient généralement une horloge dite *temps-réel* qui est en pratique construite autour d'un cristal qui oscille à une fréquence fixée. Ce cristal est piloté par un circuit électronique qui compte ses oscillations, ce qui permet de mesurer le passage du temps. Le système d'exploitation utilise cette horloge *temps réel* pour diverses fonctions et notamment la mesure du temps du niveau des applications.

Un système de type Unix maintient différentes structures qui sont associées à la mesure du temps [8]. La première sert à mesurer le nombre de secondes et de microsecondes qui se sont écoulées depuis le premier janvier 1970. Cette structure, baptisée `struct timeval` est définie dans sys/time.h comme suit :

```
struct timeval {
  time_t       tv_sec;   /* seconds since Jan. 1, 1970 */
  suseconds_t  tv_usec;  /* and microseconds */
};
```

Cette structure est utilisée par des appels système tels que gettimeofday(2) pour notamment récupérer l'heure courante ou les appels de manipulation de temporisateurs (*timers* en anglais) tels que getitimer(2) / setitimer(2). Elle est aussi utilisée par la fonction time(3posix) de la librairie standard et est très utile pour mesurer les performances d'un programme.

Les structures sont également fréquemment utilisées pour représenter des formats de données spéciaux sur disque comme le format des répertoires [9] ou les formats de paquets qui sont échangés sur le réseau [10].

La définition de `struct timeval` utilise une fonctionnalité fréquemment utilisée du C : la possibilité de définir des alias pour des noms de type de données existants. Cela se fait en utilisant l'opérateur `typedef`. En C, il est possible de renommer des types de données existants. Ainsi, l'exemple ci-dessous utilise `typedef` pour définir `Entier` comme alias pour le type `int` et `Fraction` pour la structure `struct fraction`.

```
// structure pour stocker une fraction
typedef struct fraction {
  double numerator;
  double denominator;
} Fraction ;

typedef int Entier;

int main(int argc, char *argv[])
{
  Fraction demi = {1, 2};
  Entier i = 2;
  // ...
```

8. Une description plus détaillée des différentes possibilités de mesure du temps via les fonctions de la librairie standard est disponible dans le chapitre 21 du manuel de la *libc*.

9. Voir notamment fs(5) pour des exemples relatifs aux systèmes de fichiers. Une analyse détaillée des systèmes de fichiers sort du cadre de ce cours.

10. Parmi les exemples simples, on peut citer la structure `struct ipv6hdr` qui correspond à l'entête du protocole IP version 6 et est définie dans linux/ipv6.h.

```
  return EXIT_SUCCESS;
}
```

Les types `Entier` et `int` peuvent être utilisés de façon interchangeable à l'intérieur du programme une fois qu'ils ont été définis.

Note : `typedef` en pratique

Renommer des types de données a des avantages et des inconvénients dont il faut être conscient pour pouvoir l'utiliser à bon escient. L'utilisation de `typedef` peut faciliter la lecture et la portabilité de certains programmes. Lorsqu'un `typedef` est associé à une structure, cela facilite la déclaration de variables de ce type et permet le cas échéant de modifier la structure de données ultérieurement sans pour autant devoir modifier l'ensemble du programme. Cependant, contrairement aux langages orientés objet, des méthodes ne sont pas directement associées aux structures et la modification d'une structure oblige souvent à vérifier toutes les fonctions qui utilisent cette structure. L'utilisation de `typedef` permet de clarifier le rôle de certains types de données ou valeurs de retour de fonctions. À titre d'exemple, l'appel système read(2) qui permet notamment de lire des données dans un fichier retourne le nombre d'octets qui ont été lus après chaque appel. Cette valeur de retour est de type `ssize_t`. L'utilisation de ces types permet au compilateur de vérifier que les bons types de données sont utilisés lors des appels de fonctions.

`typedef` est souvent utilisé pour avoir des identifiants de types de données plus courts. Par exemple, il est très courant de remplacer le types `unsigned` par les abréviations ci-dessous.

```
typedef unsigned int u_int_t;
typedef unsigned long u_long_t;
```

Soyez prudents si vous utilisez des `typedef` pour redéfinir des pointeurs. En C, il est tout à fait valide d'écrire les lignes suivantes.

```
typedef int * int_ptr;
typedef char * string;
```

Malheureusement, il y a un risque dans un grand programme que le développeur oublie que ces types de données correspondent à des pointeurs qui doivent être manipulés avec soin. Le Linux kernel coding style contient une discussion intéressante sur l'utilisation des `typedef`.

Les pointeurs sont fréquemment utilisés lors de la manipulation de structures. Lorsqu'un pointeur pointe vers une structure, il est utile de pouvoir accéder facilement aux éléments de la structure. Le langage C supporte deux notations pour représenter ces accès aux éléments d'une structure. La première notation est `(*ptr).elem` où `ptr` est un pointeur et `elem` l'identifiant d'un des éléments de la structure pointée par `ptr`. Cette notation est en pratique assez peu utilisée. La notation la plus fréquente est `ptr->elem` dans laquelle `ptr` et `->elem` sont respectivement un pointeur et un identifiant d'élément. L'exemple ci-dessous illustre l'initialisation de deux fractions en utilisant ces notations.

```
struct fraction demi, quart;
struct fraction *demi_ptr;
struct fraction *quart_ptr;

demi_ptr = &demi;
quart_ptr = &quart;

(*demi_ptr).num = 1;
(*demi_ptr).den = 2;

quart_ptr->num = 1;
quart_ptr->den = 4;
```

Les pointeurs sont fréquemment utilisés en combinaison avec des structures et on retrouve très souvent la seconde notation dans des programmes écrits en C.

3.2.7 Les fonctions

Comme la plupart des langages, le C permet de faciliter la compréhension d'un programme en le découpant en de nombreuses fonctions. Chacune réalise une tâche simple. Tout comme Java, C permet la définition de fonctions qui ne retournent aucun résultat. Celles-ci sont de type `void` comme l'exemple trivial ci-dessous.

```
void usage()
{
  printf("Usage : ...\n");
}
```

La plupart des fonctions utiles retournent un résultat qui peut être une donnée d'un des types standard ou une structure. Cette utilisation est similaire à ce que l'on trouve dans des langages comme Java. Il faut cependant être attentif à la façon dont le langage C traite les arguments des fonctions. Le langage C utilise le *passage par valeur* des arguments. Lorsqu'une fonction est exécutée, elle reçoit les valeurs de ces arguments. Ces valeurs sont stockées dans une zone mémoire qui est locale à la fonction. Toute modification faite sur la valeur d'une variable à l'intérieur d'une fonction est donc locale à cette fonction. Les deux fonctions ci-dessous ont le même résultat et aucune des deux n'a d'effet de bord.

```
int twotimes(int n)
{
  return 2 * n;
}

int two_times(int n)
{
  n = 2 * n;
  return n;
}
```

Il faut être nettement plus attentif lorsque l'on écrit des fonctions qui utilisent des pointeurs comme arguments. Lorsqu'une fonction a un argument de type pointeur, celui-ci est passé par valeur, mais connaissant la valeur du pointeur, il est possible à la fonction de modifier le contenu de la zone mémoire pointée par le pointeur. Ceci est illustré par l'exemple ci-dessous.

```
int times_two(int *n)
{
  return (*n) + (*n);
}

int timestwo(int *n)
{
  *n = (*n) + (*n);
  return *n;
}

void f()
{
  int i = 1252;
  printf("i:%d\n", i);
  printf("times_two(&i)=%d\n", times_two(&i));
  printf("après times_two, i:%d\n", i);
  printf("timestwo(&i)=%d\n", timestwo(&i));
  printf("après timestwo, i:%d\n", i);
}
```

Lors de l'exécution de la fonction `f`, le programme ci-dessus affiche à la console la sortie suivante :

```
i:1252
times_two(&i)=2504
après times_two, i:1252
timestwo(&i)=2504
après timestwo, i:2504
```

Cet exemple illustre aussi une contrainte imposée par le langage C sur l'ordre de définition des fonctions. Pour que les fonctions `times_two` et `timestwo` puissent être utilisées à l'intérieur de la fonction `f`, il faut qu'elles aient été préalablement définies. Dans l'exemple ci-dessus, cela s'est fait en plaçant la définition des deux fonctions avant leur utilisation. C'est une règle de bonne pratique utilisable pour de petits programmes composés de quelques fonctions. Pour des programmes plus larges, il est préférable de placer au début du code source la signature des fonctions qui y sont définies. La signature d'une fonction comprend le type de valeur de retour de la fonction, son nom et les types de ses arguments. Généralement, ces déclarations sont regroupées à l'intérieur d'un *fichier header* dont le nom se termine par `.h`.

```
int times_two(int *);
int timestwo(int *);
```

Les fonctions peuvent évidemment recevoir également des tableaux comme arguments. Cela permet par exemple d'implémenter une fonction qui calcule la longueur d'une chaîne de caractères en itérant dessus jusqu'à trouver le caractère de fin de chaîne.

```
int length(char *s)
{
  int i = 0;
  while (*(s+i) != '\0')
    i++;
  return i;
}
```

Tout comme cette fonction peut accéder au *ième* caractère de la chaîne passée en argument, elle peut également et sans aucune restriction modifier chacun des caractères de cette chaîne. Par contre, comme le pointeur vers la chaîne de caractères est passé par valeur, la fonction ne peut pas modifier la zone mémoire qui est pointée par l'argument.

Un autre exemple de fonctions qui manipulent les tableaux sont des fonctions mathématiques qui traitent des vecteurs par exemple.

```
void plusun(int size, int *v)
{
  int i;
  for (i = 0; i < size; i++)
    v[i]++;
}

void print_vecteur(int size, int*v) {
  int i;
  printf("v={");
  for (i = 0;i < size - 1; i++)
    printf("%d,", v[i]);

  if (size > 0)
    printf("%d}", v[size - 1]);
  else
    printf("}");
}
```

Ces deux fonctions peuvent être utilisées par le fragment de code ci-dessous :

```
int vecteur[N] = {1, 2, 3, 4, 5};
plusun(N, vecteur);
print_vecteur(N, vecteur);
```

Note : Attention à la permissivité du compilateur C

Certains langages comme Java sont fortement typés et le compilateur contient de nombreuses vérifications, notamment sur les types de données utilisés, qui permettent d'éviter un grand nombre d'erreurs. Le langage C est

lui nettement plus libéral. Les premiers compilateurs C étaient très permissifs notamment sur les types de données passés en arguments. Ainsi, un ancien compilateur C accepterait probablement sans broncher les appels suivants :

```
plusun(vecteur,N);
print_vecteur(N,vecteur);
```

Dans ce fragment de programme, l'appel à `print_vecteur` est tout à fait valide. Par contre, l'appel à `plusun` est lui erroné puisque le premier argument est un tableau d'entiers (ou plus précisément un pointeur vers le premier élément d'un tableau d'entiers) alors que la fonction `plusun` attend un entier. Inversement, le second argument est un entier à la place d'un tableau d'entiers. Cette erreur n'empêche pas le compilateur gcc(1) de compiler le programme correspondant. Il émet cependant le *warning* suivant :

```
warning: passing argument 1 of 'plusun' makes integer from pointer without a cast
warning: passing argument 2 of 'plusun' makes pointer from integer without a cast
```

De nombreux programmeurs débutants ignorent souvent les warnings émis par le compilateur et se contentent d'avoir un programme compilé. C'est la source de nombreuses erreurs et de nombreux problèmes. Dans l'exemple ci-dessus, l'exécution de l'appel `plusun(vecteur,N)` provoquera une tentative d'accès à la mémoire dans une zone qui n'est pas allouée au processus. Dans ce cas, la tentative d'accès est bloquée par le système et provoque l'arrêt immédiat du programme sur une *segmentation fault*. Dans d'autres cas, des erreurs plus subtiles mais du même type ont provoqué des problèmes graves de sécurité dans des programmes écrits en langage C. Nous y reviendrons ultérieurement.

Pour terminer, mentionnons que les fonctions écrites en C peuvent utiliser des structures et des pointeurs vers des structures comme arguments. Elles peuvent aussi retourner des structures comme résultat. Ceci est illustré par deux variantes de fonctions permettant d'initialiser une fraction et de déterminer si deux fractions sont égales [11].

```
struct fraction init(int num, int den)
{
  struct fraction f;
  f.numerator = num;
  f.denominator = den;
  return f;
}

int equal(struct fraction f1, struct fraction f2)
{
  return ((f1.numerator == f2.numerator)
          && (f1.denominator == f2.denominator));
}

int equalptr(struct fraction *f1, struct fraction *f2)
{
  return ((f1->numerator==f2->numerator)
          && (f1->denominator==f2->denominator));
}

void initptr(struct fraction *f, int num, int den)
{
  f->numerator = num;
  f->denominator = den;
}
```

Considérons d'abord les fonctions `init` et `equal`. `init` est une fonction qui construit une structure sur base d'arguments entiers et retourne la valeur construite. `equal` quant à elle reçoit comme arguments les valeurs de deux structures. Elle peut accéder à tous les éléments des structures passées en argument. Comme ces structures sont passées par valeur, toute modification aux éléments de la structure est locale à la fonction `equal` et n'est pas répercutée sur le code qui a appelé la fonction.

11. Cette définition de l'égalité entre fractions suppose que les fractions à comparer sont sous forme irréductible. Le lecteur est invité à écrire la fonction générale permettant de tester l'égalité entre fractions réductibles.

Les fonctions `initptr` et `equalptr` utilisent toutes les deux des pointeurs vers des `struct fraction` comme arguments. Ce faisant, elles ne peuvent modifier la valeur de ces pointeurs puisqu'ils sont passés comme valeurs. Par contre, les deux fonctions peuvent bien entendu modifier les éléments de la structure qui se trouvent dans la zone de mémoire pointée par le pointeur. C'est ce que `initptr` fait pour initialiser la structure. `equalptr` par contre se contente d'accéder aux éléments des structures passées en argument sans les modifier. Le fragment de code ci-dessous illustre comment ces fonctions peuvent être utilisées en pratique.

```
struct fraction quart;
struct fraction tiers;
quart = init(1, 4);
initptr(&tiers, 1, 3);
printf("equal(tiers,quart)=%d\n", equal(tiers, quart));
printf("equalptr(&tiers,&quart)=%d\n", equalptr(&tiers, &quart));
```

3.2.8 Les expressions de manipulation de bits

La plupart des langages de programmation sont spécialisés dans la manipulation des types de données classiques comme les entiers, les réels et les chaînes de caractères. Comme nous l'avons vu, le langage C permet de traiter ces types de données. En outre, il permet au programmeur de pouvoir facilement manipuler les bits qui se trouvent en mémoire. Pour cela, le langage C définit des expressions qui correspondent à la plupart des opérations de manipulation de bits que l'on retrouve dans les langages d'assemblage. Les premières opérations sont les opérations logiques.

La première opération logique est la négation *négation* (*NOT* en anglais). Elle prend comme argument un bit et retourne le bit inverse. Comme toutes les opérations logiques, elle peut se définir simplement sous la forme d'une table de vérité. Dans des formules mathématiques, la négation est souvent représentée sous la forme $\neg A$.

A	NOT(A)
0	1
1	0

La deuxième opération est la *conjonction logique* (*AND* en anglais). Cette opération prend deux arguments binaires et retourne un résultat binaire. Dans des formules mathématiques, la conjonction logique est souvent représentée sous la forme $A \wedge B$. Elle se définit par la table de vérité suivante :

A	B	A AND B
0	0	0
0	1	0
1	0	0
1	1	1

La troisième opération est la *disjonction logique* (*OR* en anglais). Cette opération prend deux arguments binaires. Dans des formules mathématiques, la disjonction logique est souvent représentée sous la forme $A \vee B$. Elle se définit par la table de vérité suivante.

A	B	A OR B
0	0	0
0	1	1
1	0	1
1	1	1

Enfin, une dernière opération logique intéressante est le *ou exclusif* (*XOR* en anglais). Celle-ci se définit par la table de vérité ci-dessous. Cette opération est parfois représentée mathématiquement comme $A \oplus B$.

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

Ces opérations peuvent être combinées entre elles. Pour des raisons technologiques, les circuits logiques implémentent plutôt les opérations NAND (qui équivaut à AND suivi de NOT) ou NOR (qui équivaut à OR suivi de

NOT). Il est également important de mentionner les lois formulées par De Morgan qui peuvent se résumer par les équations suivantes :
– $\neg(A \wedge B) = \neg A \vee \neg B$
– $\neg(A \vee B) = \neg A \wedge \neg B$
Ces opérations binaires peuvent s'étendre à des séquences de bits. Voici quelques exemples qui permettent d'illustrer ces opérations sur des octets.

```
~ 00000000 = 11111111
11111010 & 01011111 = 01011010
11111010 | 01011111 = 11111111
11111010 ^ 01011111 = 10100101
```

En C, ces expressions logiques s'utilisent comme dans le fragment de code suivant. En général, elles s'utilisent sur des représentations non signées, souvent des `unsigned char` ou des `unsigned int`.

```
r = ~a;     // négation bit à bit
r = a & b; // conjonction bit à bit
r = a | b; // disjonction bit à bit
r = a ^ b; // xor bit à bit
```

En pratique, les opérations logiques sont utiles pour effectuer des manipulations au niveau des bits de données stockées en mémoire. Une utilisation fréquente dans certaines applications réseaux ou systèmes est de forcer certains bits à prendre la valeur `0` ou `1`. La conjonction logique permet de forcer facilement un bit à zéro tandis que la disjonction logique permet de forcer facilement un bit à un. L'exemple ci-dessous montre comment forcer les valeurs de certains bits dans un `unsigned char`. Il peut évidemment se généraliser à des séquences de bits plus longues.

```
r = c & 0x7E; // 0b01111110 force les bits de poids faible et fort à 0
r = d | 0x18; // 0b00011000 force les bits 4 et 3 à 1
```

L'opération XOR joue un rôle important dans certaines applications. La plupart des méthodes de chiffrement et de déchiffrement utilisent de façon extensive cette opération. Une des propriétés intéressantes de l'opération XOR est que $(A \oplus B) \oplus B = A$. Cette propriété est largement utilisée par les méthodes de chiffrement. La méthode développée par Vernam au début du vingtième siècle s'appuie sur l'opération XOR. Pour transmettre un message *M* de façon sûre, elle applique l'opération XOR bit à bit entre tous les bits du message *M* et une clé *K* doit avoir au moins le même nombre de bits que *M*. Si cette clé *K* est totalement aléatoire et n'est utilisée qu'une seule fois, alors on parle de *one-time-pad*. On peut montrer que dans ce cas, la méthode de chiffrement est totalement sûre. En pratique, il est malheureusement difficile d'avoir une clé totalement aléatoire qui soit aussi longue que le message à transmettre. Le programme ci-dessous implémente cette méthode de façon triviale. La fonction memfrob(3) de la librairie *GNU* utilise également un chiffrement via un XOR.

```
int main(int argc, char* argv[])
{
  if (argc != 2)
    usage("ce programme prend une clé comme argument");

  char *key = argv[1];
  char c;
  int i = 0;
  while (((c = getchar()) != EOF) && (i < strlen(key))) {
    putchar(c ^ *(key + i));
    i++;
  }
  return EXIT_SUCCESS;
}
```

Note : Ne pas confondre expressions logiques et opérateurs binaires.

En C, les symboles utilisés pour les expressions logiques (`||` et `&&`) sont très proches de ceux utilisés pour représenter les opérateurs binaires (`|` et `&`). Il arrive parfois qu'un développeur confonde `&` avec `&&`. Malheureusement, le compilateur ne peut pas détecter une telle erreur car dans les deux cas le résultat attendu est généralement du même type.

```
0b0100 & 0b0101 = 0b0100
0b0100 && 0b0101 = 0b0001
0b0100 | 0b0101 = 0b0101
0b0100 || 0b0101 = 0b0001
```

Un autre point important à mentionner concernant les expressions logiques est qu'en C celles-ci sont évaluées de gauche à droite. Cela implique que dans l'expression `( expr1 && expr2 )`, le compilateur C va d'abord évaluer l'expression `expr1`. Si celle-ci est évaluée à la valeur `0`, la seconde expression ne sera pas évaluée. Cela peut être très utile lorsque l'on doit exécuter du code si un pointeur est non-`NULL` et qu'il pointe vers une valeur donnée. Dans ce cas, la condition sera du type `((ptr != NULL) && (ptr->den > 0))`.

Pour terminer, le langage C supporte des expressions permettant le décalage à gauche ou à droite à l'intérieur d'une suite de bits non signée.
– `a = n >> B` décale les bits représentants `n` de `B` bits vers la droite et place le résultat dans la variable `a`.
– `a = n << B` décale les bits représentants `n` de `B` bits vers la gauche et place le résultat dans la variable `a`.
Ces opérations de décalage permettent différentes manipulations de bits. À titre d'exemple, la fonction `int2bin` utilise à la fois des décalages et des masques pour calculer la représentation binaire d'un entier non signé et la placer dans une chaîne de caractères.

```c
#define BITS_INT 32
// str[BITS_INT]
void int2bin(unsigned int num, char *str)
{
  int i;
  str[BITS_INT] = '\0';
  for (i = BITS_INT - 1; i >= 0; i--) {
    if ((num & 1) == 1)
      str[i] = '1';
    else
      str[i] = '0';
    num = num >> 1;
  }
}
```

3.3 Déclarations

Durant les chapitres précédents, nous avons principalement utilisé des variables locales. Celles-ci sont déclarées à l'intérieur des fonctions où elles sont utilisées. La façon dont les variables sont déclarées est importante dans un programme écrit en langage C. Dans cette section nous nous concentrerons sur des programmes C qui sont écrits sous la forme d'un seul fichier source. Nous verrons plus tard comment découper un programme en plusieurs modules qui sont répartis dans des fichiers différents et comment les variables peuvent y être déclarées.

La première notion importante concernant la déclaration des variables est leur *portée*. La portée d'une variable peut être définie comme étant la partie du programme où la variable est accessible et où sa valeur peut être modifiée. Le langage C définit deux types de portée à l'intérieur d'un fichier C. La première est la *portée globale*. Une variable qui est définie en dehors de toute définition de fonction a une portée globale. Une telle variable est accessible dans toutes les fonctions présentes dans le fichier. La variable `g` dans l'exemple ci-dessous a une portée globale.

```c
float g;   // variable globale

int f(int i) {
int n;   // variable locale
// ...
for(int j=0;j<n;j++) {  // variable locale
  // ...
  }
//...
for(int j=0;j<n;j++) {  // variable locale
  // ...
```

```
  }
}
```

Dans un fichier donné, il ne peut évidemment pas y avoir deux variables globales qui ont le même identifiant. Lorsqu'une variable est définie dans un *bloc*, la portée de cette variable est locale à ce bloc. On parle dans ce cas de *portée locale*. La variable locale n'existe pas avant le début du bloc et n'existe plus à la fin du bloc. Contrairement aux identifiants de variables globales qui doivent être uniques à l'intérieur d'un fichier, il est possible d'avoir plusieurs variables locales qui ont le même identifiant à l'intérieur d'un fichier. C'est fréquent notamment pour les définitions d'arguments de fonction et les variables de boucles. Dans l'exemple ci-dessus, les variables `n` et `j` ont une portée locale. La variable `j` est définie dans deux blocs différents à l'intérieur de la fonction `f`.

Le programme `/C/S3-src/portee.c` illustre la façon dont le compilateur C gère la portée de différentes variables.

```
int g1;
int g2=1;

void f(int i) {
  int loc;    //def1a
  int loc2=2; //def2a
  int g2=-i*i;
  g1++;

  printf("[f-%da] \t\t %d \t %d \t %d \t %d\n",i,g1,g2,loc,loc2);
  loc=i*i;
  g1++;
  g2++;
  printf("[f-%db] \t\t %d \t %d \t %d \t %d\n",i,g1,g2,loc,loc2);

}


int main(int argc, char *argv[]) {
  int loc; //def1b
  int loc2=1; //def2b

  printf("Valeurs de : \t g1 \t g2\t loc\t loc2\n");
  printf("================================================\n");

  printf("[main1] \t %d \t %d \t %d \t %d\n",g1,g2,loc,loc2);

  loc=1252;
  loc2=1234;
  g1=g1+1;
  g1=g1+2;

  printf("[main2] \t %d \t %d \t %d \t %d\n",g1,g2,loc,loc2);

  for(int i=1;i<3;i++) {
    int loc=i; //def1c
    int g2=-i;
    loc++;
    g1=g1*2;
    f(i);
    printf("[main-for-%d] \t %d \t %d \t %d \t %d\n",i,g1,g2,loc,loc2);
  }
  f(7);
  g1=g1*3;
  g2=g2+2;
  printf("[main3] \t %d \t %d \t %d \t %d\n",g1,g2,loc,loc2);

  return(EXIT_SUCCESS);
}
```

Ce programme contient deux variables qui ont une portée globale : `g1` et `g2`. Ces deux variables sont définies en dehors de tout bloc. En pratique, elles sont généralement déclarées au début du fichier, même si le compilateur C accepte un définition en dehors de tout bloc et donc par exemple en fin de fichier. La variable globale `g1` n'est définie qu'une seule fois. Par contre, la variable `g2` est définie avec une portée globale et est redéfinie à l'intérieur de la fonction `f` ainsi que dans la boucle `for` de la fonction `main`. Redéfinir une variable globale de cette façon n'est pas du tout une bonne pratique, mais cela peut arriver lorsque par mégarde on importe un fichier header qui contient une définition de variable globale. Dans ce cas, le compilateur C utilise la variable qui est définie dans le bloc le plus proche. Pour la variable `g2`, c'est donc la variable locale `g2` qui est utilisée à l'intérieur de la boucle `for` ou de la fonction `f`.

Lorsqu'un identifiant de variable locale est utilisé à plusieurs endroits dans un fichier, c'est la définition la plus proche qui est utilisée. L'exécution du programme ci-dessus illustre cette utilisation des variables globales et locales.

```
Valeurs de :           g1           g2          loc          loc2
================================================
[main1]           0           1           0           1
[main2]           3           1        1252        1234
[f-1a]                7           -1           0           2
[f-1b]                8           0           1           2
[main-for-1]          8          -1           2        1234
[f-2a]                17           -4           0           2
[f-2b]                18           -3           4           2
[main-for-2]          18          -2           3        1234
[f-7a]                19           -49           0           2
[f-7b]                20           -48           49           2
[main3]           60           3        1252        1234
```

Note : Utilisation des variables

En pratique, les variables globales doivent être utilisées de façon parcimonieuse et il faut limiter leur utilisation aux données qui doivent être partagées par plusieurs fonctions à l'intérieur d'un programme. Lorsqu'une variable globale a été définie, il est préférable de ne pas réutiliser son identifiant pour une variable locale. Au niveau des variables locales, les premières versions du langage C imposaient leur définition au début des blocs. Les standards récents [C99] autorisent la déclaration de variables juste avant leur première utilisation un peu comme en Java.

Les versions récentes de C [C99] permettent également de définir des variables dont la valeur sera constante durant toute l'exécution du programme. Ces déclarations de ces constants sont préfixées par le mot-clé `const` qui joue le même rôle que le mot clé `final` en Java.

```
// extrait de <math.h>
#define M_PI   3.14159265358979323846264338327950288;

const double pi=3.14159265358979323846264338327950288;

const struct fraction {
  int num;
  int denom;
}  demi={1,2};
```

Il y a deux façons de définir des constantes dans les versions récentes de C [C99]. La première est via la macro `#define` du préprocesseur. Cette macro permet de remplacer une chaîne de caractères (par exemple `M_PI` qui provient de math.h) par un nombre ou une autre chaîne de caractères. Ce remplacement s'effectue avant la compilation. Dans le cas de `M_PI` ci-dessus, le préprocesseur remplace toute les occurrences de cette chaîne de caractères par la valeur numérique de π. Lorsqu'une variable `const` est utilisée, la situation est un peu différente. Le préprocesseur n'intervient pas. Par contre, le compilateur réserve une zone mémoire pour la variable qui a été définie comme constante. Cela a deux avantages par rapport à l'utilisation de `#define`. Premièrement, il est possible de définir comme constante n'importe quel type de données en C, y compris des structures ou des pointeurs alors qu'avec un `#define` on ne peut définir que des nombres ou des chaînes de caractères. Ensuite, comme une `const` est stockée en mémoire, il est possible d'obtenir son adresse et de l'examiner via un *debugger*.

3.4 Unions et énumérations

Les structures que nous avons présentées précédemment permettent de combiner plusieurs données de types primitifs différents entre elles. Outre ces structures (`struct`), le langage C supporte également les `enum` et les `union`. Le mot-clé `enum` est utilisé pour définir un type énuméré, c'est-à-dire un type de donnée qui permet de stocker un nombre fixe de valeurs. Quelques exemples classiques sont repris dans le fragment de programme ci-dessous :

```
// les jours de la semaine
typedef enum  {
      monday, tuesday, wednesday, thursday, friday, saturday, sunday
   } day;

// jeu de carte
typedef enum {
  coeur, carreau, trefle, pique
} carte;

// bits
typedef enum {
        BITRESET = 0,
        BITSET = 1
} bit_t;
```

Le premier `enum` permet de définir le type de données `day` qui contient une valeur énumérée pour chaque jour de la semaine. L'utilisation d'un type énuméré rend le code plus lisible que simplement l'utilisation de constantes définies via le préprocesseur.

```
  bit_t bit=BITRESET;
  day jour=monday;
  if(jour==saturday||jour==sunday)
    printf("Congé\n");
```

En pratique, lors de la définition d'un type énuméré, le compilateur C associe une valeur entière à chacune des valeurs énumérées. Une variable permettant de stocker la valeur d'un type énuméré occupe la même zone mémoire qu'un entier.

Outre les structures, le langage C supporte également les unions. Alors qu'une structure permet de stocker plusieurs données dans une même zone mémoire, une `union` permet de réserver une zone mémoire pour stocker une données parmi plusieurs types possibles. Une `union` est parfois utilisée pour minimiser la quantité de mémoire utilisée pour une structure de données qui peut contenir des données de plusieurs types. Pour bien comprendre la différence entre une `union` et une `struct`, considérons l'exemple ci-dessous.

```
struct s_t {
  int i;
  char c;
} s;

union u_t {
  int i;
  char c;
} u;
```

Une union, `u` et une structure, `s` sont déclarées dans ce fragment de programme.

```
  // initialisation
  s.i=1252;
  s.c='A';
  u.i=1252;
  // u contient un int
  u.c='Z';
  // u contient maintenant un char (et u.i est perdu)
```

La structure `s` peut contenir à la fois un entier et un caractère. Par contre, l'`union u`, peut elle contenir un entier (`u.i`) *ou* un caractère (`u.c`), mais jamais les deux en même temps. Le compilateur C alloue la taille pour l'`union` de façon à ce qu'elle puisse contenir le type de donnée se trouvant dans l'`union` nécessitant le plus de mémoire. Si les unions sont utiles dans certains cas très particulier, il faut faire très attention à leur utilisation. Lorsqu'une `union` est utilisée, le compilateur C fait encore moins de vérifications sur les types de données et le code ci-dessous est considéré comme valide par le compilateur :

```
  u.i=1252;
  printf("char : %c\n", u.c);
```

Lors de son exécution, la zone mémoire correspondant à l'union `u` sera simplement interprétée comme contenant un `char`, même si on vient d'y stocker un entier. En pratique, lorsqu'une `union` est vraiment nécessaire pour des raisons d'économie de mémoire, on préférera la placer dans une `struct` en utilisant un type énuméré qui permet de spécifier le type de données qui est présent dans l'`union`.

```
typedef enum { INTEGER, CHAR } Type;

typedef struct
{
  Type type;
  union {
  int i;
  char c;
  } x;
} Value;
```

Le programmeur pourra alors utiliser cette structure en indiquant explicitement le type de données qui y est actuellement stocké comme suit.

```
Value v;
v.type=INTEGER;
v.x.i=1252;
```

3.5 Organisation de la mémoire

Lors de l'exécution d'un programme en mémoire, le système d'exploitation charge depuis le système de fichier le programme en langage machine et le place à un endroit convenu en mémoire. Lorsqu'un programme s'exécute sur un système Unix, la mémoire peut être vue comme étant divisée en six zones principales. Ces zones sont représentées schématiquement dans la figure ci-dessous.

La figure ci-dessus présente une vision schématique de la façon dont un processus Linux est organisé en mémoire centrale. Il y a d'abord une partie de la mémoire qui est réservée au système d'exploitation (OS dans la figure). Cette zone est représentée en grisé dans la figure.

3.5.1 Le segment text

La première zone est appelée par convention le *segment text*. Cette zone se situe dans la partie basse de la mémoire [12]. C'est dans cette zone que sont stockées toutes les instructions qui sont exécutées par le micro-processeur. Elle est généralement considérée par le système d'exploitation comme étant uniquement accessible en lecture. Si un programme tente de modifier son *segment text*, il sera immédiatement interrompu par le système d'exploitation. C'est dans le segment text que l'on retrouvera les instructions de langage machine correspondant aux fonctions de calcul et d'affichage du programme. Nous en reparlerons lorsque nous présenterons le fonctionnement du langage d'assemblage.

12. Dans de nombreuses variantes de Unix, il est possible de connaître le sommet du segment *text* d'un processus grâce à la variable *etext*. Cette variable, de type `char` est initialisée par le système au chargement du processus. Elle doit être déclarée comme variable de type `extern char etext` et son adresse (`&etext`) correspond au sommet du segment text.

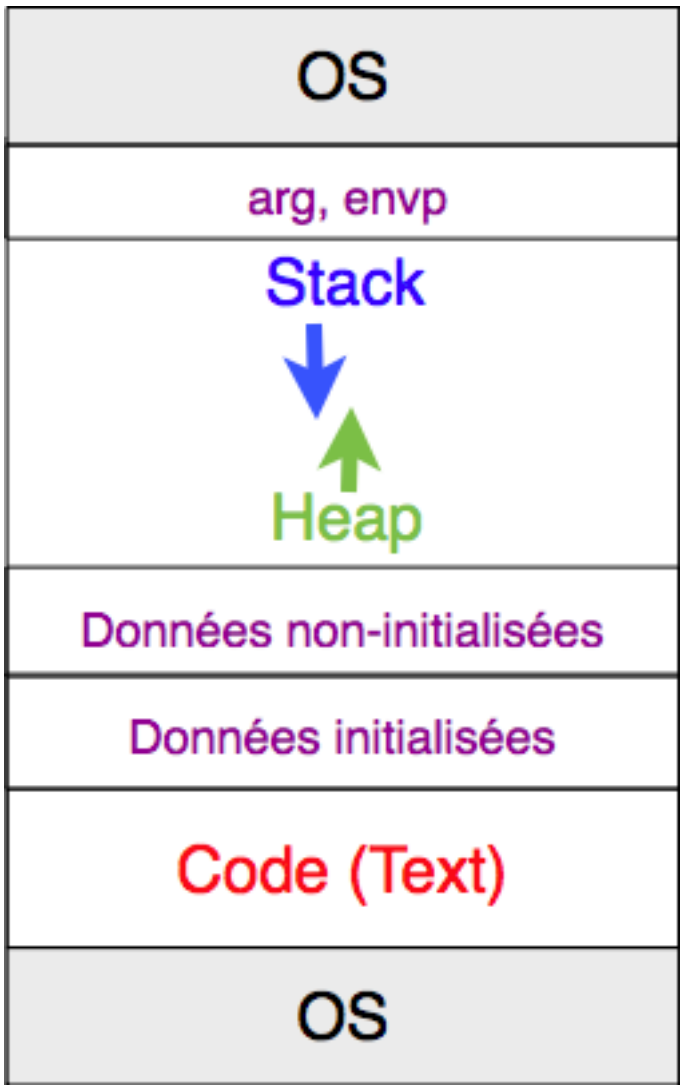


FIGURE 3.2 – Organisation d'un programme Linux en mémoire

3.5.2 Le segment des données initialisées

La deuxième zone, baptisée *segment des données initialisées*, contient l'ensemble des données et chaînes de caractères qui sont utilisées dans le programme. Ce segment contient deux types de données. Tout d'abord, il comprend l'ensemble des variables globales explicitement initialisées par le programme (dans le cas contraire, elles sont initialisées à zéro par le compilateur et appartiennent alors au *segment des données non-initialisées*). Ensuite, les constantes et les chaînes de caractères utilisées par le programme.

```
#define MSG_LEN 10
int g;   // initialisé par le compilateur
int g_init=1252;
const int un=1;
int tab[3]={1,2,3};
int array[10000];
char cours[]="SINF1252";
char msg[MSG_LEN]; // initialisé par le compilateur

int main(int argc, char *argv[]) {
  int i;
  printf("g est à l'adresse %p et initialisée à %d\n",&g,g);
  printf("msg est à l'adresse %p contient les caractères :",msg);
  for(i=0;i<MSG_LEN;i++)
    printf(" %x",msg[i]);
  printf("\n");
  printf("Cours est à l'adresse %p et contient : %s\n",&cours,cours);
  return(EXIT_SUCCESS);
}
```

Dans le programme ci-dessus, la variable `g_init`, la constante `un` et les tableaux `tab` et `cours` sont dans la zone réservée aux variables initialisées. En pratique, leur valeur d'initialisation sera chargée depuis le fichier exécutable lors de son chargement en mémoire. Il en va de même pour toutes les chaînes de caractères qui sont utilisées comme arguments aux appels à `printf(3)`.

L'exécution de ce programme produit la sortie standard suivante.

```
g est à l'adresse 0x60aeac et initialisée à 0
msg est à l'adresse 0x60aea0 contient les caractères : 0 0 0 0 0 0 0 0 0 0
Cours est à l'adresse 0x601220 et contient : SINF1252
```

Cette sortie illustre bien les adresses où les variables globales sont stockées. La variable globale `msg` fait notamment partie du *segment des données non-initialisées*.

3.5.3 Le segment des données non-initialisées

La troisième zone est le *segment des données non-initialisées*, réservée aux variables non-initialisées. Cette zone mémoire est initialisée à zéro par le système d'exploitation lors du démarrage du programme. Dans l'exemple ci-dessus, c'est dans cette zone que l'on stockera les valeurs de la variable `g` et des tableaux `array` et `msg`.

Note : Initialisation des variables

Un point important auquel tout programmeur C doit faire attention est l'initialisation correcte de l'ensemble des variables utilisées dans un programme. Le compilateur C est nettement plus permissif qu'un compilateur Java et il autorisera l'utilisation de variables avant qu'elles n'aient été explicitement initialisées, ce qui peut donner lieu à des erreurs parfois très difficiles à corriger.

En C, par défaut les variables globales qui ne sont pas explicitement initialisées dans un programme sont initialisées à la valeur zéro par le compilateur. Plus précisément, la zone mémoire qui correspond à chaque variable globale non-explicitement initialisée contiendra des bits valant 0. Pour les variables locales, le langage C n'impose aucune initialisation par défaut au compilateur. Par souci de performance et sachant qu'un programmeur ne devrait jamais utiliser de variable locale non explicitement initialisée, le compilateur C n'initialise pas par défaut la valeur de ces variables. Cela peut avoir des conséquences ennuyeuses comme le montre l'exemple ci-dessous.

```
#define ARRAY_SIZE 1000

// initialise un tableau local
void init(void) {
  long a[ARRAY_SIZE];
  for(int i=0;i<ARRAY_SIZE;i++) {
    a[i]=i;
  }
}

// retourne la somme des éléments
// d'un tableau local
long read(void) {
  long b[ARRAY_SIZE];
  long sum=0;
  for(int i=0;i<ARRAY_SIZE;i++) {
    sum+=b[i];
  }
  return sum;
}
```

Cet extrait de programme contient deux fonctions erronées. La seconde, baptisée `read(void)` déclare un tableau local et retourne la somme des éléments de ce tableau sans l'initialiser. En Java, une telle utilisation d'un tableau non-initialisé serait détectée par le compilateur. En C, elle est malheureusement valide (mais fortement découragée évidemment). La première fonction, `init(void)` se contente d'initialiser un tableau local mais ne retourne aucun résultat. Cette fonction ne sert a priori à rien puisqu'elle n'a aucun effet sur les variables globales et ne retourne aucun résultat. L'exécution de ces fonctions via le fragment de code ci-dessous donne cependant un résultat interpellant.

```
  printf("Résultat de read() avant init(): %ld\n",read());
  init();
  printf("Résultat de read() après init() : %ld\n",read());
```

```
Résultat de read() avant init(): 7392321044987950589
Résultat de read() après init() : 499500
```

3.5.4 Le tas (ou *heap*)

La quatrième zone de la mémoire est le *tas* (ou *heap* en anglais). Vu l'importance pratique de la terminologie anglaise, c'est celle-ci que nous utiliserons dans le cadre de ce document. C'est une des deux zones dans laquelle un programme peut obtenir de la mémoire supplémentaire pour stocker de l'information. Un programme peut y réserver une zone permettant de stocker des données et y associer un pointeur.

Le système d'exploitation mémorise, pour chaque processus en cours d'exécution, la limite supérieure de son *heap*. Le système d'exploitation permet à un processus de modifier la taille de son heap via les appels systèmes brk(2) et sbrk(2). Malheureusement, ces deux appels systèmes se contentent de modifier la limite supérieure du *heap* sans fournir une API permettant au processus d'y allouer efficacement des blocs de mémoire. Rares sont les processus qui utilisent directement brk(2) si ce n'est sous la forme d'un appel à `sbrk(0)` de façon à connaître la limite supérieure actuelle du *heap*.

En C, la plupart des processus allouent et libèrent de la mémoire en utilisant les fonctions malloc(3) et free(3) qui font partie de la librairie standard.

La fonction malloc(3) prend comme argument la taille (en bytes) de la zone mémoire à allouer. La signature de la fonction malloc(3) demande que cette taille soit de type `size_t`, c'est-à-dire le type retourné par l'expression `sizeof`. Il est important de toujours utiliser `sizeof` lors du calcul de la taille d'une zone mémoire à allouer. malloc(3) retourne normalement un pointeur de type `(void *)`. Ce type de pointeur est le type par défaut pour représenter dans un programme C une zone mémoire qui ne pointe pas vers un type de données particulier. En

pratique, un programme va généralement utiliser malloc(3) pour allouer de la mémoire pour stocker différents types de données et le pointeur retourné par malloc(3) sera *casté* dans un pointeur du bon type.

Note : `typecast` en langage C

Comme le langage Java, le langage C supporte des conversions implicites et explicites entre les différents types de données. Ces conversions sont possibles entre les types primitifs et les pointeurs. Nous les rencontrerons régulièrement, par exemple lorsqu'il faut récupérer un pointeur alloué par malloc(3) ou le résultat de `sizeof`. Contrairement au compilateur Java, le compilateur C n'émet pas toujours de message de *warning* lors de l'utilisation de typecast qui risque d'engendrer une perte de précision. Ce problème est illustré par l'exemple suivant avec les nombres.

```
int i=1;
float f=1e20;
double d=1e100;

printf("i [int]: %d, [float]:%f, [double]:%f\n",i,(float)i,(double)i);
printf("f [int]: %d, [float]:%f, [double]:%f\n",(int)f,f,(double)f);
printf("d [int]: %d, [float]:%f, [double]:%f\n",(int)d,(float)d,d);
printf("sizeof -> int:%d float:%d double:%d\n",(int)sizeof(int), (int)sizeof(float), (int)s
```

La fonction de la librairie free(3) est le pendant de malloc(3). Elle permet de libérer la mémoire qui a été allouée par malloc(3). Elle prend comme argument un pointeur dont la valeur a été initialisée par malloc(3) et libère la zone mémoire qui avait été allouée par malloc(3) pour ce pointeur. La valeur du pointeur n'est pas modifiée, mais après libération de la mémoire il n'est évidemment plus possible [13] d'accéder aux données qui étaient stockées dans cette zone.

Le programme ci-dessous illustre l'utilisation de malloc(3) et free(3).

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>


typedef struct fraction {
  int num;
  int den;
} Fraction;

void error(char *msg) {
  fprintf(stderr,"Erreur :%s\n",msg);
  exit(EXIT_FAILURE);
}

int main(int argc, char *argv[]) {
  int size=1;
  if(argc==2)
    size=atoi(argv[1]);

  char * string;
  printf("Valeur du pointeur string avant malloc : %p\n",string);
  string=(char *) malloc((size+1)*sizeof(char));
  if(string==NULL)
    error("malloc(string)");

  printf("Valeur du pointeur string après malloc : %p\n",string);
  int *vector;
  vector=(int *)malloc(size*sizeof(int));
```

13. Pour des raisons de performance, le compilateur C ne génère pas de code permettant de vérifier automatiquement qu'un accès via un pointeur pointe vers une zone de mémoire qui est libre. Il est donc parfois possible d'accéder à une zone mémoire qui a été libérée, mais le programme n'a aucune garantie sur la valeur qu'il y trouvera. Ce genre d'accès à des zones mémoires libérées doit bien entendu être complètement proscrit.

```
  if(vector==NULL)
    error("malloc(vector)");

  Fraction * fract_vect;
  fract_vect=(Fraction *) malloc(size*sizeof(Fraction));
  if(fract_vect==NULL)
    error("malloc(fract_vect)");

  free(string);
  printf("Valeur du pointeur string après free   : %p\n",string);
  string=NULL;
  free(vector);
  vector=NULL;
  free(fract_vect);
  fract_vect=NULL;

  return(EXIT_SUCCESS);
}
```

Ce programme alloue trois zones mémoires. Le pointeur vers la première est sauvé dans le pointeur `string`. Elle est destinée à contenir une chaîne de `size` caractères (avec un caractère supplémentaire pour stocker le caractère `\0` de fin de chaîne). Il y a deux points à remarquer concernant cette allocation. Tout d'abord, le pointeur retourné par malloc(3) est "casté" en un `char *`. Cela indique au compilateur que `string` va bien contenir un pointeur vers une chaîne de caractères. Cette conversion explicite rend le programme plus clair. Ensuite, la valeur de retour de malloc(3) est systématiquement testée. malloc(3) peut en effet retourner `NULL` lorsque la mémoire est remplie. Cela a peu de chance d'arriver dans un programme de test tel que celui-ci, mais tester les valeurs de retour des fonctions de la librairie est une bonne habitude à prendre lorsque l'on programme sous Unix. Le second pointeur, `vector` pointe vers une zone destiné à contenir un tableau d'entiers. Le dernier pointeur, `fract_vect` pointe vers une zone qui pourra stocker un tableau de `Fraction`. Lors de son exécution, le programme affiche la sortie suivante.

```
Valeur du pointeur string avant malloc : 0x7fff5fbfe1d8
Valeur du pointeur string après malloc : 0x100100080
Valeur du pointeur string après free   : 0x100100080
```

Dans cette sortie, on remarque que l'appel à fonction free(3) libère la zone mémoire, mais ne modifie pas la valeur du pointeur correspondant. Le programmeur doit explicitement remettre le pointeur d'une zone mémoire libérée à `NULL`.

Un autre exemple d'utilisation de malloc(3) est la fonction `duplicate` ci-dessous qui permet de retourner une copie d'une chaîne de caractères. Il est important de noter qu'en C la fonction strlen(3) retourne la longueur de la chaîne de caractères passée en argument sans prendre en compte le caractère `\0` qui marque sa fin. C'est la raison pour laquelle malloc(3) doit réserver un bloc de mémoire en plus. Même si généralement les `char` occupent un octet en mémoire, il est préférable d'utiliser explicitement `sizeof(char)` lors du calcul de l'espace mémoire nécessaire pour un type de données.

```
#include <string.h>

char *duplicate(char * str) {
  int i;
  size_t len=strlen(str);
  char *ptr=(char *)malloc(sizeof(char)*(len+1));
  if(ptr!=NULL) {
    for(i=0;i<len+1;i++) {
        *(ptr+i)=*(str+i);
      }
  }
  return ptr;
}
```

malloc(3) et free(3) sont fréquemment utilisés dans des programmes qui manipulent des structures de données dont la taille varie dans le temps. C'est le cas pour les différents sortes de listes chaînées, les piles, les queues, les

arbres, ... L'exemple ci-dessous (`/C/S3-src/stack.c`) illustre une implémentation d'une pile simple en C. Le pointeur vers le sommet de la pile est défini comme une variable globale. Chaque élément de la pile est représenté comme un pointeur vers une structure qui contient un pointeur vers la donnée stockée (dans cet exemple des fractions) et l'élément suivant sur la pile. Les fonctions `push` et `pop` permettent respectivement d'ajouter un élément et de retirer un élément au sommet de la pile. La fonction `push` alloue la mémoire nécessaire avec malloc(3) tandis que la fonction `pop` utilise free(3) pour libérer la mémoire dès qu'un élément est retiré.

```
typedef struct node_t
{
  struct fraction_t *data;
  struct node_t *next;
} node;

struct node_t *stack; // sommet de la pile

// ajoute un élément au sommet de la pile
void push(struct fraction_t *f)
{
  struct node_t *n;
  n=(struct node_t *)malloc(sizeof(struct node_t));
  if(n==NULL)
    exit(EXIT_FAILURE);
  n->data = f;
  n->next = stack;
  stack = n;
}
// retire l'élément au sommet de la pile
struct fraction_t * pop()
{
  if(stack==NULL)
    return NULL;
  // else
  struct fraction_t *r;
  struct node_t *removed=stack;
  r=stack->data;
  stack=stack->next;
  free(removed);
  return (r);
}
```

Ces fonctions peuvent être utilisées pour empiler et dépiler des fractions sur une pile comme dans l'exemple ci-dessous. La fonction `display` permet d'afficher sur *stdout* le contenu de la pile.

```
// affiche le contenu de la pile
void display()
{
  struct node_t *t;
  t = stack;
  while(t!=NULL) {
    if(t->data!=NULL) {
      printf("Item at addr %p  :  Fraction %d/%d   Next %p\n",t,t->data->num,t->data->den,t->next
    }
    else {
      printf("Bas du stack %p\n",t);
    }
    t=t->next;
  }
}

// exemple
int main(int argc, char *argv[]) {

  struct fraction_t demi={1,2};
```

```
  struct fraction_t tiers={1,3};
  struct fraction_t quart={1,4};
  struct fraction_t zero={0,1};

  // initialisation
  stack = (struct node_t *)malloc(sizeof(struct node_t));
  stack->next=NULL;
  stack->data=NULL;

  display();
  push(&zero);
  display();
  push(&demi);
  push(&tiers);
  push(&quart);
  display();

  struct fraction_t *f=pop();
  if(f!=NULL)
    printf("Popped : %d/%d\n",f->num,f->den);

  return(EXIT_SUCCESS);
}
```

Lors de son exécution le programme `/C/S3-src/stack.c` présenté ci-dessus affiche les lignes suivantes sur sa sortie standard.

```
Bas du stack 0x100100080
Item at addr 0x100100090  :  Fraction 0/1   Next 0x100100080
Bas du stack 0x100100080
Item at addr 0x1001000c0  :  Fraction 1/4   Next 0x1001000b0
Item at addr 0x1001000b0  :  Fraction 1/3   Next 0x1001000a0
Item at addr 0x1001000a0  :  Fraction 1/2   Next 0x100100090
Item at addr 0x100100090  :  Fraction 0/1   Next 0x100100080
Bas du stack 0x100100080
Popped : 1/4
```

Le tas (ou *heap*) joue un rôle très important dans les programmes C. Les données qui sont stockées dans cette zone de la mémoire sont accessibles depuis toute fonction qui possède un pointeur vers la zone correspondante

---

Note : Ne comptez jamais sur les free(3) implicites

Un programmeur débutant qui expérimente avec malloc(3) pourrait écrire le code ci-dessous et conclure que comme celui-ci s'exécute correctement, il n'est pas nécessaire d'utiliser free(3). Lors de l'exécution d'un programme, le système d'exploitation réserve de la mémoire pour les différents segments du programme et ajuste si nécessaire cette allocation durant l'exécution du programme. Lorsque le programme se termine, via `return` dans la fonction `main` ou par un appel explicite à exit(2), le système d'exploitation libère tous les segments utilisés par le programme, le text, les données, le tas et la pile. Cela implique que le système d'exploitation effectue un appel implicite à free(3) à la terminaison d'un programme.

```
#define LEN 1024
int main(int argc, char *argv[]) {

  char *str=(char *) malloc(sizeof(char)*LEN);
  for(int i=0;i<LEN-1;i++) {
    *(str+i)='A';
  }
  *(str+LEN)='\0';  // fin de chaîne
  return(EXIT_SUCCESS);
}
```

Un programmeur ne doit cependant *jamais* compter sur cet appel implicite à free(3). Ne pas libérer la mémoire lorsqu'elle n'est plus utilisée est un problème courant qui est généralement baptisé *memory leak*. Ce problème est particulièrement gênant pour les processus tels que les serveurs Internet qui ne se terminent pas ou des processus qui s'exécutent longtemps. Une petite erreur de programmation peut causer un *memory leak* qui peut après quelque temps consommer une grande partie de l'espace mémoire inutilement. Il est important d'être bien attentif à l'utilisation correcte de malloc(3) et de free(3) pour toutes les opérations d'allocation et de libération de la mémoire.

---

malloc(3) est la fonction d'allocation de mémoire la plus fréquemment utilisée [14]. La librairie standard contient cependant d'autres fonctions permettant d'allouer de la mémoire mais aussi de modifier des allocations antérieures. calloc(3) est nettement moins utilisée que malloc(3). Elle a pourtant un avantage majeur par rapport à malloc(3) puisqu'elle initialise à zéro la zone de mémoire allouée. malloc(3) se contente d'allouer la zone de mémoire mais n'effectue aucune initialisation. Cela permet à malloc(3) d'être plus rapide, mais le programmeur ne doit jamais oublier qu'il ne peut pas utiliser malloc(3) sans initialiser la zone mémoire allouée. Cela peut s'observer en pratique avec le programme ci-dessous. Il alloue une zone mémoire pour `v1`, l'initialise puis la libère. Ensuite, le programme alloue une nouvelle zone mémoire pour `v2` et y retrouve les valeurs qu'il avait stocké pour `v1` précédemment. En pratique, n'importe quelle valeur pourrait se trouver dans la zone retournée par *malloc(3)*.

```
#define LEN 1024

int main(int argc, char *argv[]) {

  int *v1;
  int *v2;
  int sum=0;
  v1=(int *)malloc(sizeof(int)*LEN);
  for(int i=0;i<LEN;i++) {
    sum+=*(v1+i);
    *(v1+i)=1252;
  }
  printf("Somme des éléments de v1 : %d\n", sum);
  sum=0;
  free(v1);
  v2=(int *)malloc(sizeof(int)*LEN);
  for(int i=0;i<LEN;i++) {
    sum+=*(v2+i);
  }

  printf("Somme des éléments de v2 : %d\n", sum);
  free(v2);

  return(EXIT_SUCCESS);
}
```

L'exécution du programme ci-dessus affiche le résultat suivant sur la sortie standard. Ceci illustre bien que la fonction malloc(3) n'initialise pas les zones de mémoire qu'elle alloue.

```
Somme des éléments de v1 : 0
Somme des éléments de v2 : 1282048
```

Lors de l'exécution du programme, on remarque que la première zone mémoire retournée par malloc(3) a été initialisée à zéro. C'est souvent le cas en pratique pour des raisons de sécurité, mais ce serait une erreur de faire cette hypothèse dans un programme. Si la zone de mémoire doit être initialisée, la mémoire doit être allouée par calloc(3) ou via une initialisation explicite ou en utilisant des fonctions telles que bzero(3) ou memset(3).

3.5.5 Les arguments et variables d'environnement

Lorsque le système d'exploitation charge un programme Unix en mémoire, il initialise dans le haut de la mémoire une zone qui contient deux types de variables. Cette zone contient tout d'abord les arguments qui ont été passés

14. Il existe différentes alternatives à l'utilisation de malloc(3) pour l'allocation de mémoire comme Hoard ou gperftools

via la ligne de commande. Le système d'exploitation met dans `argc` le nombre d'arguments et place dans `char *argv[]` tous les arguments passés avec dans `argv[0]` le nom du programme qui est exécuté.

Cette zone contient également les variables d'environnement. Ces variables sont généralement relatives à la configuration du système. Leurs valeurs sont définies par l'administrateur système ou l'utilisateur. De nombreuses variables d'environnement sont utilisées dans les systèmes Unix. Elles servent à modifier le comportement de certains programmes. Une liste exhaustive de toutes les variables d'environnement est impossible, mais en voici quelques unes qui sont utiles en pratique [15] :

- `HOSTNAME` : le nom de la machine sur laquelle le programme s'exécute. Ce nom est fixé par l'administrateur système via la commande hostname(1)
- `SHELL` : l'interpréteur de commande utilisé par défaut pour l'utilisateur courant. Cet interpréteur est lancé par le système au démarrage d'une session de l'utilisateur. Il est stocké dans le fichier des mots de passe et peut être modifié par l'utilisateur via la commande passwd(1)
- `USER` : le nom de l'utilisateur courant. Sous Unix, chaque utilisateur est identifié par un numéro d'utilisateur et un nom uniques. Ces identifiants sont fixés par l'administrateur système via la commande passwd(1)
- `HOME` : le répertoire d'accueil de l'utilisateur courant. Ce répertoire d'accueil appartient à l'utilisateur. C'est dans ce répertoire qu'il peut stocker tous ses fichiers.
- `PRINTER` : le nom de l'imprimante par défaut qui est utilisée par la commande lp(1posix)
- `PATH` : cette variable d'environnement contient la liste ordonnée des répertoires que le système parcourt pour trouver un programme à exécuter. Cette liste contient généralement les répertoires dans lesquels le système stocke les exécutables standards, comme `/usr/local/bin:/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/sbin:` ainsi que des répertoires relatifs à des programmes spécialisés comme `/usr/lib/mozart/bin:/opt/python3/bin`. L'utilisateur peut ajouter des répertoires à son `PATH` avec bash(1) en incluant par exemple la commande `PATH=$PATH:$HOME/local/bin:.` dans son fichier `.profile`. Cette commande ajoute au `PATH` par défaut le répertoire `$HOME/local/bin` et le répertoire courant. Par convention, Unix utilise le caractère `.` pour représenter ce répertoire courant.

La librairie standard contient plusieurs fonctions qui permettent de manipuler les variables d'environnement d'un processus. La fonction getenv(3) permet de récupérer la valeur associée à une variable d'environnement. La fonction unsetenv(3) permet de supprimer une variable de l'environnement du programme courant. La fonction setenv(3) permet elle de modifier la valeur d'une variable d'environnement. Cette fonction alloue de la mémoire pour stocker la nouvelle variable d'environnement et peut échouer si il n'y a pas assez de mémoire disponible pour stocker de nouvelles variables d'environnement. Ces fonctions sont utilisées notamment par l'interpréteur de commande mais parfois par des programmes dont le comportement dépend de la valeur de certaines variables d'environnement. Par exemple, la commande man(1) utilise différentes variables d'environnement pour déterminer par exemple où les pages de manuel sont stockées et la langue (variable `LANG`) dans laquelle il faut afficher les pages de manuel.

Le programme ci-dessous illustre brièvement l'utilisation de getenv(3), unsetenv(3) et setenv(3). Outre ces fonctions, il existe également clearenv(3) qui permet d'effacer complètement toutes les variables d'environnement du programme courant et putenv(3) qui était utilisé avant setenv(3).

```
#include <stdio.h>
#include <stdlib.h>

// affiche la valeur de la variable d'environnement var
void print_var(char *var) {
  char *val=getenv(var);
  if(val!=NULL)
    printf("La variable %s a la valeur : %s\n",var,val);
  else
    printf("La variable %s n'a pas été assignée\n",var);
}

int main(int argc, char *argv[]) {
```

15. Il possible de lister les définitions actuelles des variables d'environnement via la commande printenv(1). Les interpréteurs de commande tels que bash(1) permettent de facilement modifier les valeurs de ces variables. La plupart d'entre elles sont initialisées par le système ou via les fichiers qui sont chargés automatiquement au démarrage de l'interpréteur comme `/etc/profile` qui contient les variables fixées par l'administrateur système ou le fichier `.profile` du répertoire d'accueil de l'utilisateur qui contient les variables d'environnement propres à cet utilisateur.

```
  char *old_path=getenv("PATH");

  print_var("PATH");

  if(unsetenv("PATH")!=0) {
    fprintf(stderr,"Erreur unsetenv\n");
    exit(EXIT_FAILURE);
  }

  print_var("PATH");

  if(setenv("PATH",old_path,1)!=0) {
    fprintf(stderr,"Erreur setenv\n");
    exit(EXIT_FAILURE);
  }

  print_var("PATH");

  return(EXIT_SUCCESS);
}
```

3.5.6 La pile (ou stack)

La *pile* ou *stack* en anglais est la dernière zone de mémoire utilisée par un processus. C'est une zone très importante car c'est dans cette zone que le processus va stocker l'ensemble des variables locales mais également les valeurs de retour de toutes les fonctions qui sont appelées. Cette zone est gérée comme une pile, d'où son nom. Pour comprendre son fonctionnement, nous utiliserons le programme `/C/S3-src/fact.c` qui permet de calculer une factorielle de façon récursive.

```
// retourne i*j
int times(int i, int j) {
  int m;
  m=i*j;
  printf("\t[times(%d,%d)] : return(%d)\n",i,j,m);
  return m;
}
// calcul récursif de factorielle
// n>0
int fact(int n) {
  printf("[fact(%d)]: Valeur de n:%d, adresse: %p\n",n,n,&n);
  int f;
  if(n==1) {
    printf("[fact(%d)]: return(1)\n",n);
    return(n);
  }
  printf("[fact(%d)]: appel à fact(%d)\n",n,n-1);
  f=fact(n-1);
  printf("[fact(%d)]: calcul de times(%d,%d)\n",n,n,f);
  f=times(n,f);
  printf("[fact(%d)]: return(%d)\n",n,f);
  return(f);
}

void compute() {
  int nombre=3;
  int f;
  printf("La fonction fact est à l'adresse : %p\n",fact);
  printf("La fonction times est à l'adresse : %p\n",times);
  printf("La variable nombre vaut %d et est à l'adresse %p\n",nombre,&nombre);
  f=fact(nombre);
```

```
  printf("La factorielle de %d vaut %d\n",nombre,f);
}
```

Lors de l'exécution de la fonction `compute()`, le programme ci-dessus produit la sortie suivante.

```
La fonction fact est à l'adresse : 0x100000a0f
La fonction times est à l'adresse : 0x1000009d8
La variable nombre vaut 3 et est à l'adresse 0x7fff5fbfe1ac
[fact(3)]: Valeur de n:3, adresse: 0x7fff5fbfe17c
[fact(3)]: appel à fact(2)
[fact(2)]: Valeur de n:2, adresse: 0x7fff5fbfe14c
[fact(2)]: appel à fact(1)
[fact(1)]: Valeur de n:1, adresse: 0x7fff5fbfe11c
[fact(1)]: return(1)
[fact(2)]: calcul de times(2,1)
        [times(2,1)] : return(2)
[fact(2)]: return(2)
[fact(3)]: calcul de times(3,2)
        [times(3,2)] : return(6)
[fact(3)]: return(6)
La factorielle de 3 vaut 6
```

Il est intéressant d'analyser en détails ce calcul récursif de la factorielle car il illustre bien le fonctionnement du stack et son utilisation.

Tout d'abord, il faut noter que les fonctions `fact` et `times` se trouvent, comme toutes les fonctions définies dans le programme, à l'intérieur du *segment text*. La variable `nombre` quant à elle se trouve sur la pile en haut de la mémoire. Il s'agit d'une variable locale qui est allouée lors de l'exécution de la fonction `compute`. Il en va de même des arguments qui sont passés aux fonctions. Ceux-ci sont également stockés sur la pile. C'est le cas par exemple de l'argument `n` de la fonction `fact`. Lors de l'exécution de l'appel à `fact(3)`, la valeur `3` est stockée sur la pile pour permettre à la fonction `fact` d'y accéder. Ces accès sont relatifs au sommet de la pile comme nous aurons l'occasion de le voir dans la présentation du langage d'assemblage. Le premier appel récursif se fait en calculant la valeur de l'argument (`2`) et en appelant la fonction. L'argument est placé sur la pile, mais à une autre adresse que celle utilisée pour `fact(3)`. Durant son exécution, la fonction `fact(2)` accède à ses variables locales sur la pile sans interférer avec les variables locales de l'exécution de `fact(3)` qui attend le résultat de `fact(2)`. Lorsque `fact(2)` fait l'appel récursif, la valeur de son argument (`1`) est placée sur la pile et l'exécution de `fact(1)` démarre. Celle-ci a comme environnement d'exécution le sommet de la pile qui contient la valeur `1` comme argument et la fonction retourne la valeur `1` à l'exécution de `fact(2)` qui l'avait lancée. Dès la fin de `fact(1)`, `fact(2)` reprend son exécution où elle avait été interrompue et applique la fonction `times` avec `2` et `1` comme arguments. Ces deux arguments sont placés sur le pile et `times` peut y accéder au début de son exécution pour calculer la valeur `2` et retourner le résultat à la fonction qui l'a appelé, c'est-à-dire `fact(2)`. Cette dernière retrouve son environnement d'exécution sur la pile. Elle peut maintenant retourner son résultat à la fonction `fact(3)` qui l'avait appelée. Celle-ci va appeler la fonction `times` avec `3` et `2` comme arguments et finira par retourner la valeur `6`.

La pile joue un rôle essentiel lors de l'exécution de programmes en C puisque toutes les variables locales, y compris celles de la fonction `main` y sont stockées. Comme nous le verrons lorsque nous aborderons le langage assembleur, la pile sert aussi à stocker l'adresse de retour des fonctions. C'est ce qui permet à l'exécution de `fact(2)` de se poursuivre correctement après avoir récupéré la valeur calculée par l'appel à `fact(1)`. L'utilisation de la pile pour stocker les variables locales et les arguments de fonctions a une conséquence importante. Lorsqu'une variable est définie comme argument ou localement à une fonction `f`, elle n'est accessible que durant l'exécution de la fonction `f`. Avant l'exécution de `f` cette variable n'existe pas et si `f` appelle la fonction `g`, la variable définie dans `f` n'est plus accessible à partir de la fonction `g`.

En outre, comme le langage C utilise le passage par valeur, les valeurs des arguments d'une fonction sont copiés sur la pile avant de démarrer l'exécution de cette fonction. Lorsque la fonction prend comme argument un entier, cette copie prend un temps très faible. Par contre, lorsque la fonction prend comme argument une ou plusieurs structures de grand taille, celles-ci doivent être entièrement copiées sur la pile. A titre d'exemple, le programme ci-dessous définit une très grande structure contenant un entier et une zone permettant de stocker un million de caractères. Lors de l'appel à la fonction `sum`, les structures `one` et `two` sont entièrement copiées sur la pile. Comme chaque structure occupe plus d'un million d'octets, cela prend plusieurs centaines de microsecondes. Cette copie est

nécessaire pour respecter le passage par valeur des structures à la fonction `sum`. Celle-ci ne peut pas modifier le contenu des structures qui lui sont passées en argument. Par comparaison, lors de l'appel à `sumptr`, seules les adresses de ces deux structures sont copiées sur la pile. Un appel à `sumptr` prend moins d'une microseconde, mais bien entendu la fonction `sumptr` a accès via les pointeurs passés en argument à toute la zone de mémoire qui leur est associée.

```
#define MILLION 1000000

struct large_t {
  int i;
  char str[MILLION];
};

int sum(struct large_t s1, struct large_t s2) {
  return (s1.i+s2.i);
}

int sumptr(struct large_t *s1, struct large_t *s2) {
  return (s1->i+s2->i);
}

int main(int argc, char *argv[]) {
  struct timeval tvStart, tvEnd;
  int err;
  int n;
  struct large_t one={1,"one"};
  struct large_t two={1,"two"};

  n=sum(one,two);
  n=sumptr(&one,&two);
```

Certaines variantes de Unix et certains compilateurs permettent l'allocation de mémoire sur la pile via la fonction alloca(3). Contrairement à la mémoire allouée par malloc(3) qui doit être explicitement libérée en utilisant free(3), la mémoire allouée par alloca(3) est libérée automatiquement à la fin de l'exécution de la fonction dans laquelle la mémoire a été allouée. Cette façon d'allouer de la mémoire sur la pile n'est pas portable et il est préférable de n'allouer de la mémoire que sur le tas en utilisant malloc(3).

Les versions récentes du C et notamment [C99] permettent d'allouer de façon dynamique un tableau sur la pile. Cette fonctionnalité peut être utile dans certains cas, mais elle peut aussi être la source de nombreuses erreurs et difficultés. Pour bien comprendre ce problème, considérons à nouveau la fonction `duplicate` qui a été définie précédemment en utilisant malloc(3) et des pointeurs.

```
#include <string.h>

char *duplicate(char * str) {
  int i;
  size_t len=strlen(str);
  char *ptr=(char *)malloc(sizeof(char)*(len+1));
  if(ptr!=NULL) {
    for(i=0;i<len+1;i++) {
        *(ptr+i)=*(str+i);
      }
  }
  return ptr;
}
```

Un étudiant pourrait vouloir éviter d'utiliser malloc(3) et écrire plutôt la fonction suivante.

```
char *duplicate2(char * str) {
  int i;
  size_t len=strlen(str);
  char str2[len+1];
  for(i=0;i<len+1;i++) {
```

```
    str2[i]=*(str+i);
  }
  return str2;
}
```

Lors de la compilation, gcc(1) affiche le *warning* `In function 'duplicate2': warning: function returns address of local variable`. Ce warning indique que la ligne `return str2;` retourne l'adresse d'une variable locale qui n'est plus accessible à la fin de la fonction `duplicate2`. En effet, l'utilisation de tableaux alloués dynamiquement sur la pile est équivalent à une utilisation implicite de alloca(3). La déclaration `char str2[len];` est équivalente à `char *str2 =(char *)alloca(len*sizeof(char));` et la zone mémoire allouée sur la pile pour `str2` est libérée lors de l'exécution de `return str2;` puisque toute mémoire allouée sur la pile est implicitement libérée à la fin de l'exécution de la fonction durant laquelle elle a été allouée. Donc, une fonction qui appelle `duplicate2` ne peut pas récupérer les données se trouvant dans la zone mémoire qui a été allouée par `duplicate2`.

3.6 Gestion de la mémoire dynamique

Nous avons vu dans la section précédente comment allouer et libérer de la mémoire dans le *heap* (tas) en utilisant les fonctions de la librairie standard malloc(3) et free(3), ainsi que leurs dérivées.

Pour rappel, les signatures de ces deux fonctions sont les suivantes :

```
void *malloc(size_t size);
void free(void *ptr);
```

malloc(3) renvoie un pointeur vers une zone de mémoire du *heap* de taille *minimum* `size` octets. free(3) permet de libérer une zone mémoire précédemment réservée indiquée par le pointeur `ptr`. Cette dernière fonction a un comportement indéterminé si elle est appelée avec un pointeur ne correspondant pas à une zone mémoire réservée et non encore libérée.

La gestion du *heap* est sous la responsabilité d'un algorithme de *gestion de mémoire dynamique*. L'objectif de cet algorithme est double. Premièrement, il doit retourner des zones réservées qui ne se chevauchent pas entre elles, et contiennent *au moins* le nombre d'octets demandés. Deuxièmement, il doit permettre de *recycler* la mémoire des zones libérées pour pouvoir les utiliser de nouveau pour héberger de nouvelles zones réservées.

Dans cette section, nous étudierons les principes et la mise en œuvre des algorithmes de gestion de mémoire dynamique.

Nous ne couvrirons pas la mise en œuvre de l'appel realloc(3) dans le cadre de ce cours. L'appel calloc(3) peut être mis en œuvre en utilisant malloc(3), ce qui est laissé en exercice.

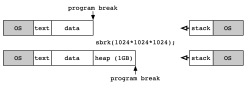
3.6.1 Gestion du heap

L'adresse de départ du *heap* est toujours l'octet qui suit le segment des données non-initialisées. Son adresse de fin est appelée le *program break* sous Linux. Au démarrage d'un programme sous Linux, l'adresse de départ et l'adresse de fin sont identiques. La taille du *heap* initiale est donc de 0.

Traditionnellement sous UNIX, deux appels système permettent de changer la valeur du *program break* et donc d'augmenter la taille du *heap*. Ces deux appels sont définis dans `<unistd.h>` :

```
#include <unistd.h>
int brk(void *addr);
void *sbrk(intptr_t increment);
```

L'appel brk(2) permet de fixer le *program break* a une valeur arbitraire. Sa valeur de retour est 0 lorsque l'opération est un succès, -1 sinon. L'appel sbrk(2) permet d'incrémenter la valeur du *program break* d'un nombre d'octets fourni en argument, et retourne la nouvelle valeur du *program break*. Appeler sbrk(2) avec un argument de 0 permet de lire la valeur actuelle du *program break* sans la modifier. La figure suivante illustre le fonctionnement de l'appel sbrk(2) pour réserver 1 Giga-octet de mémoire pour le *heap*.



Les deux appels brk(2) et sbrk(2) peuvent échouer, en particulier lorsque la valeur demandée pour le *program break* résulte en un dépassement de la taille maximale autorisée pour le programme appelant. Cette taille maximale dépend des paramètres du système et des autorisations de l'utilisateur. On peut connaître ces dernières en utilisant l'utilitaire ulimit(1posix).

En pratique, un programme utilisateur utilise très rarement les appels brk(2) et sbrk(2) mais fait appel aux fonction de l'algorithme de gestion de mémoire dynamique, c'est à dire malloc(3) et free(3). La mise en oeuvre de malloc(3) détermine ainsi quand il est nécessaire d'utiliser sbrk(2) pour étendre la taille de la *heap* afin de répondre à une demande d'allocation, et la mise en oeuvre de free(3) peut de façon similaire décider de réduire la taille du *heap* en appelant sbrk(2) avec un argument négatif.

Note : On notera que des alternatives à sbrk(2) existent pour réserver de la mémoire dynamiquement pour le *heap* d'un programme, et en particulier l'appel système mmap(2) que nous couvrirons lorsque nous aborderons la mise en œuvre de la mémoire virtuelle. C'est la méthode qui est désormais utilisée par Linux, même si sbrk(2) reste supporté pour assurer la compatibilité. Les principes de mise en œuvre de la gestion de la mémoire dynamique présentés ci-dessous sont d'application dans les deux cas.

3.6.2 Contraintes

Un algorithme de gestion de mémoire dynamique obéit aux besoins et contraintes suivants :
– Il est nécessaire de conserver de l'information (des méta-données) sur les blocs alloués et libérés ;
– Le segment *heap* doit être utilisé pour stocker ces méta-données. Les autres segments de la mémoire sont en effet dédiés aux données du programme lui-même (segments *text*, segments de données initialisées et non initialisées, etc.). Il n'est donc possible de stocker les méta-données utilisées par l'algorithme que dans le segment *heap* lui même. Les méta-données doivent donc être *intercalées* avec les zones de mémoire allouées par l'application.

Par ailleurs, il est généralement nécessaire que les zones mémoires allouées soient *alignées*. Cela veut dire que l'adresse de début de chaque zone, ainsi que la taille de la zone, doivent être des multiples d'un *facteur d'alignement* propre au système. Ce facteur est de 8 octets sous Linux. Une zone réservée sera toujours d'une taille multiple du facteur d'alignement. Par exemple, sous Linux, une demande pour 17 octets réservera en réalité 24 octets, le multiple de 8 supérieur le plus proche. On appelle cette extension de la zone demandé le *padding*.

L'alignement permet tout d'abord de faire des hypothèses sur les adresses retournées (les bits de poids faibles sont toujours à 0 : avec un facteur d'alignement de 8 les trois derniers bits des adresses retournées par malloc(3) valent ainsi toujours 0). L'alignement facilite aussi comme nous allons le voir la mise en oeuvre et l'efficacité des algorithmes de gestion de mémoire dynamique. L'exemple ci-dessous illustre l'alignement utilisé par malloc(3) sous Linux.

```
char *a, *b, *c;

a = (char *) malloc(sizeof(char)*1);
b = (char *) malloc(sizeof(char)*9);
c = (char *) malloc(sizeof(char)*1);

printf("Adresse de a : %p.\n",a);
printf("Adresse de b : %p.\n",b);
printf("Adresse de c : %p.\n",c);
```

L'exécution de ce programme produit la sortie standard suivante.

```
Adresse de a : 0x8e29008.
Adresse de b : 0x8e29018.
Adresse de c : 0x8e29028.
```

On peut observer que la zone réservée pour `b` est située 16 octets plus loin que celle pour `a` même si cette dernière ne demandait qu'une zone de 1 octet. Il est intéressant d'analyser ce résultat. On observe tout d'abord que, bien qu'elle soit une multiple de 8, le facteur d'alignement, l'adresse `0x8e29010` n'ait pas été retournée pour `b`, ce

qui aurait pourtant laissé un espace de 8 octets pour `a`. La raison est que la zone nécessaire pour `a` ne contient pas seulement les octets retournés mais aussi des métadonnées nécessaires à l'algorithme de gestion de la mémoire dynamique. La zone réservée pour `c` est elle encore 16 octets plus loin que celle pour `b`, son adresse de démarrage est ainsi alignée sur le facteur d'alignement de 8.

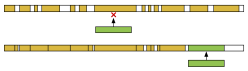
3.6.3 Objectifs

On mesure la qualité d'un algorithme de gestion de mémoire dynamique selon **trois critères** principaux.

Premièrement, les appels aux fonctions malloc(3) et free(3) doivent idéalement s'exécuter le plus rapidement possible, et ce temps d'exécution doit varier le moins possible entre plusieurs appels. Ces fonctions sont effectivement utilisées de manière intensive par de nombreux programmes, et les appels à malloc(3) et free(3) peuvent se trouver dans des chemins de code critiques dont la performance ne doit pas varier au cours du temps ou ne doit pas varier en fonction de la quantité de données manipulées par le programme.

Deuxièmement, l'algorithme doit utiliser la mémoire disponible de manière *efficace*. Il doit pour cela réduire la *fragmentation*. On distingue la fragmentation externe et la fragmentation interne :

– La fragmentation externe mesure à quel point l'espace mémoire complet est fragmenté avec de nombreuses zones libres intercalées entre des zones réservées. On peut voir un exemple de deux heaps dans l'illustration ci-dessous. Les espaces alloués sont représentés en jaune. Dans la heap du haut, on observe que l'espace disponible est fragmenté en de nombreux *trous* entre les espaces alloués. Une requête d'allocation, représentée en vert, ne peut pas être servie car il n'existe pas de trou de taille suffisante pour la placer. Il est donc nécessaire, dans ce cas, d'augmenter la taille du *heap*. En revanche, dans la heap du dessous, l'espace disponible est réparti en de moins nombreux trous et il est possible de répondre à la demande d'allocation directement.



– La fragmentation interne mesure l'espace *perdu* pour chaque allocation, qui n'est pas utilisé pour stocker des donnés. Cela inclut l'espace de *padding*, mais aussi l'espace utilisé pour stocker les métadonnées. Dans l'exemple plus haut, l'espace nécessaire pour la zone `a` de 1 octet demandé fait 16 octets, ce qui résulte en une fragmentation interne de 15 octets.

Note : La défragmentation n'est pas une option

On pourrait être tenté de chercher un mécanisme pour revisiter l'allocation des zones allouées dans le but de réduire la fragmentation, par exemple en décalant les blocs pour éliminer zones vides. Cela n'est malheureusement pas possible dans ce contexte : les pointeurs vers les zones allouées ont déjà été retournés à l'application par malloc(3) et il n'est plus possible de les changer. Il faut donc prendre en compte l'objectif de réduction de la fragmentation dès le départ, lors des appels à malloc(3) et free(3).

Troisièmement, les espaces mémoires réservés par des appels malloc(3) successifs doivent être idéalement proches les uns des autres. Cette propriété de *localité* est importante pour maximiser l'utilisation du cache du processeur, dont l'utilité dépend de cette notion de localité. De façon générale, il est recommandé de suivre le principe : les données allouées de façon proche dans le temps doivent être proches en mémoire, et les données similaires (de même taille) doivent aussi être proches en mémoire car dans les deux cas il y a une probabilité importante qu'elles soient accédées ensemble. Les principes (simplifiés) du fonctionnement d'un cache sont détaillés ci-dessous pour en comprendre les raisons.

Note : Le principe de localité et le cache du processeur

Pour comprendre le principe de localité il nous faut comprendre le principe de cache. Dans notre modèle de système informatique présenté précédemment, nous avons considéré que le processeur effectuait des opérations de lecture et d'écriture directement vers la mémoire principale. L'évolution de la technologie a été telle que désormais la vitesse d'exécution d'un processeur est très largement supérieure à la vitesse à laquelle ou peut accéder à la mémoire : un processeur peut ainsi attendre des centaines de cycles avant de recevoir le résultat d'une opération de lecture en mémoire. Pour pallier ce problème, les processeurs sont équipés de *mémoire cache*. Cette mémoire est plus performante que la mémoire principale : sa latence d'accès est plus faible. Elle est aussi beaucoup plus chère. La mémoire cache ne contient donc qu'un petit sous-ensemble des données utilisées par le programme, sous

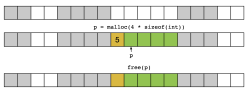
forme de lignes de cache dont la taille est généralement de quelques douzaines d'octets (par exemple, 64 octets). La mémoire principale n'est utilisée que si l'adresse lue n'est pas déjà présente dans le cache. En pratique, une grande partie des accès à la mémoire est servie par le cache grâce à la localité des accès : localité temporelle (une même donnée est lue plusieurs fois dans un intervalle de temps court) et la localité spatiale (si une donnée est lue alors il y a une forte probabilité que la donnée présente dans les octets suivants le soit aussi – par exemple lors du parcours d'une structure de données).

Enfin, il est préférable qu'un algorithme de gestion de la mémoire dynamique soit robuste et qu'il facilite le déboguage. Par exemple, il est préférable que l'on puisse vérifier, lors d'un appel à free(3) que l'adresse soit vérifiable comme étant effectivement une adresse précédemment retournée par un appel à malloc(3).

3.6.4 Algorithmes

Il existe de très nombreux algorithmes de gestion de la mémoire dynamique, dont certains sont très sophistiqués. L'objectif de ce cours n'est pas de les présenter de façon exhaustive mais d'illustrer le compromis entre performance, localité, et efficacité de l'utilisation de la mémoire avec des exemples simples. Ce sujet permet par ailleurs de montrer un exemple concret de la séparation entre mécanisme et politique, typique de la philosophie des systèmes UNIX.

Dans les descriptions ci-dessous, on considèrera que la mémoire est divisée en cases pouvant contenir chacune un mot long, c'est à dire soit un entier soit une adresse. Un *bloc* est composé de plusieurs mots contigus. Pour chaque bloc, il est nécessaire de conserver des méta-données de deux types : la longueur de ce bloc, et un *drapeau* indiquant s'il s'agit d'un bloc réservé ou d'un bloc libre. Une méthode simple pour stocker les méta-données est de réserver un mot dédié sité avant le bloc de données proprement dit, comme l'illustre la figure ci-dessous. Dans cet exemple, le mot de méta-données (*header* en anglais), en jaune, augmente donc de une case l'espace nécessaire pour héberger la zone demandée de 4 blocs, en vert, donc l'adresse retournée sera `p`. Ici la taille du bloc stockée dans le header sera de 5 mots (on indique 5 ici, mais c'est bien `5*sizeof(int*)` qui est stocké), et le bloc sera marquée comme réservée.

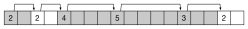


Note : Utilisation du bit de poids faible pour stocker l'état d'un bloc

On note que si la taille des blocs en octets est toujours un multiple de 2 (ce qui est le cas dans notre exemple ou chaque mot fait 8 octets ou 64 bits), alors on a l'assurance que le bit de poids faible sera de valeur 0. On peut tirer partie de cela pour stocker le drapeau indiquant s'il s'agit d'un bloc libre ou d'un bloc réservé : le bit de poids faible peut être positionné à 0 pour indiquer un bloc libre, et positionné à 1 pour indiquer un bloc réservé. On peut alors utiliser les opérations de manipulations de bit pour forcer à 1 ou 0 la valeur de ce bit, et un masque binaire pour lire sa valeur. Ainsi, si `c` est la valeur du compteur stockée dans le bloc on peut obtenir son état en utilisant `c & 0x1`, forcer sa valeur à 1 en utilisant `c = c | 0x1;` ou enfin forcer sa valeur à 0 avec `c = c & ~0x1;`. Attention toutefois, si on souhaite utiliser la valeur du compteur, il faut penser à masquer la valeur du bit de poids faible en lisant `c & ~0x1`.

Utilisation d'une liste implicite

L'algorithme le plus simple utilisant le principe de header dédiés utilise une liste implicite. Cet algorithme peut trouver un bloc libre d'une taille demandé, s'il existe, en suivant un à un les cases de header. Il parcoure ainsi l'ensemble des blocs réservés et libres, comme illustré par la figure ci-dessous.



Le parcours de cette liste peut ressembler alors au pseudo-code suivant, où `p` est un pointeur vers une case mémoire et `start` le début du heap :

```
p = start;
while (p < end &&            // fin de liste ?
       ((*p & 0x1) != 0 ||   // déjà alloué
        *p <= len))          // trou trop petit
    p = p + (*p & ~0x1);     // progresse vers le prochain bloc
```

On notera ici l'utilisation conjointe d'opérateurs binaires (`&`) pour accéder à la valeur du bit de poids faible du header, et des opérateurs logiques (`&&` et `||`). Il ne faut pas les confondre !

Le parcours de la liste va trouver, s'il existe, un espace assez grand pour accueillir la zone dont la création est demandée. Cette zone peut être trop grande et doit alors être scindé en une zone réservée et une nouvelle zone libre, comme illustré par la figure ci-dessous.

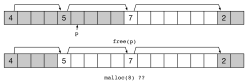
La fonction de placement d'un nouveau bloc de taille `len` au pointeur pointé par `p` retourné précédemment peut ressembler au pseudo-code suivant :

```
void place_block(ptr p, int len) {
  int newsize = len + 1;                // ajoute 1 mot pour le header
  int oldsize = *p & ~0x1;              // récupère taille actuelle sans bit de poids faible
  *p = newsize | 0x1;                   // nouvelle taille avec bit de poids faible à 1
  if (newsize < oldsize)                // s'il reste de la place ...
    *(p + newsize) = oldsize - newsize; // nouveau bloc vide avec la taille restante
}
```

L'exécution de ce code est assez simple. Il faut simplement faire attention ici à ne pas laisser le bit de poids fort pour le header du nouveau bloc. Bien entendu, l'adresse qui est retournée au code client n'est pas `p` mais `p-1` : le header n'a pas a être visible par l'application qui n'a accès qu'au bloc de données proprement dit.

La manière la plus simple de libérer un bloc avec une liste implicite consiste à passer le drapeau, stocké dans le bit de poids faible du header, à 0. Ainsi, un appel à `free(p)` peut simplement exécuter `*(p-1)= *(p-1) & ~0x1;`. Cette méthode naïve présente deux désavantages majeurs :

– Tout d'abord, elle ne vérifie pas que le pointeur passé à free() est un pointeur valide. Il est possible de simplement modifier un bit dans un mot utilisé pour stocker des données, ce qui peut entraîner des bogues particulièrement difficile à découvrir. Il est préférable de stopper l'exécution du programme si free() est appelé pour une adresse invalide. Pour cela toutefois, il est nécessaire de parcourir de nouveau la liste depuis le début pour savoir si l'adresse du header `p-1` est bien valide. Le coût de free() devient donc O(n) où est le nombre de blocs, et plus O(1).
– Ensuite, cette méthode entraîne un problème de *fausse fragmentation*. On peut ainsi plusieurs blocs libres les uns à la suite des autres, mais aucun de ces blocs n'est suffisant pour accueillir une demande d'allocation. Cette situation est illustrée par la figure suivante. Une requête `malloc(8)` arrive, et bien qu'il existe deux blocs de tailles 5 et 7 contiguës, il n'y a pas de bloc de taille 9 disponible.



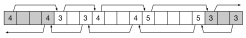
Il y a deux manières de répondre au problème de la fausse fragmentation. Une première serait de laisser les blocs contigus tels quels, et de modifier la fonction de recherche de bloc pour qu'elle prenne en compte leur existence lors de la recherche d'un bloc libre. Cette approche est dite paresseuse (lazy en anglais) car elle procrastine le travail jusqu'au dernier moment. Elle a ses avantages, mais elle complique la mise en œuvre de la fonction malloc(3). Nous laissons son développement en exercice au lecteur.

Une seconde approche est de regrouper les zones libres lors du traitement du free(3). Dans l'exemple de la figure précédente, il est possible de regrouper les deux blocs libres pour en former un seul.

Le pseudo-code de la fonction free_block, prenant en argument l'adresse du *header* à libérer, peut être :

```
void free_block(ptr p) {
  *p = *p & ~0x1;           // remise à 0 du drapeau
  next = p + *p;            // calcul de l'adresse du header du bloc suivant
  if ((*next & 0x1) == 0)   // si ce bloc est aussi libre ...
    *p = *p + *next;        // combiner les blocs
}
```

On remarque toutefois que cette méthode est incomplète. En effet, elle permet de fusionner un bloc libre avec les blocs qui le suive, mais pas avec les blocs qui le précèdent. Ce problème est caractéristique d'une liste chaînée simple : on ne sait simplement pas revenir en arrière. Il y a deux solutions possibles à ce problème. Une première solution est de garder un *historique* des adresses des derniers headers vus lors du parcours (par exemple sur la pile), et de re-parcourir les blocs libres jusqu'à atteindre celui sélectionné, puis ses successeurs. Cette solution impose un surcoût en mémoire lors de l'exécution de free(3) mais ne nécessite pas de modifier Une deuxième méthode, plus simple à mettre en œuvre mais entrainant un surcoût d'utilisation mémoire, est d'utiliser un liste chaînée bi-directionnelle. Cette méthode est illustrée ci-dessous. Elle consiste simplement à répliquer le header avant et après le bloc de données (nécessitant ainsi 2 mots supplémentaires et plus un seul), ce qui permet le parcours, et donc la fusion des blocs libres dans les deux sens. Par exemple ici, la réservation du troisième bloc libre entraînera la fusion avec les deux blocs libres précédents. La figure suivante illustre ce principe.



Politique de placement

L'algorithme de sélection du bloc libre à utiliser présenté et utilisé jusqu'ici est assez basique. L'opération de recherche s'arrête dès qu'un bloc vide de taille *suffisante* pour héberger la nouvelle allocation est trouvée. On appelle cette politique de placement *first fit*. Celle-ci peut prendre un temps linéaire en le nombre de blocs (alloués et libres) mais termine dès qu'un bloc valide est trouvé, et donc assez rapidement en pratique. Toutefois, elle présente des désavantages en pratique, car elle entraîne de la fragmentation est n'a pas de bonnes propriétés de localité. En effet, les allocations de petits objets ont tendance à s'accumuler au début de la zone mémoire, et lorsque ceux-ci sont libérés il subsiste de nombreux petits blocs libres qui ralentissent les recherches ultérieures. Les allocations d'objets de taille plus importante, et en particulier les allocations successives, ne sont pas nécessairement placées dans des zones proches, ce qui entraîne une faible localité.

Il existe d'autres politiques, comme par exemple :
– La politique *next fit* se comporte de la même manière que *first fit* mais démarre la recherche depuis le dernier bloc alloué. Bien entendu, si aucun bloc de taille suffisante n'est trouvé, il est nécessaire de recommencer depuis le début du heap. Cette politique a de meilleures propriétés de localité que *first fit* mais les évaluations montrent qu'elle conduit à une encore plus grande fragmentation.
– La politique *best fit* cherche a répondre à une demande d'allocation avec un bloc donc la taille est la plus proche de celle demandée (juste nécessaire). Cette politique a l'avantage d'être optimale en terme de fragmentation. Toutefois, elle a deux désavantages : elle a potentiellement une mauvaise localité, les blocs libres de taille équivalente n'étant pas nécessairement proches les uns des autres, et elle nécessite un parcours complet de la liste dans tous les cas (sauf, bien sûr, si un bloc de la taille exactement demandée est trouvé).

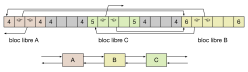
Ces politiques représentent des compromis différents entre les critères que nous avons défini plus haut : rapidité d'exécution, fragmentation de la mémoire, et localité.

Utilisation d'une liste explicite

Nous allons maintenant étudier une autre alternative, qui consiste à créer une liste explicite liant *uniquement* les blocs libres entre eux. Le principe est ici d'utiliser l'espace disponible dans chaque blocs libre pour y stocker des pointeurs vers un prochain bloc libre. Le parcours de la liste chaînée permet donc de ne considérer que les blocs libres, ce qui est plus rapide en particulier lorsque de nombreux blocs sont déjà alloués. La figure suivante présente le principe avec une liste explicite *simplement* chaînée.

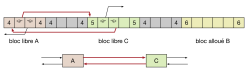
On notera l'utilisation du premier mot pour stocker un pointeur vers un premier bloc libre. Pour chaque bloc libre, deux mots sont utilisés. Le premier mot contient comme auparavant la taille du bloc libre. Le deuxième mot contient un pointeur vers le prochain bloc libre. Le parcours de la liste des blocs libres se fait en suivant ces pointeurs.

En pratique, une liste explicite est toujours une liste *doublement chaînée*, c'est à dire que chaque bloc libre contient un pointeur vers le prochain bloc dans la liste ainsi qu'un pointeur vers le bloc précédent. Par ailleurs, l'ordre des blocs libres dans la liste explicite ne suit pas forcément l'ordre des adresses de début de ces blocs. Une telle liste est illustrée par la figure suivante, qui présente seulement une partie du segment heap.



Les blocs libres A, B et C ne sont pas liés dans cet exemple dans l'ordre de leurs adresses. Il est donc nécessaire de dupliquer le header de bloc vide au début et à la fin de chaque bloc comme expliqué précédemment, afin de permettre la fusion de blocs vides contigus dans les deux sens. L'espace mémoire utilisée (fragmentation interne) est le même que dans le cas d'une liste implicite avec la duplication du header : deux mots supplémentaires sont nécessaires au début et à la fin de chaque bloc alloué. En revanche, quatre mots sont nécessaires pour chaque bloc libre. Ils servent à stocker les pointeurs vers le bloc successeur et prédécesseur de chaque bloc libre. Cela n'a pas d'impact sur la fragmentation interne, puisque ces méta-données sont stockées dans les blocs libres. En revanche, la taille minimum de bloc (libre ou alloué) est désormais de 4 mots. Il est donc nécessaire, dans cet exemple, d'aligner les allocations de blocs pour qu'ils utilisent un multiple de 4 mots, en utilisant du *padding* si nécessaire.

L'allocation d'un bloc de données avec une liste explicite suit le même principe qu'avec une liste implicite, et peut obéir à des politiques similaires. Il est nécessaire, bien entendu, de maintenir les propriétés de la liste doublement chaînée. Ceci nécessite de mettre à jour le pointeur vers le bloc successeur du bloc libre précédent le bloc alloué, et le pointeur vers le bloc prédécesseur du bloc libre suivant le bloc alloué. Ces opérations sont illustrés sur notre exemple dans la figure suivante, où les pointeurs en rouge représentent les pointeurs modifiés.



La libération d'un bloc est plus complexe qu'avec une liste implicite. Cela est du au fait que la liste des blocs libres ne représente pas nécessairement l'ordre des adresses de ces blocs en mémoire. Deux politiques sont possibles pour pallier cette difficulté :

– Une première politique, appelée *address-ordered* consiste à s'assurer que cette propriété d'ordre est assurée. Lors de la libération d'un bloc, celui-ci est inséré dans la liste chaînée exactement entre les deux blocs libres le précédent et le succédant dans l'ordre des adresses. Cette politique permet de fusionner très facilement les zones libres contigus (elles sont soit deux, soit trois, jamais plus). En revanche, elle a un défaut majeur : elle nécessite de parcourir la liste pour chercher l'endroit dans la liste où insérer le bloc vide.
– Une deuxième politique est de simplement insérer le bloc vide au début de la liste des blocs vides (on parle alors de politique LIFO, pour *Last-In-First-Out*). Cette approche a l'avantage indéniable que l'opération de libération est en temps constant. Toutefois, la fusion de zones libres contigus devient plus complexes et nécessite d'opérer des changements dans au plus trois éléments de la liste chaînée et leurs voisins respectifs. Nous n'en couvrirons pas les détails dans le cadre de ce cours.

En règle générale, la politique *address-ordered* entraîne une fragmentation externe moins importante que la politique LIFO. Son principal désavantage réside donc dans son temps d'exécution, qui peut devenir important lorsque de nombreux blocs vides sont présents.

Utilisation de listes multiples

Les mécanismes et politiques de gestion de la mémoire dynamique qui ont été présentées ne font pas de distinction entre les demandes d'allocation de blocs de petite et de grande tailles. Cela pose un inconvénient double. Tout d'abord, le temps de recherche d'un bloc vide augmente linéairement avec le nombre de blocs (total pour une liste implicite, ou seulement libres pour une liste explicite). Par ailleurs, le temps de recherche d'un bloc vide de grande taille peut prendre un temps plus important que la recherche d'un bloc vide de petite taille.

Une solution à ces défauts est d'utiliser non pas une, mais plusieurs listes. Chaque liste correspond alors à une classe de taille. Il peut ainsi y avoir une liste pour les blocs de taille 1, 2, 3, ou 4 octets, puis des listes pour les

tailles correspondant aux puissances de 2 : 8, 16, 32, 64, etc. octets. La recherche d'une zone libre est faite dans la liste correspondant à la taille immédiatement supérieure à la taille demandée : une demande d'allocation pour un bloc de taille 67 octets sera ainsi servi en utilisant la liste des blocs libres de taille 128, c'est à dire parmi les blocs libres de 65 à 128 octets.

On notera que l'utilisation de listes multiples a une synergie avec l'utilisation de listes explicites. Il suffit, en effet, d'utiliser autant de pointeurs de début de liste que de listes, et de parcourir la liste adéquate. Les opérations de réservation et de libération de blocs peuvent entraîner, en revanche, des opérations de recherche sur *plusieurs* listes. Cela rend alors la politique *address-ordered* significativement plus coûteuse que la politique LIFO, qui peut alors être préférée dans ce cas.

3.7 Compléments de C

Dans les sections précédentes, nous n'avons pas pu couvrir l'ensemble des concepts avancés qui sont relatifs à une bonne utilisation du langage C. Cette section contient quelques notions plus avancées qui sont importantes en pratique.

3.7.1 Pointeurs

Les pointeurs sont très largement utilisés dans les programmes écrits en langage C. Nous avons utilisé des pointeurs vers des types de données primitifs tel que les `int`, `char` ou `float` et des pointeurs vers des structures. En pratique, il est possible en C de définir des pointeurs vers n'importe quel type d'information qui est manipulée par un programme C.

Un premier exemple sont les pointeurs vers des fonctions. Comme nous l'avons vu dans le chapitre précédent, une fonction est une séquence d'instructions assembleur qui sont stockées à un endroit bien précis de la mémoire. Cette localisation précise des instructions qui implémentent la fonction permet d'appeler une fonction avec l'instruction `calll`. En C, il est parfois aussi souhaitable de pouvoir appeler une fonction via un pointeur vers cette fonction plutôt qu'en nommant la fonction directement. Cela peut rendre le code plus flexible et plus facile à adapter. Nous avons déjà utilisé des pointeurs vers des fonctions sans le savoir lorsque nous avons utilisé `printf("fct : %p\n",f)` où `f` est un nom de fonction. L'exemple ci-dessous montre une autre utilisation intéressante des pointeurs vers des fonctions. Lorsque l'on écrit du code C, il est parfois utile d'ajouter des commandes qui permettent d'afficher à l'écran des informations de débogage. L'exemple ci-dessous est une application qui supporte trois niveaux de débogage. Rien n'est affiché au niveau `0`, une ligne s'affiche au niveau `1` et des informations plus détaillées sont affichées au niveau `2`. Lors de son exécution , l'application affiche la sortie suivante.

```
$./fctptr 0
fct debug_print : 0x100000d28
$ ./fctptr 1
fct debug_print : 0x100000d32
debug: Hello
$ ./fctptr 2
fct debug_print : 0x100000d5f
debug: Hello
g=1
```

Cette application qui supporte plusieurs niveaux de débogage utilise pourtant toujours le même appel pour afficher l'information de débogage : `(debug_print[debug_level])(...);`. Cet appel profite des pointeurs vers les fonctions. Le tableau `debug_print` est un tableau de pointeurs vers des fonctions qui chacune prend comme argument un `char *`. La variable globale `debug_level` est initialisée sur base de l'argument passé au programme.

```
int g=1;
int debug_level;

void empty (char *str) {
  return;
}
```

```
void oneline(char *str) {
  fprintf(stderr,"debug: %s\n",str);
}

void detailed(char *str) {
  fprintf(stderr, "debug: %s\n",str);
  fprintf(stderr,"g=%d\n",g);
}

void (* debug_print[])(char *) = { empty,
                                   oneline,
                                   detailed };

int main(int argc, char *argv[]) {

  if(argc!=2)
    return(EXIT_FAILURE);

  debug_level=atoi(argv[1]);
  if((debug_level<0) || (debug_level>2) )
    return(EXIT_FAILURE);
  printf("fct debug_print : %p\n",debug_print[debug_level]);
  (debug_print[debug_level])("Hello");

   return(EXIT_SUCCESS);
}
```

Ce n'est pas la seule utilisation des pointeurs vers des fonctions. Il y a notamment la fonction de la librairie qsort(3) qui permet de trier un tableau contenant n'importe quel type d'information. Cette fonction prend plusieurs arguments :

```
void qsort(void *base, size_t nel, size_t width,
           int (*compar)(const void *, const void *));
```

Le premier est un pointeur vers le début de la zone mémoire à trier. Le second est le nombre d'éléments à trier. Le troisième contient la taille des éléments stockés dans le tableau. Le quatrième argument est un pointeur vers la fonction qui permet de comparer deux éléments du tableau. Cette fonction retourne un entier négatif si son premier argument est inférieur au second et positif ou nul sinon. Un exemple de fonction de comparaison est la fonction strcmp(3) de la librairie standard. Un autre exemple est repris ci-dessous avec une fonction de comparaison simple qui permet d'utiliser qsort(3) pour trier un tableau de `double`.

```
#define SIZE 5
double array[SIZE]= { 1.0, 7.32, -3.43, 8.7, 9.99 };

void print_array() {
  for(int i=0;i<SIZE;i++)
    printf("array[i]:%f\n",array[i]);
}

int cmp(const void *ptr1, const void *ptr2) {
  const double *a=ptr1;
  const double *b=ptr2;
  if(*a==*b)
    return 0;
  else
    if(*a<*b)
      return -1;
    else
      return +1;
}

int main(int argc, char *argv[]) {
```

```
  printf("Avant qsort\n\n");
  print_array();
  qsort(array,SIZE,sizeof(double),cmp);
  printf("Après qsort\n\n");
  print_array();

  return(EXIT_SUCCESS);
}
```

Il est utile d'analyser en détails les arguments de la fonction de comparaison utilisée par qsort(3). Celle-ci prend deux arguments de type `const void *`. L'utilisation de pointeurs `void *` est nécessaire car la fonction doit être générique et pouvoir traiter n'importe quel type de pointeurs. `void *` est un pointeur vers une zone quelconque de mémoire qui peut être casté vers n'importe quel type de pointeur par la fonction de comparaison. `const` indique que la fonction n'a pas le droit de modifier la donnée référencée par ce pointeur, même si elle reçoit un pointeur vers cette donnée. On retrouvera régulièrement cette utilisation de `const` dans les signatures des fonctions de la librairie pour spécifier des contraintes sur les arguments passés à une fonction [16].

Le second type de pointeurs que nous n'avons pas encore abordé en détails sont les pointeurs vers des pointeurs. En fait, nous les avons utilisés sans vraiment le savoir dans la fonction `main`. En effet, le second argument de cette fonction est un tableau de pointeurs qui pointent chacun vers des chaînes de caractères différentes. La notation `char *argv[]` est équivalente à la notation `char **argv`. `**argv` est donc un pointeur vers une zone qui contient des pointeurs vers des chaînes de caractères. Ce pointeur vers un pointeur doit être utilisé avec précaution. `argv[0]` est un pointeur vers une chaîne de caractères. La construction `&(argv[0])` permet donc d'obtenir un pointeur vers un pointeur vers une chaîne de caractères, ce qui correspond bien à la déclaration `char **`. Ensuite, l'utilisation de `*p` pourrait surprendre. `*p` est un pointeur vers une chaîne de caractères. Il peut donc être comparé à `NULL` qui est aussi un pointeur, incrémenté et la chaîne de caractères qu'il référence peut être affichée par printf(3).

```
int main(int argc, char **argv) {

  char **p;
  p=argv;
  printf("Arguments :");
  while(*p!=NULL) {
    printf(" %s",*p);
    p++;
  }
  printf("\n");
  return(EXIT_SUCCESS);
}
```

En pratique, ces pointeurs vers des pointeurs se retrouveront lorsque l'on doit manipuler des structures multidimensionnelles, mais aussi lorsqu'il faut qu'une fonction puisse modifier une adresse qu'elle a reçue en argument.

Un autre exemple d'utilisation de pointeurs vers des pointeurs est la fonction strtol(3) de la librairie standard. Cette fonction est une généralisation des fonctions comme atoi(3). Elle permet de convertir une chaîne de caractères en un nombre. La fonction strtol(3) prend trois arguments et retourne un `long`. Le premier argument est un pointeur vers la chaîne de caractères à convertir. Le troisième argument est la base utilisée pour cette conversion.

```
#include <stdlib.h>
long
strtol(const char *restrict str, char **restrict endptr, int base);
```

L'utilisation principale de strtol(3) est de convertir une chaîne de caractères en un nombre. La fonction atoi(3) fait de même et l'expression `atoi("1252")` retourne l'entier `1252`. Malheureusement, la fonction atoi(3) ne traite pas correctement les chaînes de caractères qui ne contiennent pas un nombre. Elle ne retourne pas de code d'erreur et ne permet pas de savoir quelle partie de la chaîne de caractères passée en argument était en erreur.

strtol(3) est un exemple de fonction qui doit retourner deux types d'informations. Tout d'abord, strtol(3) retourne un résultat (dans ce cas un nombre). Si la chaîne de caractères à convertir est erronée, strtol(3) convertit le début de

16. `restrict` est également parfois utilisé pour indiquer des contraintes sur les pointeurs passés en argument à une fonction [Walls2006].

la chaîne et retourne un pointeur indiquant le premier caractère en erreur. Pour bien comprendre le fonctionnement de strtol(3), considérons l'exemple ci-dessous.

```
#include <stdlib.h>
#include <stdio.h>

int main(int argc, char *argv[]) {

  char *p, *s;
  long li;
  s = "1252";
  li = strtol(s,&p,10);
  if(*p != '\0') {
    printf("Caractère erronné : %c\n",*p);
      // p pointe vers le caractère en erreur
  }
  printf("Valeur convertie : %s -> %ld\n",s,li);

  s = "12m52";
  li = strtol(s,&p,10);
  if(*p != '\0') {
    printf("Caractère erronné : %c\n",*p);
  }
  printf("Valeur convertie : %s -> %ld\n",s,li);

  return(EXIT_SUCCESS);
}
```

Lors de son exécution, ce programme affiche la sortie suivante.

```
Valeur convertie : 1252 -> 1252
Caractère erronné : m
Valeur convertie : 12m52 -> 12
```

L'appel à strtol(3) prend trois arguments. Tout d'abord un pointeur vers la chaîne de caractères à convertir. Ensuite l'adresse d'un pointeur vers une chaîne de caractères. Enfin la base de conversion. La première chaîne de caractères est correcte. Elle est convertie directement. La seconde par contre contient un caractère erroné. Lors de son exécution, strtol(3) va détecter la présence du caractère `m` et placera un pointeur vers ce caractère dans `*p`. Pour que la fonction strtol(3) puisse retourner un pointeur de cette façon, il est nécessaire que son second argument soit de type `char **`. Si le second argument était de type `char *`, la fonction strtol(3) recevrait l'adresse d'une zone mémoire contenant un caractère. Comme le langage C utilise le passage par valeur, strtol(3) pourrait modifier la caractère pointé par ce pointeur mais pas son adresse. En utilisant un second argument de type `char **`, strtol(3) a la possibilité de modifier la valeur pointée par ce pointeur.

Une implémentation partielle de strtol(3) pourrait être la suivante.

```
#include <stdlib.h>
#include <stdio.h>
#include <ctype.h>
#include <stdbool.h>

int mystrtol(const char *restrict str,
             char **restrict endptr,
             int base) {

  int val;
  int i=0;
  int err=false;
  while(!err && *(str+i)!='\0')
    {
      if(!isdigit(*(str+i))) {
        err=true;
        *endptr=(char *)(str+i);
```

```
        }
        i++;
      }
    // ...
    return val;
}
```

Cette partie de code utilise la fonction isdigit(3) pour vérifier si les caractères présents dans la chaîne de caractères sont des chiffres. Sinon, elle fixe via son second argument la valeur du pointeur vers le caractère en erreur. Cela est réalisé par l'expression `*endptr=(char *)(str+i);`. Il faut noter que `*endptr` est bien une zone mémoire pointée par le pointeur `endptr` reçu comme second argument. Cette valeur peut donc être modifiée.

Il existe d'autres fonctions de la librairie standard qui utilisent des pointeurs vers des pointeurs comme arguments dont notamment strsep(3) et strtok_r(3).

3.7.2 De grands programmes en C

Lorsque l'on développe de grands programmes en C, il est préférable de découper le programme en modules. Chaque module contient des fonctions qui traitent d'un même type de problème et sont fortement couplées. A titre d'exemple, un module `stack` pourrait regrouper différentes fonctions de manipulation d'une pile. Un autre module pourrait regrouper les fonctions relatives au dialogue avec l'utilisateur, un autre les fonctions de gestion des fichiers, ...

Pour comprendre l'utilisation de ces modules, considérons d'abord un programme trivial composé de deux modules. Le premier module est celui qui contient la fonction `main`. Tout programme C doit contenir une fonction `main` pour pouvoir être exécuté. C'est en général l'interface avec l'utilisateur. Le second module contient une fonction générique qui est utilisée par le module principal.

```
/*************************************
 * main.c
 *
 * Programme d'exemple pour le linker
 *
 *************************************/

#include "min.h"
#include <stdio.h>
#include <stdlib.h>

int main(int argc, char *argv[]) {
  float f1=3.45;
  float f2=-4.12;
  printf("Minimum(%f,%f)=%f\n",f1,f2,min(f1,f2));
  return(EXIT_SUCCESS);
}
```

Un module d'un programme C est en général décomposé en deux parties. Tout d'abord, le fichier *fichier header* contient les définitions de certaines constantes et les signatures des fonctions exportées par ce module. Ce fichier est en quelque sorte un résumé du module, ou plus précisément de son interface externe. Il doit être inclus dans tout fichier qui utilise les fonctions du module correspondant. Dans un tel fichier *fichier header*, on retrouve généralement trois types d'informations :

– les signatures des fonctions qui sont définies dans le module. En général, seules les fonctions qui sont destinées à être utilisées par des modules extérieures sont reprises dans le *fichier header*
– les constantes qui sont utilisées à l'intérieur du module et doivent être visibles en dehors de celui-ci, notamment par les modules qui utilisent les fonctions du module. Ces constantes peuvent être définies en utilisant des directives `#define` du préprocesseur
– les variables globales qui sont utilisées par les fonctions du module et doivent être accessibles en dehors de celui-ci

```c
/************************************
 * min.h
 *
 ************************************/

#ifndef _MIN_H_
#define _MIN_H_

float min(float, float);

#endif /* _MIN_H */
```

Note : Un *fichier header* ne doit être inclus qu'une seule fois

L'exemple de *fichier header* ci-dessus illustre une convention courante dans l'écriture de ces fichiers. Parfois, il est nécessaire d'inclure un *fichier header* dans un autre fichier header. Suite à cela, il est possible que les mêmes définitions d'un *fichier header* soient incluses deux fois ou plus dans le même module. Cela peut causer des erreurs de compilation qui risquent de perturber certains programmeurs. Une règle de bonne pratique pour éviter ce problème est d'inclure le contenu du *fichier header* de façon conditionnelle comme présenté ci-dessus. Une constante, dans ce cas `_MIN_H_`, est définie pour le *fichier header* concerné. Cette constante est définie dans la première ligne effective du *fichier header*. Celui-ci n'est inclus dans un module que si cette constante n'a pas été préalablement définie. Si cette constante est connue par le préprocesseur, cela indique qu'un autre *fichier header* a déjà inclus les définitions de ce fichier et qu'elles ne doivent pas être incluses une seconde fois.

```c
/************************************
 * min.c
 *
 * Programme d'exemple pour le linker
 *
 ************************************/

#include "min.h"

float min(float a, float b) {
  if(a<b)
    return a;
  else
    return b;
}
```

Note : Localisation des fichiers header

Un programmeur C peut utiliser deux types de fichiers header. Il y a tout d'abord les fichiers headers standards qui sont fournis avec le système. Ce sont ceux que nous avons utilisés jusque maintenant. Ces headers standards se reconnaissent car ils sont entourés des caractères < et > dans la directive `#include`. Ceux-ci se trouvent dans des répertoires connus par le compilateur, normalement `/usr/include`. Les fichiers headers qui accompagnent un module se trouvent eux généralement dans le même répertoire que le module. Dans l'exemple ci-dessus, le header `min.h` est inclus via la directive `#include "min.h"`. Lorsque le préprocesseur rencontre une telle directive, il cherche le fichier dans le répertoire courant. Il est possible de spécifier des répertoires qui contiennent des fichiers headers via l'argument `-I` de gcc(1) ou en utilisant les variables d'environnement `GCC_INCLUDE_DIR` ou `CPATH`.

Lorsque l'on doit compiler un programme qui fait appel à plusieurs modules, quelle que soit sa taille, il est préférable d'utiliser make(1) pour automatiser sa compilation. Le fichier ci-dessous est un petit exemple de *Makefile* utilisable pour un tel projet.

```
myprog: main.o min.o
        gcc -std=c99 -o myprog main.o min.o
```

```
main.o: main.c min.h
        gcc -std=c99 -c main.c

min.o: min.c min.h
        gcc -std=c99 -c min.c
```

La compilation d'un tel programme se déroule en plusieurs étapes. La première étape est celle du préprocesseur. Celui-ci est directement appelé par le compilateur gcc(1) mais il est également possible de l'invoquer directement via cpp(1). Le préprocesseur remplace toutes les macros telles que les `#define` et insère les fichiers headers sur base des directives `#include`. La sortie du préprocesseur est utilisée directement par le compilateur. Celui-ci transforme le module en langage C en langage assembleur. Ce module en assembleur est ensuite assemblé par as(1) pour produire un *fichier objet*. Ce *fichier objet* n'est pas directement exécutable. Il contient les instructions en langage machine pour les différentes fonctions définies dans le module, les définitions des constantes et variables globales ainsi qu'une table reprenant tous les symboles (noms de fonction, noms de variables globales, ...) définis dans ce module. Ces phases sont exécutées pour chaque module utilisé. Par convention, les fichiers objets ont en général l'extension `.o`. Ces fichiers objet sont créés par les deux dernières cibles du fichier `Makefile` ci-dessus. L'option `-c` passée à gcc(1) indique à gcc(1) qu'il doit générer un fichier objet sans le transformer en exécutable. Cette dernière opération est réalisée par la première cible du `Makefile`. Dans cette cible, gcc(1) fait office d'éditeur de liens ou de *linker* en anglais. Le *linker* combine différents fichiers objets en faisant les liens nécessaires entre les fichiers. Dans notre exemple, le fichier `main.o` contient une référence vers la fonction `min` qui n'est pas connue lors de la compilation de `main.c`. Par contre, cette référence est connue dans le fichier `min.o`. L'éditeur de liens va combiner ces références de façon à permettre aux fonctions d'un module d'exécuter n'importe quelle fonction définie dans un autre module.

La figure ci-dessous représente graphiquement les différentes étapes de compilation des modules `min.c` et `main.c`.

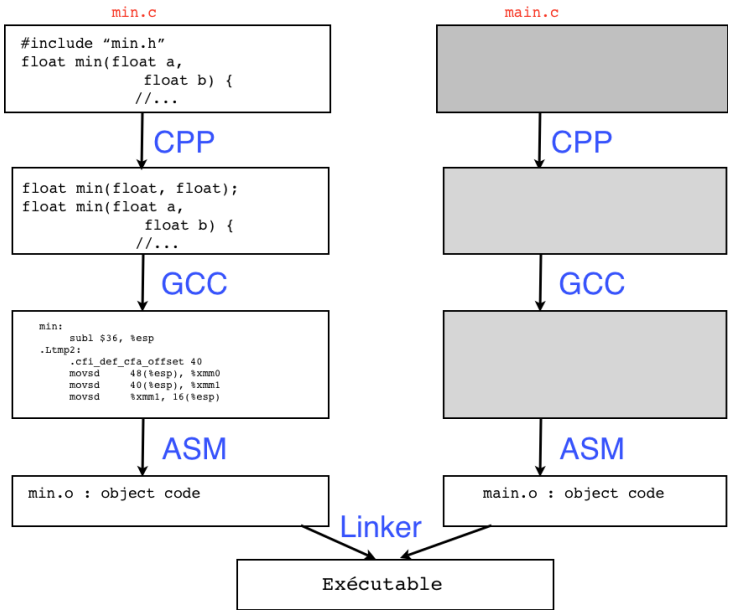


FIGURE 3.3 – Etapes de compilation

Lorsque plusieurs modules, potentiellement développés par des programmeurs qui ne se sont pas concertés, doivent être intégrés dans un grand programme, il y a des risques de conflits entre des variables et fonctions qui pourraient être définies dans plusieurs modules différents. Ainsi, deux modules pourraient définir la fonction `int min(int, int)` ou la variable globale `float dist`. Le langage C intègre des facilités qui permettent d'éviter ou de contrôler ces problèmes.

Tout d'abord, les variables locales sont locales au bloc dans lequel elles sont définies. Ce principe permet d'utiliser le même nom de variable dans plusieurs blocs d'un même fichier. Il s'étend naturellement à l'utilisation de variables locales dans des fichiers différents.

Pour les variables globales, la situation est différente. Si une variable est définie en dehors d'un bloc dans un fichier,

cette variable est considérée comme étant globale. Par défaut, elle est donc accessible depuis tous les modules qui composent le programme. Cela peut en pratique poser des difficultés si le même nom de variable est utilisé dans deux modules différents. Pour contourner ce problème, le langage C utilise `static`. Lorsque `static` est placé devant une déclaration de variable en dehors d'un bloc dans un module, il indique que la variable doit être accessible à toutes les fonctions du module mais pas en dehors du module. Lorsqu'un module utilise des variables qui sont communes à plusieurs fonctions mais ne doivent pas être visibles en dehors du module, il est important de les déclarer comme étant `static`. Lorsqu'une déclaration de variable globale est préfixée par `extern`, cela indique au compilateur que la variable est définie dans un autre module qui sera lié ultérieurement. Le compilateur réserve une place pour cette variable dans la table des symboles du fichier objet, mais cette place ne pourra être liée à la zone mémoire qui correspond à cette variable que lorsque l'éditeur de liens combinera les différents fichiers objet entre eux.

Note : Les deux utilisations de `static` pour des variables

`static` peut être utilisé à la fois pour des variables qui sont définies en dehors d'un bloc et dans un bloc. Lorsqu'une variable est définie comme étant `static` hors d'un bloc dans un module, elle n'est accessible qu'aux fonctions de ce module. Par contre, lorsqu'une variable est définie comme étant `static` à l'intérieur d'un bloc, par exemple dans une fonction, cela indique que cette variable doit toujours se trouver à la même localisation en mémoire, quel que soit le moment où elle est appelée. Ces variables `static` sont placées par le compilateur dans le bas de la mémoire, avec les variables globales. Contrairement aux variables locales traditionnelles, une variable locale `static` garde sa valeur d'une invocation de la fonction à l'autre. En pratique, les variables locales `static` doivent être utilisées avec précaution et bien documentées. Un de leurs intérêt est qu'elles ne sont initialisées qu'au lancement du programme et pas à chaque invocation de la fonction où elles sont définies.

Il faut noter que `static` peut aussi précéder des déclarations de fonctions. Dans ce cas, il indique que la fonction ne doit pas être visible en dehors du module dans lequel elle est définie. Sans `static`, une fonction déclarée dans un module est accessible depuis n'importe quel autre module.

Afin d'illustrer l'utilisation de `static` et `extern`, considérons le programme `prog.c` ci-dessous qui inclut le module `module.c` et également le module `min.c` présenté plus haut.

```
/*************************************
 * module.h
 *
 *************************************/
#ifndef _MODULE_H_
#define _MODULE_H_

float vmin(int, float *);

#endif /* _MODULE_H */

/*************************************
 * module.c
 *
 *************************************/

#include "module.h"

static float min(float, float);

int num1=0;  // accessible hors de module.c
extern int num2; // définie dans un autre module
static int num3=1252; // accessible uniquement dans ce module

float vmin(int n, float *ptr) {
  float *p=ptr;
  float m=*ptr;
  for(int i=1;i<n;i++) {
    m=min(m,*p);
    p++;
```

```
  }
  return m;
}

static float min(float a, float b) {
  if(a<b)
    return a;
  else
    return b;
}
```

Ce module contient deux fonctions, `vmin` et `min`. `vmin` est accessible depuis n'importe quel module. Sa signature est reprise dans le *fichier header* `module.h`. La fonction `min` par contre est déclarée comme étant `static`. Cela implique qu'elle n'est utilisable qu'à l'intérieur de ce module et invisible de tout autre module. La variable globale `num1` est accessible depuis n'importe quel module. La variable `num2` également, mais elle est initialisée dans un autre module. Enfin, la variable `num3` n'est accessible qu'à l'intérieur de ce module.

```
#include "min.h"
#include "module.h"

#define SIZE 4

extern int num1; // définie dans un autre module
int num2=1252;   // accessible depuis un autre module
static int num3=-1; // accessible uniquement dans ce module

void f() {
  static int n=0;
  int loc=2;
  if(n==0)
    printf("n est à  l'adresse %p et loc à l'adresse %p\n",&n,&loc);
  printf("f, n=%d\n",n);
  n++;
}

int main(int argc, char* argv[]) {

  float v[SIZE]={1.0, 3.4, -2.4, 9.9};
  printf("Minimum: %f\n",vmin(SIZE,v));
  f();
  f();
  printf("Minimum(0.0,1.1)=%f\n",min(0.0,1.1));
  return(EXIT_SUCCESS);
}
```

Ce module inclus les fichiers `min.h` et `module.h` qui contiennent les signatures des fonctions se trouvant dans ces deux modules. Trois variables globales sont utilisées par ce module. `num1` est définie dans un autre module (dans ce cas `module.c`). `num2` est initialisée dans ce module mais accessible depuis n'importe quel autre module. `num3` est une variable globale qui est accessible uniquement depuis le module `prog.c`. Même si cette variable porte le même nom qu'une autre variable déclarée dans `module.c`, il n'y aura pas de conflit puisque ces deux variables sont `static`.

La fonction `f` mérite que l'on s'y attarde un peu. Cette fonction contient la définition de la variable `static n`. Même si cette variable est locale à la fonction `f` et donc invisible en dehors de cette fonction, le compilateur va lui réserver une place dans la même zone que les variables globales. La valeur de cette variable `static` sera initialisée une seule fois : au démarrage du programme. Même si cette variable parait être locale, elle ne sera jamais réinitialisée lors d'un appel à la fonction `f`. Comme cette variable est stockée en dehors de la pile, elle conserve sa valeur d'une invocation à l'autre de la fonction `f`. Ceci est illustré par l'exécution du programme qui produit la sortie suivante.

```
Minimum: -2.400000
n est à  l'adresse 0x100001078 et loc à l'adresse 0x7fff5fbfe1cc
```

```
f, n=0
f, n=1
Minimum(0.0,1.1)=0.000000
```

Le dernier point à mentionner concernant cet exemple est relatif à la fonction `min` qui est utilisée dans la fonction `main`. Le module `prog.c` étant lié avec `module.c` et `min.c`, le linker associe à ce nom de fonction la déclaration qui se trouve dans le fichier `min.c`. La déclaration de la fonction `min` qui se trouve dans `module.c` est `static`, elle ne peut donc pas être utilisée en dehors de ce module.

3.7.3 Traitement des erreurs

Certaines fonctions de la librairie et certains appels systèmes réussissent toujours. C'est le cas par exemple pour getpid(2). D'autres fonctions peuvent échouer et il est important de tester la valeur de retour de chaque fonction/appel système utilisé pour pouvoir réagir correctement à toute erreur. Pour certaines fonctions ou appels systèmes, il est parfois nécessaire de fournir à l'utilisateur plus d'information sur l'erreur qui s'est produite. La valeur de retour utilisée pour la plupart des fonctions de la librairie et appels systèmes (souvent un `int` ou un pointeur), ne permet pas de fournir de l'information précise sur l'erreur qui s'est produite.

Les systèmes Unix utilisent la variable globale *errno* pour résoudre ce problème et permettre à une fonction de la librairie ou un appel système qui a échoué de donner plus de détails sur les raisons de l'échec. Cette variable globale est définie dans errno.h qui doit être inclus par tout programme voulant tester ces codes d'erreur. Cette variable est de type `int` et *errno.h* contient les définitions des constantes correspondants aux cas d'erreurs possibles. Il faut noter que la librairie standard fournit également les fonctions perror(3) et strerror(3) qui facilitent l'écriture de messages d'erreur compréhensibles pour l'utilisateur.

A titre d'exemple, le programme ci-dessous utilise strerror(3) pour afficher un message d'erreur plus parlant lors d'appels erronés à la fonction setenv(3).

```
#include <errno.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>

int main(int argc, char *argv[]) {

  if(setenv(NULL,NULL,1)!=0) {
    fprintf(stderr,"Erreur : errno=%d %s\n",errno,strerror(errno));
  }
  if(setenv("PATH=","/usr/bin",1)!=0) {
    fprintf(stderr,"Erreur : errno=%d %s\n",errno,strerror(errno));
   }


}
```

Note : La valeur de `errno` n'indique pas la réussite ou l'échec d'une fonction

Il faut noter que la variable `errno` n'est modifiée par les fonctions de la librairie ou les appels systèmes que si l'appel échoue. Si l'appel réussit, la valeur de `errno` n'est pas modifiée. Cela implique qu'il ne faut surtout pas tester la valeur de `errno` pour déterminer si une fonction de la librairie a échoué ou réussi. Il ne faut surtout pas utiliser le pattern suivant :

```
setenv("PATH","/usr/bin",1);
if(errno!=0) {
  fprintf(stderr,"Erreur : errno=%d %s\n",errno,strerror(errno);
}
```

Le code ci-dessus est erroné car il ne teste pas la valeur de retour de setenv(3). Comme les fonctions de la librairie et les appels systèmes ne modifient `errno` que lorsqu'une erreur survient, le code ci-dessus pourrait afficher un message d'erreur relatif à un appel système précédent qui n'a absolument rien à voir avec l'appel à la fonction setenv(3). Le code correct est évidemment de tester la valeur de retour de setenv(3) :

```
err=setenv("PATH","/usr/bin",1);
if(err!=0) {
  fprintf(stderr,"Erreur : errno=%d %s\n",errno,strerror(errno));
}
```

CHAPITRE 4

Structure des ordinateurs

4.1 Organisation des ordinateurs

Pour bien comprendre la façon dont les programmes s'exécutent sur un ordinateur, il est nécessaire de connaitre quelques principes de base sur l'architecture des ordinateurs et de leur organisation.

Comme nous l'avons déjà abordé dans l'introduction, un des premiers principes fondateurs est le modèle d'architecture de *von Neumann*. Ce modèle d'architecture a été introduit durant le développement des premiers ordinateurs pendant la seconde guerre mondiale mais reste tout à fait valide aujourd'hui [Krakowiak2011]. La partie la plus intéressante de ce modèle concerne les fonctions de calcul d'un ordinateur. Il postule qu'un ordinateur est organisé autour de deux types de dispositifs :

– L'unité centrale ou *processeur*. Cette unité centrale peut être décomposée en deux parties : l'unité de commande et l'unité arithmétique et logique. L'unité arithmétique et logique regroupe les circuits électroniques qui permettent d'effectuer les opérations arithmétiques (addition, soustraction, division, ...) et logiques. C'est cette unité qui réalise les calculs proprement dits. L'unité de commande permet quant à elle de charger, décoder et exécuter les instructions du programme qui sont stockées en mémoire.
– La *mémoire*. Celle-ci joue un double rôle. Elle stocke à la fois les données qui sont traitées par le programme mais aussi les instructions qui composent celui-ci. Cette utilisation de la mémoire pour stocker les données et une représentation binaire du programme à exécuter sont un des principes fondamentaux du fonctionnement des ordinateurs actuels.

La figure ci-dessous illustre les principaux éléments du modèle de von Neumann.

Dans cette section du cours, nous allons étudier plus en détails le fonctionnement d'un système informatique moderne et introduire des considérations de technologie et de performance. Ensuite, nous verrons comment un exemple de jeu d'instruction illustrant le fonctionnement au plus bas niveau du couple processeur et mémoire.

Les technologies utilisées pour construire les processeurs et la mémoire ont fortement évolué depuis les premiers ordinateurs, mais les principes fondamentaux restent applicables. En première approximation, on peut considérer la mémoire comme étant un dispositif qui permet de stocker des données binaires. La mémoire est découpée en blocs d'un octet. Chacun de ces blocs est identifié par une adresse, qui est elle aussi représentée sous la forme d'un nombre binaire. Une mémoire qui permet de stocker 2^k bytes de données utilisera au minimum k bits pour représenter l'adresse d'une zone mémoire. Ainsi, une mémoire pouvant stocker 64 millions de bytes doit utiliser au moins 26 bits d'adresse. En pratique, les processeurs des ordinateurs de bureau utilisent 32 ou 64 bits pour représenter les adresses en mémoire. D'anciens processeurs utilisaient 16 ou 20 bits d'adresse. Le nombre de bits utilisés pour représenter une adresse en mémoire limite la capacité totale de mémoire adressable par un processeur. Ainsi, un processeur qui utilise des adresses sur 32 bits n'est pas capable physiquement d'adresser plus de 4 GBytes de mémoire.

En pratique, l'organisation physique d'un ordinateur actuel est plus complexe que le modèle de von Neumann. Schématiquement, on peut considérer l'organisation présentée dans la figure ci-dessous. Le processeur est directement connecté à la mémoire via un *bus* de communication rapide. Ce bus permet des échanges de données et d'instructions efficaces entre la mémoire et le processeur. (Le bus est aussi utilisé pour gérer les échanges dans les systèmes équipés de plusieurs processeurs.) Outre le processeur et la mémoire, un troisième dispositif, souvent

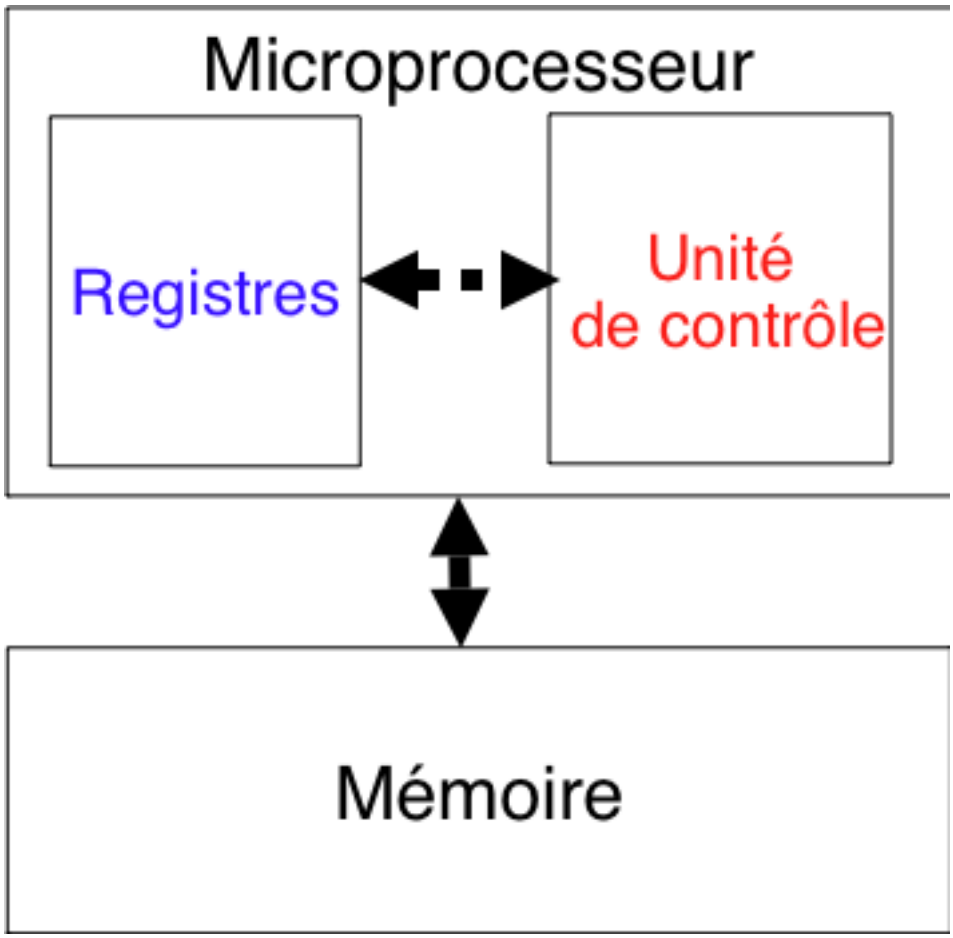


FIGURE 4.1 – Modèle de von Neumann

baptisé adaptateur de bus est connecté au bus processeur-mémoire. Cet adaptateur permet au processeur d'accéder aux gestionnaire de périphériques pour interagir avec les dispositifs de stockage et avec les dispositifs d'entrées-sorties tels que le clavier, la souris ou les cartes réseau. En pratique, cela se réalise en connectant les différents dispositifs à un autre bus de communication (PCI, SCSI, ...) et en utilisant un adaptateur de bus qui est capable de traduire les commandes venant du processeur.

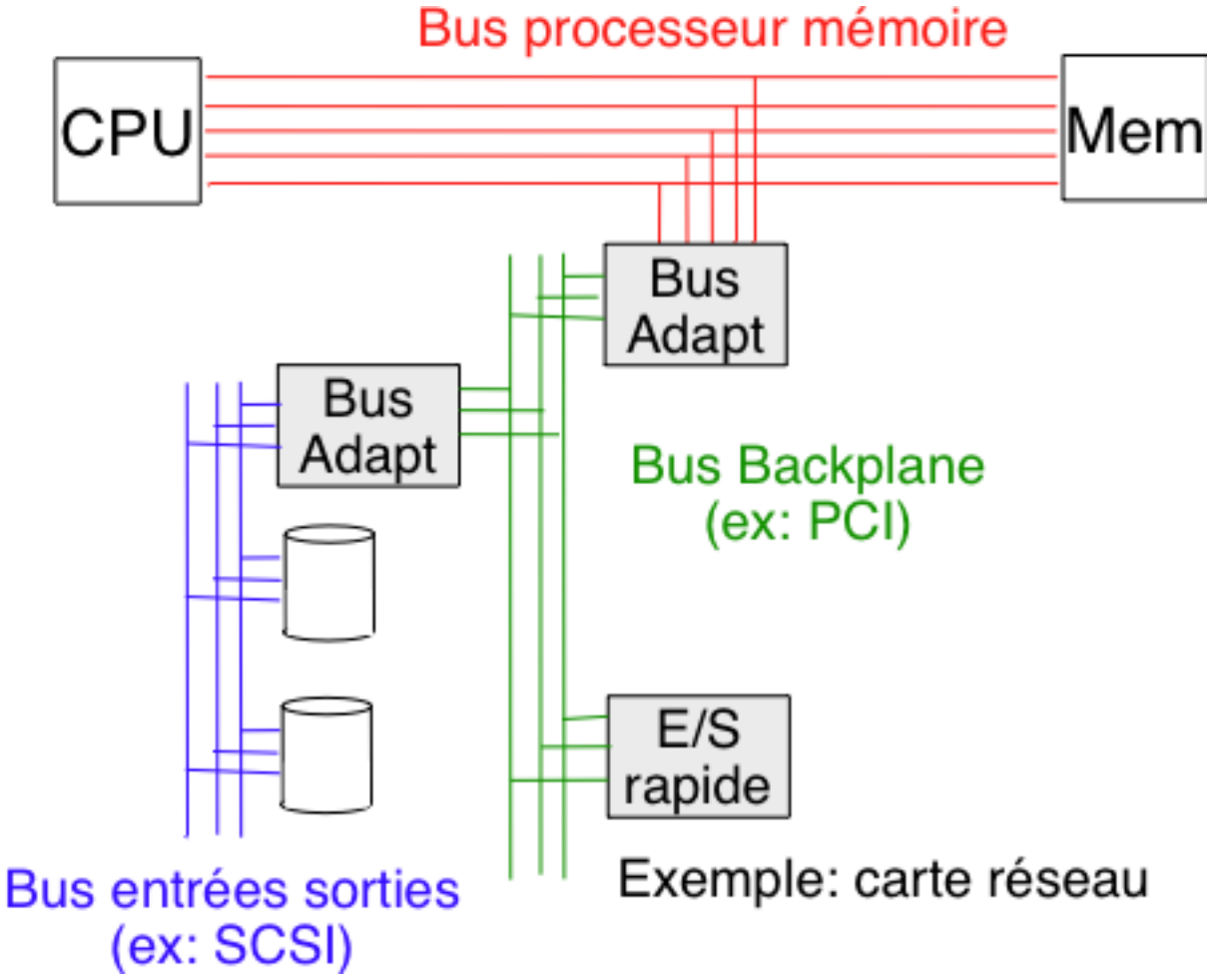


FIGURE 4.2 – Architecture d'un ordinateur actuel

Différentes technologies ont été mises en oeuvre pour construire les mémoires utilisées dans les ordinateurs. Aujourd'hui, les technologies les plus courantes sont les mémoires de type *SRAM* et les mémoires de type *DRAM*. Dans une *SRAM*, chaque bit d'information est stocké par un circuit composé de 4 à 6 transistors, et de quelques autres composants électroniques (résistances), mettant en oeuvre le principe de *bascule*. Une ligne d'information permet d'évaluer la valeur du bit stocké en observant le passage (ou non) d'un courant électrique dans ce circuit bascule, qui n'est possible que si le bit stocké est 1. Une ligne de commande permet par ailleurs de mettre la valeur stockée à 0 ou 1. Avec une *SRAM*, les données restent stockées de manière statique tant que le circuit de mémoire est sous tension (d'où leur nom de *static memory*–SRAM, par opposition à la mémoire dynamique). Malheureusement, les deux inconvénients majeurs de la *SRAM* est sa grande consommation électrique et le nombre important de composants électronique par bit stocké, qui empêche de développer des mémoires de grande capacité à un prix raisonnable. Aujourd'hui, les *SRAM* les plus grandes ont une capacité de seulement quelques douzaines de MBytes [HennessyPatterson].

Les *DRAM* sont totalement différentes des *SRAM* d'un point de vue de leur conception électronique. Dans une mémoire de type *DRAM*, c'est la présence ou l'absence d'une charge (de quelques électrons à quelques dizaines d'électrons) dans un condensateur qui représente la valeur `0` ou `1`. Comme la charge du condensateur est perdue au cours du temps à cause des courants de fuite, il est nécessaire de la remettre au maximum périodiquement. C'est la raison pour laquelle on parle de mémoire *dynamique* : il est nécessaire de parcourir, avec un certaine fréquence, l'ensemble des bits pour recharger les condensateurs représentant des `1`. Comme pour la mémoire

SRAM les données sont perdues lorsque l'alimentation en électricité est coupée.

Il est possible de construire des *DRAM* de très grande taille, jusqu'à 8 voire même 32 GByte par chip [Hennessy-Patterson]. C'est la raison pour laquelle on retrouve très largement des mémoires de type *DRAM* dans les ordinateurs. Malheureusement, leurs performances sont nettement moins bonnes que celles des mémoires de type *SRAM*. En pratique, une mémoire *DRAM* actuelle peut être vue comme étant équivalente à une grille [Drepper2007]. Les adresses peuvent être vues comme étant composées d'un numéro de colonne et d'un numéro de ligne. Pour lire ou écrire une donnée en mémoire *DRAM*, le processeur doit d'abord indiquer la ligne qu'il souhaite lire et ensuite la colonne. Ces deux opérations sont successives. Lorsque la mémoire a reçu la ligne et la colonne demandées, elle peut commencer le transfert de la donnée. En pratique, les mémoires *DRAM* sont optimisées pour fournir un débit de transfert élevé, mais elles ont une latence élevée. Cela implique que dans une mémoire *DRAM*, il est plus rapide de lire ou d'écrire un bloc de 128 bits successifs que quatre blocs de 32 bits à des endroits différents en mémoire. A titre d'exemple, le tableau ci-dessous, extrait de [HP] fournit le taux de transfert maximum de différentes technologies de *DRAM*.

Technologie	Fréquence [MHz]	Débit [MB/sec]
SDRAM	200	1064
RDRAM	400	1600
DDR-1	266	2656
DDR-2	333	5328
DDR-2	400	6400
DDR-3	400-667	6400-10600
DDR-4	800-1600	12800-25600

Le processeur interagit en permanence avec la mémoire, que ce soit pour charger des données à traiter ou pour charger les instructions à exécuter. Tant les données que les instructions sont représentées sous la forme de nombres en notation binaire. Certains processeurs utilisent des instructions de taille fixe. Par exemple, chaque instruction peut être encodée sous la forme d'un mot de 32 bits (4 octets). D'autres processeurs, comme ceux qui implémentent l'architecture [IA32], utilisent des instructions qui sont encodées sous la forme d'un nombre variable d'octets. Ces choix d'encodage des instructions influencent la façon dont les processeurs sont implémentés d'un point de vue microélectronique, mais ont assez peu d'impact sur le développeur de programmes. L'élément qu'il est important de bien comprendre est que le processeur doit en permanence charger des données et des instructions depuis la mémoire lors de l'exécution d'un programme.

Outre des unités de calcul, un processeur contient plusieurs registres. Un *registre* est une zone de mémoire très rapide se trouvant sur le processeur. Sur les processeurs actuels, cette zone de mémoire permet de stocker un mot de 32 bits ou un long mot de 64 bits. Les premiers processeurs disposaient d'un registre unique baptisé l'*accumulateur*. Les processeurs actuels en contiennent généralement une ou quelques dizaines. Chaque registre est identifié par un nom ou un numéro et les instructions du processeur permettent d'accéder directement aux données se trouvant dans un registre particulier. Les registres sont les mémoires les plus rapides qui sont disponibles sur un ordinateur. Malheureusement, ils sont en nombre très limité et il est impossible de faire fonctionner un programme non trivial en utilisant uniquement des registres.

Du point de vue des performances, il serait préférable de pouvoir construire un ordinateur équipé uniquement de *SRAM*. Malheureusement, au niveau de la capacité et du prix, c'est impossible sauf pour de rares applications bien spécifiques qui nécessitent de hautes performances et se contentent d'une capacité limitée. Les ordinateurs actuels utilisent en même temps de la mémoire *SRAM* et de la mémoire *DRAM*. Avec les registres, les *SRAM* et les *DRAM* composent les trois premiers niveaux de la *hiérarchie de mémoire*.

Le tableau ci-dessous, extrait de [BryantOHallaron2011] , compare les temps d'accès entre les mémoires *SRAM* et les mémoires *DRAM* à différentes périodes.

Année	Accès SRAM	Accès DRAM
1980	300 ns	375 ns
1985	150 ns	200 ns
1990	35 ns	100 ns
1995	15 ns	70 ns
2000	3 ns	60 ns
2005	2 ns	50 ns
2010	1.5 ns	40 ns

Cette évolution des temps d'accès doit être mise en parallèle avec l'évolution des performances des processeurs.

En 1980, le processeur Intel 8080 fonctionnait avec une horloge de 1 MHz et accédait à la mémoire toutes les 1000 ns. A cette époque, la mémoire était nettement plus rapide que le processeur. En 1990, par contre, le processeur Intel 80386 accédait à la mémoire en moyenne toutes les 50 ns. Couplé à une mémoire uniquement de type DRAM, il était ralenti par cette mémoire. En 2000, le Pentium-III avait un cycle de 1.6 ns, plus rapide que les meilleures mémoires disponibles à l'époque. Il en va de même aujourd'hui où les temps de cycle sont inférieurs au temps d'accès des mémoires. Même s'il existe des solutions techniques pour mitiger ce problème, la différence de performance croissante entre la mémoire et le processeur est un des facteurs qui limitent les améliorations des performances de nombreux programmes.

Une première solution pour combiner la *SRAM* et la *DRAM* serait de réserver par exemple les adresses basses à la *SRAM* qui est plus performante et d'utiliser la *DRAM* pour les adresses hautes. Avec cette solution, le programme stocké dans la *SRAM* pourrait s'exécuter plus rapidement que le programme stocké en *DRAM*. Afin de permettre à tous les programmes de pouvoir utiliser la *SRAM*, on pourrait imaginer que le système d'exploitation fournisse des fonctions qui permettent aux applications de demander la taille de la mémoire *SRAM* disponible et de déplacer des parties d'un programme et des données en *SRAM*. Ce genre de solution obligerait chaque application à pouvoir déterminer quelles sont les instructions à exécuter et quelles données doivent être placées en mémoire *SRAM* pour obtenir les meilleures performances. Même si en théorie, ce genre de solution est envisageable, en pratique, elle a très peu de chances de pouvoir fonctionner.

La deuxième solution est d'utiliser le principe de la *mémoire cache*. Une *mémoire cache* est une mémoire de faible capacité mais rapide. La mémoire cache peut stocker des données provenant de mémoires de plus grande capacité mais plus lentes. Cette mémoire cache sert d'interface entre le processeur et la mémoire principale. Toutes les demandes d'accès à la mémoire principale passent par la mémoire cache comme illustré dans la figure ci-dessous.

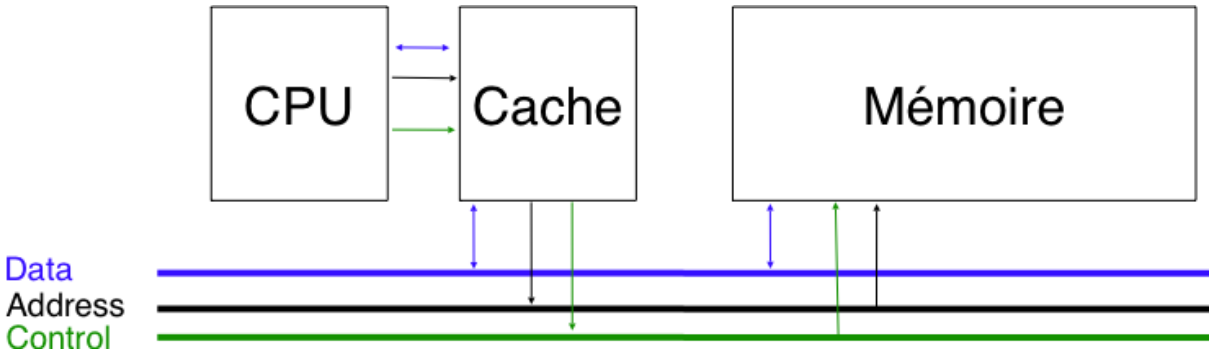


FIGURE 4.3 – Le processeur, la mémoire cache et la mémoire principale

On utilise des mémoires caches dans de nombreux systèmes informatiques de façon à améliorer leurs performances. Ces mémoires caches utilisent en fait le *principe de localité*. En pratique, deux types de localité doivent être considérés. Tout d'abord, il y a la *localité temporelle*. Si un processeur a accédé à la mémoire à l'adresse *A* à l'instant *t*, il est fort probable qu'il accédera encore à cette adresse dans les instants qui suivent. La localité temporelle apparaît notamment lors de l'exécution de longues boucles qui exécutent à de nombreuses reprises les mêmes instructions. Le second type de localité est la *localité spatiale*. Celle-ci implique que si un programme a accédé à l'adresse *A* à l'instant *t*, il est fort probable qu'il accédera aux adresses proches de *A* comme *A+4*, *A-4* dans les instants qui suivent. Cette localité apparaît par exemple lorsqu'un programme traite un vecteur stocké en mémoire.

Les mémoires caches exploitent ces principes de localité en stockant de façon transparente les instructions et les données les plus récemment utilisées. D'un point de vue physique, on peut voir le processeur comme étant connecté à la (ou parfois les) mémoire cache qui est elle-même connectée à la mémoire *RAM*. Les opérations de lecture en mémoire se déroulent généralement comme suit. Chaque fois que le processeur a besoin de lire une donnée se trouvant à une adresse, il fournit l'adresse demandée à la mémoire cache. Si la donnée correspondant à cette adresse est présente en mémoire cache, celle-ci répond directement au processeur. Sinon, la mémoire cache interroge la mémoire *RAM*, se met à jour et ensuite fournit la donnée demandée au processeur. Ce mode de fonctionnement permet à la mémoire cache de se mettre à jour au fur et à mesure des demandes faites par le processeur afin de profiter de la localité temporelle. Pour profiter de la localité spatiale, la plupart des caches se mettent à jour en chargeant directement une *ligne de cache* qui peut compter jusqu'à quelques dizaines d'adresses consécutives en mémoire (par exemple, 16 octets sur la plupart des processeurs modernes). Ce chargement d'une

ligne complète de cache permet également de profiter des mémoires *DRAM* récentes qui sont souvent optimisées pour fournir des débits de transfert élevés pour de longs blocs consécutifs en mémoire. La figure ci-dessous illustre graphiquement la hiérarchie de mémoires dans un ordinateur.

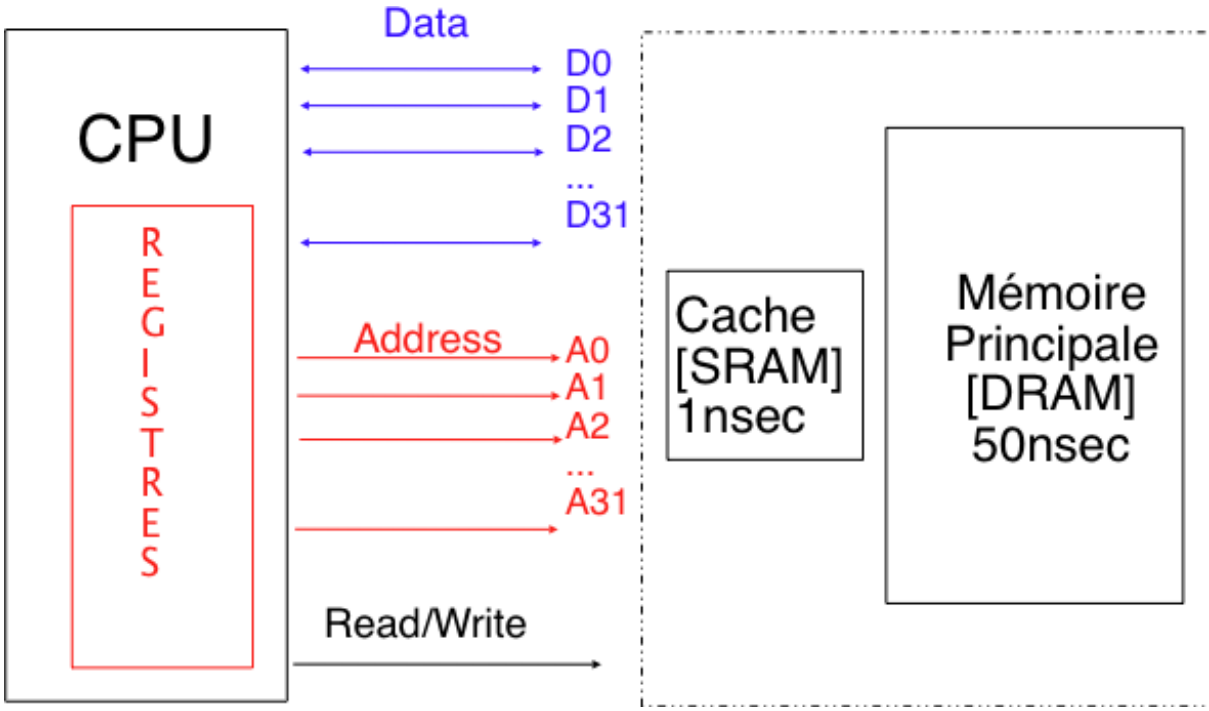


FIGURE 4.4 – La hiérarchie de mémoires

Pour les opérations d'écriture, la situation est plus compliquée. Si le processeur écrit l'information x à l'adresse A en mémoire, il faudrait idéalement que cette valeur soit écrite simultanément en mémoire cache et en mémoire *RAM* de façon à s'assurer que la mémoire *RAM* contienne toujours des données à jour. La stratégie d'écriture la plus simple est baptisée *write through*. Avec cette stratégie, toute demande d'écriture venant du processeur donne lieu à une écriture en mémoire cache et une écriture en mémoire *RAM*. Cette stratégie garantit qu'à tout moment la mémoire cache et la mémoire *RAM* contiennent la même information. Malheureusement, d'un point de vue des performances, cette technique rabaisse les performances de la mémoire cache à celles de la mémoire *RAM*. Vu la différence de performance entre les deux types de mémoires, cette stratégie n'est plus acceptable aujourd'hui. L'alternative est d'utiliser la technique du *write back*. Avec cette technique, toute écriture est faite en *mémoire cache* directement. Cela permet d'obtenir de très bonnes performances pour les écritures. Une donnée modifiée n'est réécrite en mémoire *RAM* que lorsqu'elle doit être retirée de la mémoire cache. Cette écriture est faite automatiquement par la mémoire cache. Pour la plupart des programmes, la gestion des opérations d'écriture est transparente. Il faut cependant être attentif à la technique d'écriture utilisée lorsque plusieurs dispositifs peuvent accéder directement à la mémoire *RAM* sans passer par le processeur. C'est le cas par exemple pour certaines cartes réseaux ou certains contrôleurs de disque dur. Pour des raisons de performances, ces dispositifs peuvent copier des données directement de la mémoire *RAM* vers le réseau ou un disque dur, en utilisant le principe de *Direct Memory Access* (DMA) présenté dans l'introduction de ce cours. Si une écriture de type *write-back* est utilisée, le système d'exploitation doit veiller à ce que les données écrites par le processeur en cache aient bien été écrites également en mémoire *RAM* avant d'autoriser la carte réseau ou le contrôleur de disque à effectuer un transfert en utilisant DMA.

4.2 Etude de cas : Architecture IA32

Pour comprendre le fonctionnement d'un microprocesseur, la solution la plus efficace est de considérer une architecture en particulier et de voir comment fonctionnent les processeurs qui l'implémentent. Dans cette section,

nous analysons brièvement le fonctionnement des processeurs [1] de la famille [IA32].

Cette architecture recouvre un grand nombre de variantes qui ont leur spécificités propre. Une descriptions détaillée de cette architecture est disponible dans [IA32]. Nous nous limiterons à un très petit sous-ensemble de cette architecture dans le cadre de ce cours. Une analyse complète de l'architecture [IA32] occupe plusieurs centaines de pages dans des livres de référence [BryantOHallaron2011] [Hyde2010].

L'architecture [IA32] est supportée par différents types de processeurs. Certains utilisent des registres et des bus de données de 32 bits. D'autres, plus récents utilisent des registres de 64 bits. Il y a des différences importantes entre ces deux architectures. Comme les processeurs récents supportent à la fois les modes 32 bits et 64 bits, nous nous limiterons à l'architecture 32 bits.

Un des éléments importants d'un processeur tel que ceux de l'architecture [IA32] sont ses registres. Un processeur [IA32] dispose de huit registres génériques. Ceux-ci ont été baptisés `EAX`, `EBX`, `ECX`, `EDX`, `EBP`, `ESI`, `EDI` et `ESP`. Ces registres peuvent stocker des données sous forme binaire. Dans l'architecture [IA32], ils ont une taille de 32 bits. Cela implique que chaque registre peut contenir un nombre ou une adresse puisque les entiers (`int` en C) et les adresses (pointeurs `*` en C sur [IA32]) sont tous les deux encodés sur 32 bits dans l'architecture [IA32]. Cette capacité à stocker des données ou des adresses à l'intérieur d'un même registre est un des points clés de la flexibilité des microprocesseurs.

Deux de ces registres, `EBP` et `ESP` sont utilisés dans la gestion de la pile comme nous le verrons plus tard. Les autres registres peuvent être utilisés directement par le programmeur. En outre, tout processeur contient un registre spécial qui stocke à tout moment l'adresse de l'instruction courante en mémoire. Ce registre est souvent dénommé le *compteur de programme* ou *program counter* (*PC*) en anglais. Dans l'architecture [IA32], c'est le registre `EIP` qui stocke l'*Instruction Pointer* qui joue ce rôle. Ce registre ne peut pas être utilisé pour effectuer des opérations arithmétiques. Il peut cependant être modifié par les instructions de saut comme nous le verrons plus tard et joue un rôle essentiel dans l'implémentation des instructions de contrôle.

Outre ces registres génériques, les processeurs de la famille [IA32] contiennent aussi des registres spécialisés pour manipuler les nombres en virgule flottante (`float` et `double`). Nous ne les analyserons pas dans le cadre de ce cours. Par contre, les processeurs [IA32] contiennent également des drapeaux regroupés dans le registre `eflags`. Ceux-ci sont utilisés pour implémenter différents tests et comparaisons.

Les processeurs qui implémentent les spécifications [IA32] supportent les types de données repris dans le tableau ci-dessous.

Type	Taille (bytes)	Suffixe assembleur
`char`	1	b
`short`	2	w
`int`	4	l
`long int`	4	l
`void *`	4	l

Dans les sections qui suivent, nous analysons quelques instructions de l'architecture [IA32] qui permettent de manipuler des nombres entiers en commençant par les instructions de transfert entre la mémoire et les registres.

4.2.1 Les instructions `mov`

Les instructions de la famille `mov` [2] permettent de déplacer des données entre registres ou depuis la mémoire vers un registre ou enfin d'un registre vers une zone mémoire. Ces instructions sont essentielles car elles permettent au processeur de récupérer les données qui sont stockées en mémoire mais aussi de sauvegarder en mémoire le résultat d'un calcul effectué par le processeur. Une instruction `mov` contient toujours deux arguments. Le premier spécifie la donnée à déplacer ou son adresse et la seconde l'endroit où il faut sauvegarder cette donnée ou la valeur stockée à cette adresse.

1. Pour une liste détaillée des processeurs de cette famille produits par intel, voir notamment http ://www.intel.com/pressroom/kits/quickreffam.htm D'autres fabricants produisent également des processeurs compatibles avec l'architecture [IA32].
2. On parle de famille d'instructions car il existe de nombreuses instructions de déplacement en mémoire. Les plus simples sont suffixées par un caractère qui indique le type de données transféré. Ainsi, `movb` permet le transfert d'un byte tandis que `movl` permet le transfert d'un mot de 32 bits. Des détails sur ces instructions peuvent être obtenus dans [IA32]

```
mov src, dest  ; déplacement de src vers dest
```

Il existe une instruction de la famille `mov` qui correspond à chaque type de donnée pouvant être déplacé. L'instruction `movb` est utilisée pour déplacer un byte, `movw` pour déplacer un mot de 16 bits et `movl` lorsqu'il faut déplacer un mot de 32 bits.

En pratique, il y a plusieurs façons de spécifier chaque argument d'une instruction `mov`. Certains auteurs utilisent le terme *mode d'adressage* pour représenter ces différents types d'arguments même si il ne s'agit pas toujours d'adresses. Le premier mode est le mode *registre*. La source et la destination d'une opération `mov` peuvent être un nom de registre. Ceux-ci sont en général préfixés avec le caractère `%`. Ainsi, `%eax` correspond au registre `EAX`. La première instruction ci-dessous déplace le mot de 32 bits stocké dans le registre `%eax` vers le registre `%ebx`. La seconde instruction elle n'a aucun effet puisqu'elle déplace le contenu du registre `%ecx` vers ce même registre.

```
movl %eax, %ebx   ; déplacement de %eax vers %ebx
movl %ecx, %ecx   ; aucun effet
```

Le deuxième mode d'adressage est le mode *immédiat*. Celui-ci ne peut être utilisé que pour l'argument *source*. Il permet de placer une constante dans un registre, par exemple pour initialiser sa valeur comme dans les exemples ci-dessous. Il se reconnaît à l'utilisation du symbole `$` comme préfixe de la constante.

```
movl $0, %eax      ; initialisation de %eax à 0
movl $1252, %ecx   ; initialisation de %ecx à 1252
```

Le troisième mode d'adressage est le mode *absolu*. Dans ce mode, l'un des arguments de l'instruction `mov` est une adresse en mémoire. Si la source est une adresse, alors l'instruction `mov` transfère le mot de 32 bits stocké à l'adresse spécifiée comme source vers le registre spécifié comme destination. Si la destination est une adresse, alors l'instruction `mov` sauvegarde la donnée source à cette adresse en mémoire. Pour illustrer cette utilisation de l'instruction `mov`, considérons la mémoire illustrée ci-dessous.

Adresse	Valeur
0x10	0x04
0x0C	0x10
0x08	0xFF
0x04	0x00
0x00	0x04

Les instructions ci-dessous sont un exemple de déplacement de données entre la mémoire et un registre et d'un registre vers la mémoire.

```
movl 0x04, %eax    ; place la valeur 0x00 (qui se trouve à l'adresse 0x04) dans %eax
movl $1252, %ecx   ; initialisation de %ecx à 1252
movl %ecx, 0x08    ; remplace 0xFF par le contenu de %ecx (1252) à l'adresse 0x08
```

Le quatrième mode d'adressage est le mode *indirect*. Plutôt que de spécifier directement une adresse, avec le mode indirect, on spécifie un registre dont la valeur est une adresse en mémoire. Ce mode indirect est équivalent à l'utilisation des pointeurs en langage C. Il se reconnait à l'utilisation de parenthèses autour du nom du registre source ou destination. L'exemple ci-dessous illustre l'utilisation de l'adressage indirect en considérant la mémoire présentée plus haut.

```
movl $0x08, %eax    ; place la valeur 0x08 dans %eax
movl (%eax), %ecx   ; place la valeur se trouvant à l'adresse qui est
                    ; dans %eax dans le registre %ecx : %ecx=0xFF
movl 0x10, %eax     ; place la valeur se trouvant à l'adresse 0x10 dans %eax
movl %ecx, (%eax)   ; place le contenu de %ecx, c'est-à-dire 0xFF à l'adresse qui est contenue da
```

Le cinquième mode d'adressage est le mode avec une *base* et un *déplacement*. Ce mode peut être vu comme une extension du mode *indirect*. Il permet de lire en mémoire à une adresse qui est obtenue en additionnant un entier, positif ou négatif, à une adresse stockée dans un registre. Ce mode d'adressage joue un rôle important dans le fonctionnement de la pile comme nous le verrons d'ici peu.

```
movl $0x08, %eax      ; place la valeur 0x08 dans %eax
movl 0(%eax), %ecx    ; place la valeur (0xFF) se trouvant à l'adresse
                      ; 0x08= (0x08+0) dans le registre %ecx
movl 4(%eax), %ecx    ; place la valeur (0x10) se trouvant à l'adresse
                      ; 0x0C (0x08+4) dans le registre %ecx
movl -8(%eax), %ecx   ; place la valeur (0x04) se trouvant à l'adresse
                      ; 0x00 (0x08-8) dans le registre %ecx
```

L'architecture [IA32] supporte encore d'autres modes d'adressage. Ceux sont décrits dans [IA32] ou [BryantO-Hallaron2011]. Une autre instruction permettant de déplacer de l'information est l'instruction `leal` (load effective address). Cette instruction est parfois utilisée par les compilateurs. Elle place dans le registre destination l'adresse de son argument source plutôt que sa valeur. Ainsi `leal 4(%esp) %edx` placera dans le registre `%edx` l'adresse de son argument source, c'est-à-dire l'adresse contenue dans `%esp+4`.

Pas de déplacement entre deux adresses

On pourrait vouloir écrire l'instruction `movl 0x04, 0x02` pour copier le contenu de la mémoire à l'adresse `0x04` à l'adresse `0x02`, ou bien écrire `movl (%eax), (%ebx)` pour copier le contenu de la mémoire à l'adresse pointée par le contenu du registre EAX vers l'adresse pointée par le contenu du registre EBX. Ces deux opérations ne sont pas permises dans le jeu d'instruction [IA32], et il est donc nécessaire de passer par un registre intermédiaire. D'autres architectures peuvent supporter l'utilisation de deux adresses mémoires dans une instruction, ici il s'agit d'un choix des concepteurs du jeu d'instruction qui permet de limiter la complexité de mise en oeuvre des processeurs le supportant.

4.2.2 Instructions arithmétiques et logiques

La deuxième famille d'instructions importante sur un processeur sont les instructions qui permettent d'effectuer les opérations arithmétiques et logiques. Voici quelques exemples d'instructions arithmétiques et logiques supportées par l'architecture [IA32].

Les instructions les plus simples sont celles qui prennent un seul argument. Il s'agit de :
- `inc` qui incrémente d'une unité la valeur stockée dans le registre/l'adresse fournie en argument et sauvegarde le résultat de l'incrémentation au même endroit. Cette instruction peut être utilisée pour implémenter des compteurs de boucles.
- `dec` est équivalente à `inc` mais décrémente son argument.
- `not` qui applique l'opération logique `NOT` à son argument et stocke le résultat à cet endroit

Il existe une variante de chacune de ces instructions pour chaque type de données à manipuler. Cette variante se reconnait grâce au dernier caractère de l'instruction (`b` pour byte, `w` pour un mot de 16 bits et `l` pour un mot de 32 bits). Nous nous limiterons aux instructions qui manipulent des mots de 32 bits.

```
movl    $0x12345678, %ecx   ; initialisation
notl    %ecx                ; calcul de NOT
movl    $0, %eax            ; %eax=0
incl    %eax                ; %eax++
```

L'architecture [IA32] supporte également des instructions arithmétiques et logiques prenant chacune deux arguments.
- `add` permet d'additionner deux nombres entiers. `add` prend comme arguments une source et une destination et place dans la destination la somme de ses deux arguments.
- `sub` permet de soustraire le premier argument du second et stocke le résultat dans le second
- `mul` permet de multiplier des nombres entiers non-signés (`imul` est le pendant de `mul` pour la multiplication de nombres signés)
- `div` permet la division de nombres entiers non-signés.
- `shl` (resp. `shr`) permet de réaliser un décalage logique vers la gauche (resp. droite)
- `xor` calcule un ou exclusif entre ses deux arguments et sauvegarde le résultat dans le second
- `and` calcule la conjonction logique entre ses deux arguments et sauvegarde le résultat dans le second

Pour illustrer le fonctionnement de ces instructions, considérons une mémoire hypothétique contenant les données suivantes. Supposons que la variable entière `a` est stockée à l'adresse `0x04`, `b` à l'adresse `0x08` et `c` à l'adresse `0x0C`.

Adresse	Variable	Valeur
0x0C	c	0x00
0x08	b	0xFF
0x04	a	0x02
0x00	_	0x01

Les trois premières instructions ci-dessous sont équivalentes à l'expression C `a=a+1;`. Pour implémenter une telle opération en C, il faut d'abord charger la valeur de la variable dans un registre. Ensuite le processeur effectue l'opération arithmétique. Enfin le résultat est sauvegardé en mémoire. Après ces trois instructions, la valeur `0x03` est stockée à l'adresse `0x04` qui correspond à la variable `a`. Les trois dernières instructions calculent `a=b-c;`. On remarquera que le programmeur a choisi de d'abord charger la valeur de la variable `b` dans le registre `%eax`. Ensuite il utilise l'instruction `subl` en mode d'adressage immédiat pour placer dans `%eax` le résultat de la soustraction entre `%eax` et la donnée se trouvant à l'adresse `0x0C`. Enfin, le contenu de `%eax` est sauvé à l'adresse correspondant à la variable `a`.

```
movl   0x04, %eax     ; %eax=a
addl   $1, %eax       ; %eax++
movl   %eax, 0x04     ; a=%eax
movl   0x08, %eax     ; %eax=b
subl   0x0c, %eax     ; %eax=b-c
movl   %eax, 0x04     ; a=%eax
```

L'exemple ci-dessous présente la traduction directe [3] d'un fragment de programme C utilisant des variables globales en langage assembleur.

```
int j,k,g,l;
// ...
l=g^j;
j=j|k;
g=l<<6;
```

Dans le code assembleur, les noms de variables tels que `g` ou `j` correspondent à l'adresse mémoire à laquelle la variable est stockée.

```
movl    g, %eax  ; %eax=g
xorl    j, %eax  ; %eax=g^j
movl    %eax, l  ; l=%eax
movl    j, %eax  ; %eax=j
orl     k, %eax  ; %eax=j|k
movl    %eax, j  ; j=%eax
movl    l, %eax  ; %eax=l
shll    $6, %eax ; %eax=%eax << 6
movl    %eax, g  ; g=%eax
```

Les opérations arithmétiques telles que la multiplication ou la division sont plus complexes que les opérations qui ont été présentées ci-dessus. En toute généralité, la multiplication entre deux nombres de 32 bits peut donner un résultat sur 64 bits qui ne pourra donc pas être stocké entièrement dans un registre. De la même manière, une division entière retourne un quotient et un reste qui sont tous les deux sur 32 bits. L'utilisation des instructions de division et de multiplication nécessite de prendre ces problèmes en compte. Nous ne les aborderons pas dans ce cours. Des détails complémentaires sont disponibles dans [IA32] et [BryantOHallaron2011] notamment.

3. Cette traduction et la plupart des traductions utilisées dans ce chapitre ont été obtenues en utilisant l'interface web de démo du compilateur llvm qui a été configuré pour générer du code 32 bits sans optimisation. Quelques détails ont été supprimés du code assembleur pour le rendre plus compact.

4.2.3 Les instructions de comparaison

Outre les opérations arithmétiques, un processeur doit être capable de réaliser des comparaisons. Ces comparaisons sont nécessaires pour implémenter des tests tels que `if (condition) { ... } else { ... }`. Sur les processeurs [IA32], les comparaisons utilisent des drapeaux qui sont mis à jour par le processeur après l'exécution de certaines instructions. Ceux-ci sont regroupés dans le registre `eflags`. Les principaux drapeaux sont :
– *ZF* (Zero Flag) : ce drapeau indique si le résultat de la dernière opération était zéro
– *SF* (Sign Flag) : indique si le résultat de la dernière instruction était négatif
– *CF* (Carry Flag) : indique si le résultat de la dernière instruction arithmétique non signée nécessitait plus de 32 bits pour être stocké
– *OF* (Overflow Flag) : indique si le résultat de la dernière instruction arithmétique signée a provoqué un dépassement de capacité
Nous utiliserons principalement les drapeaux *ZF* et *SF* dans ce chapitre. Ces drapeaux peuvent être fixés par les instructions arithmétiques standard, mais aussi par des instructions dédiées comme `cmp` et `test`. L'instruction `cmp` effectue l'équivalent d'une soustraction et met à jour les drapeaux *CF* et *SF* mais sans sauvegarder son résultat dans un registre. L'instruction `test` effectue elle une conjonction logique sans sauvegarder son résultat mais en mettant à jour les drapeaux.

Ces instructions de comparaison peuvent être utilisées avec les instructions `set` qui permettent de fixer la valeur d'un registre en fonction des valeurs de certains drapeaux du registre `eflags`. Chaque instruction `set` prend comme argument un registre. Pour des raisons historiques, ces instructions modifient uniquement les 8 bits de poids faible du registre indiqué et non le registre complet. C'est un détail qui est lié à l'histoire de l'architecture [IA32].
– `sete` met le registre argument à la valeur du drapeau *ZF*. Permet d'implémenter une égalité.
– `sets` met le registre argument à la valeur du drapeau *SF*
– `setg` place dans le registre argument la valeur `~SF & ~ZF` (tout en prenant en compte les dépassements éventuels avec *OF*). Permet d'implémenter la condition >.
– `setl` place dans le registre argument la valeur de `SF` (tout en prenant en compte les dépassements éventuels avec *OF*). Permet d'implémenter notamment la condition <=.
A titre d'illustration, voici quelques expressions logiques en C et leur implémentation en assembleur lorsque les variables utilisées sont toutes des variables globales.

```
r=(h>1);
r=(j==0);
r=g<=h;
r=(j==h);
```

Le programme assembleur utilise une instruction `cmpl` pour effectuer la comparaison. Ensuite, une instruction `set` permet de fixer la valeur du byte de poids faible de `%eax` et une instruction (`movzbl`) permettant de transformer ce byte en un mot de 32 bits afin de pouvoir le stocker en mémoire. Cette traduction a été obtenue avec *llvm*, d'autres compilateurs peuvent générer du code un peu différent.

```
cmpl    $1, h         ; comparaison
setg    %al           ; %al est le byte de poids faible de %eax
movzbl  %al, %ecx     ; copie le byte dans %ecx
movl    %ecx, r       ; sauvegarde du résultat dans r

cmpl    $0, j         ; comparaison
sete    %al           ; fixe le byte de poids faible de %eax
movzbl  %al, %ecx
movl    %ecx, r       ; sauvegarde du résultat dans r

movl    g, %ecx
cmpl    h, %ecx       ; comparaison entre g et h
setl    %al           ; fixe le byte de poids faible de %eax
movzbl  %al, %ecx
movl    %ecx, r

movl    j, %ecx
cmpl    h, %ecx       ; comparaison entre j et h
sete    %al
```

```
movzbl  %al, %ecx
movl    %ecx, r
```

4.2.4 Les instructions de saut

Les instructions de saut sont des instructions de base pour tous les processeurs. Elles permettent de modifier la valeur du compteur de programme `%eip` de façon à modifier l'ordre d'exécution des instructions. Elles sont nécessaires pour implémenter les tests, les boucles et les appels de fonction. Les premiers langages de programmation et des langages tels que BASIC ou FORTRAN disposent d'une construction similaire avec l'instruction `goto`. Cependant, l'utilisation de l'instruction `goto` dans des programmes de haut niveau rend souvent le code difficile à lire et de nombreux langages de programmation n'ont plus de `goto` [Dijkstra1968]. Contrairement à Java, le C contient une instruction `goto`, mais son utilisation est fortement découragée. En C, l'instruction `goto` prend comme argument une étiquette (label en anglais). Lors de l'exécution d'un `goto`, le programme saute directement à l'exécution de l'instruction qui suit le label indiqué. Ceci est illustré dans l'exemple ci-dessous :

```
  int v=0;
  for(int i=0;i<SIZE;i++)
    for(int j=0;j<SIZE;j++) {
      if(m[i][j]>MVAL) {
        v=m[i][j];
        goto suite;
      }
    }
  printf("aucune valeur supérieure à %d\n",MVAL,v);
  goto fin;
 suite:
  printf("première valeur supérieure à %d : %d\n",MVAL,v);
 fin:
  return(EXIT_SUCCESS);
```

Si l'utilisation `goto` est en pratique prohibée dans la plupart des langages de programmation, en assembleur, les instructions de saut sont inévitables. L'instruction de saut la plus simple est `jmp`. Elle prend généralement comme argument une étiquette. Dans ce cas, l'exécution du programme après l'instruction `jmp` se poursuivra par l'exécution de l'instruction qui se trouve à l'adresse correspondant à l'étiquette fournie en argument. Il est également possible d'utiliser l'instruction `jmp` avec un registre comme argument. Ainsi, l'instruction `jmp *%eax` indique que l'exécution du programme doit se poursuivre par l'exécution de l'instruction se trouvant à l'adresse qui est contenue dans le registre `%eax`.

Il existe plusieurs variantes conditionnelles de l'instruction `jmp`. Ces variantes sont exécutées uniquement si la condition correspondante est vérifiée. Les variantes les plus fréquentes sont :
– `je` : saut si égal (teste le drapeau *ZF*) (inverse : `jne`)
– `js` : saut si négatif (teste le drapeau *SF*) (inverse : `jns`)
– `jg` : saut si strictement supérieur (teste les drapeaux *SF* et *ZF* et prend en compte un overflow éventuel) (inverse : `jl`)
– `jge` : saut si supérieur ou égal (teste le drapeaux *SF* et prend en compte un overflow éventuel) (inverse : `jle`)
Ces instructions de saut conditionnel sont utilisées pour implémenter notamment des expressions `if (condition) { ... } else { ... }` en C. Voici quelques traductions réalisées par un compilateur C en guise d'exemple.

```
if(j==0)
  r=1;

if(j>g)
  r=2;
else
  r=3;

if (j>=g)
  r=4;
```

Avant d'analyser la traduction de ce programme en assembleur, il est utile de le réécrire en utilisant l'instruction `goto` afin de s'approcher du fonctionnement de l'assembleur.

```
if(j!=0) { goto diff; }
  r=1;
diff:
  // suite

if(j<=g) { goto else; }
  r=2;
  goto fin;
else:
  r=3;
fin:
  // suite

if (j<g) { goto suivant; }
  r=4;

suivant:
```

Ce code C correspond assez bien au code assembleur produit par le compilateur.

```
       cmpl    $0, j      ; j==0 ?
       jne     .LBB2_2    ; jump si j!=0
       movl    $1, r      ; r=1
.LBB2_2:

       movl    j, %eax    ; %eax=j
       cmpl    g, %eax    ; j<=g ?
       jle     .LBB2_4    ; jump si j<=g

       movl    $2, r      ; r=2
       jmp     .LBB2_5    ; jump fin expression
.LBB2_4:
       movl    $3, r      ; r=3
.LBB2_5:

       movl    j, %eax    ; %eax=j
       cmpl    g, %eax    ; j<g ?
       jl      .LBB2_7    ; jump si j<g
       movl    $4, r      ; r=4
.LBB2_7:
```

Les instructions de saut conditionnel interviennent également dans l'implémentation des boucles. Plusieurs types de boucles existent en langage C. Considérons tout d'abord une boucle `while`.

```
while(j>0)
{
  j=j-3;
}
```

Cette boucle peut se réécrire en utilisant des `goto` comme suit.

```
debut:
  if(j<=0) { goto fin; }
  j=j-3;
  goto debut;
fin:
```

On retrouve cette utilisation des instructions de saut dans la traduction en assembleur de cette boucle.

```
.LBB3_1:
      cmpl    $0, j     ; j<=0
      jle     .LBB3_3   ; jump si j<=0
      movl    j, %eax
      subl    $3, %eax
      movl    %eax, j   ; j=j-3
      jmp     .LBB3_1
.LBB3_3:
```

Les boucles `for` s'implémentent également en utilisant des instructions de saut.

```
for(j=0;j<10;j++)
  { g=g+h; }

for(j=9;j>0;j=j-1)
  { g=g-h; }
```

La première boucle démarre par l'initialisation de la variable `j` à `0`. Ensuite, la valeur de cette variable est comparée avec `10`. L'instruction `jge` fait un saut à l'adresse mémoire correspondant à l'étiquette `.LBB4_4` si la comparaison indique que `j>=10`. Sinon, les instructions suivantes calculent `g=g+h` et `j++` puis l'instruction `jmp` relance l'exécution à l'instruction de comparaison qui est stockée à l'adresse de l'étiquette `.LBB4_1`.

```
      movl    $0, j     ; j=0
.LBB4_1:
      cmpl    $10, j
      jge     .LBB4_4   ; jump si j>=10
      movl    g, %eax   ; %eax=g
      addl    h, %eax   ; %eax+=h
      movl    %eax, g   ; g=%eax
      movl    j, %eax   ; %eax=j
      addl    $1, %eax  ; %eax++
      movl    %eax, j   ; j=%eax
      jmp     .LBB4_1
.LBB4_4:

      movl    $9, j     ; j=9
.LBB4_5:
      cmpl    $0, j
      jle     .LBB4_8   ; jump si j<=0
      movl    g, %eax
      subl    h, %eax
      movl    %eax, g
      movl    j, %eax   ; %eax=j
      subl    $1, %eax  ; %eax--
      movl    %eax, j   ; j=%eax
      jmp     .LBB4_5
.LBB4_8:
```

La seconde boucle est organisée de façon similaire.

4.2.5 Manipulation de la pile

Les instructions `mov` permettent de déplacer de l'information à n'importe quel endroit de la mémoire. A côté de ces instructions de déplacement, il y a des instructions qui sont spécialisées dans la manipulation de la pile. La pile, qui dans un processus Unix est stockée dans les adresses hautes est essentielle au bon fonctionnement des programmes. Par convention dans l'architecture [IA32], l'adresse du sommet de la pile est toujours stockée dans le registre `%esp`. Deux instructions spéciales permettent de rajouter et de retirer une information au sommet de la pile.

– `pushl %reg` : place le contenu du registre `%reg` au sommet de la pile et décrémente dans le registre `%esp` l'adresse du sommet de la pile de 4 unités.

– `popl %reg` : retire le mot de 32 bits se trouvant au sommet de la pile, le sauvegarde dans le registre `%reg` et incrémente dans le registre `%esp` l'adresse du sommet de la pile de 4 unités.

En pratique, ces deux instructions peuvent également s'écrire en utilisant des instructions de déplacement et des instructions arithmétiques. Ainsi, `pushl %ebx` est équivalent à :

```
subl $4, %esp       ; ajoute un bloc de 32 bits au sommet de la pile
movl %ebx, (%esp)   ; sauvegarde le contenu de %ebx au sommet
```

Tandis que `popl %ecx` est équivalent à :

```
movl (%esp), %ecx ; sauve dans %ecx la donnée au sommet de la pile
addl $4, %esp     ; déplace le sommet de la pile de 4 unites vers le haut
```

Pour bien comprendre le fonctionnement de la pile, il est utile de considérer un exemple simple. Imaginons la mémoire ci-dessous et supposons qu'initialement le registre `%esp` contient la valeur `0x0C` et que les registres `%eax` et `%ebx` contiennent les valeurs `0x02` et `0xFF`.

Adresse	Valeur
0x10	0x04
0x0C	0x00
0x08	0x00
0x04	0x00
0x00	0x00

```
push %eax ; %esp contient 0x08 et M[0x08]=0x02
push %ebx ; %esp contient 0x04 et M[0x04]=0xFF
pop %eax  ; %esp contient 0x08 et %eax 0xFF
pop %ebx  ; %esp contient 0x0C et %ebx 0x02
pop %eax  ; %esp contient 0x10 et %eax 0x00
```

4.2.6 Les fonctions et procédures

Les fonctions et les procédures sont essentielles dans tout langage de programmation. Une procédure est une fonction qui ne retourne pas de résultat. Nous commençons par expliquer comment les procédures peuvent être implémentées en assembleur et nous verrons ensuite comment implémenter les fonctions.

Une procédure est un ensemble d'instructions qui peuvent être appelées depuis n'importe quel endroit du programme. Généralement, une procédure est appelée depuis plusieurs endroits différents d'un programme. Pour comprendre l'implémentation des procédures, nous allons considérer des procédures de complexité croissante. Nos premières procédures ne prennent aucun argument. En C, elles peuvent s'écrire sous la forme de fonctions `void` comme suit.

```
int g=0;
int h=2;

void increase() {
 g=g+h;
}

void init_g() {
 g=1252;
}

int main(int argc, char *argv[]) {
  init_g();
  increase();
  return(EXIT_SUCCESS);
}
```

Ces deux procédures utilisent et modifient des variables globales. Nous verrons plus tard comment supporter les variables locales. Lorsque la fonction `main` appelle la procédure `init_g()` ou la procédure `increase`, il y a plusieurs opérations qui doivent être effectuées. Tout d'abord, le processeur doit transférer l'exécution du code à la première instruction de la procédure appelée. Cela se fait en associant une étiquette à chaque procédure qui correspond à l'adresse de la première instruction de cette procédure en mémoire. Une instruction de saut telle que `jmp` pourrait permettre de démarrer l'exécution de la procédure. Malheureusement, ce n'est pas suffisant car après son exécution la procédure doit pouvoir poursuivre son exécution à l'adresse de l'instruction qui suit celle d'où elle a été appelée. Pour cela, il est nécessaire que la procédure qui a été appelée puisse connaître l'adresse de l'instruction qui doit être exécutée à la fin de son exécution. Dans l'architecture [IA32], cela se fait en utilisant la pile. Vu l'importance des appels de procédure et de fonctions, l'architecture [IA32] contient deux instructions dédiés pour implémenter ces appels. L'instruction `call` est une instruction de saut qui (1) sauvegarde sur la pile l'adresse de l'instruction qui suit l'instruction `call` puis (2) transfère l'exécution à l'adresse de l'étiquette passée en argument. L'adresse sauvée sur la pile est est l'adresse à laquelle la procédure doit revenir après son exécution. L'instruction `call` est donc équivalente à une instruction `push` suivie d'une instruction `jmp`. L'instruction `ret` est également une instruction de saut. Elle suppose que l'adresse de retour se trouve au sommet de la pile, retire cette adresse de la pile et effectue un saut à cette adresse. Elle est donc équivalente à une instruction `pop` suivie d'une instruction `jmp`. Dans l'architecture [IA32], le registre `%esp` contient en permanence le sommet de la pile. Les instructions `call` et `ret` modifient donc la valeur de ce registre lorsqu'elles sont exécutées. En assembleur, le programme ci-dessus se traduit comme suit :

```
increase:                   ; étiquette de la première instruction
            movl    g, %eax
            addl    h, %eax
            movl    %eax, g
            ret             ; retour à l'endroit qui suit l'appel
init_g:                     ; étiquette de la première instruction
            movl    $1252, g
            ret             ; retour à l'endroit qui suit l'appel
main:
            subl    $12, %esp
            movl    20(%esp), %eax
            movl    16(%esp), %ecx
            movl    $0, 8(%esp)
            movl    %ecx, 4(%esp)
            movl    %eax, (%esp)
            calll   init_g   ; appel à la procédure init_g
A_init_g:   calll   increase ; appel à la procédure increase
A_increase: movl    $0, %eax
            addl    $12, %esp
            ret             ; fin de la fonction main
g:                          ; étiquette, variable globale g
      .long   0             ; initialisée à 0
h:                          ; étiquette, variable globale g
      .long   2             ; initialisée à 2
```

Dans ce code assembleur, on retrouve dans le bas du code la déclaration des deux variables globales, `g` et `h` et leurs valeurs initiales. Chaque procédure a son étiquette qui correspond à l'adresse de sa première instruction. La fonction `main` débute par une manipulation de la pile qui ne nous intéresse pas pour le moment. L'appel à la procédure `init_g()` se fait via l'instruction `calll init_g` qui place sur la pile l'adresse de l'étiquette `A_init_g`. La procédure `init_g()` est très simple puisqu'elle comporte une instruction `movl` qui permet d'initialiser la variable `g` suivie d'une instruction `ret`. Celle-ci retire de la pile l'adresse `A_init_g` qui y avait été placée par l'instruction `call` et poursuit l'exécution du programme à cette adresse. L'appel à la procédure `increase` se déroule de façon similaire.

Considérons une petite variante de notre programme C dans lequel une procédure `p` appelle une procédure `q`.

```
int g=0;
int h=2;

void q() {
 g=1252;
```

```
}

void p() {
  q();
  g=g+h;
}



int main(int argc, char *argv[]) {
  p();
  return(EXIT_SUCCESS);
}
```

La compilation de ce programme produit le code assembleur suivant pour les procédures `p` et `q`.

```
q:
        movl    $1252, g
        ret                     ; retour à l'appelant
p:
        subl    $12, %esp       ; réservation d'espace sur pile
        calll   q               ; appel à la procédure q
        movl    g, %eax
        addl    h, %eax
        movl    %eax, g
        addl    $12, %esp       ; libération espace réservé sur pile
        ret                     ; retour à l'appelant
```

La seule différence par rapport au programme précédent est que la procédure `p` descend le sommet de la pile de 12 unités au début de son exécution et l'augmente de 12 unités à la fin. Ces manipulations sont nécessaires pour respecter une convention de l'architecture [IA32] qui veut que les adresses de retour des procédures soient alignées sur des blocs de 16 bytes.

Considérons maintenant une procédure qui prend un argument. Pour qu'une telle procédure puisse utiliser un argument, il faut que la procédure appelante puisse placer sa valeur à un endroit où la procédure appelée peut facilement y accéder. Dans l'architecture [IA32], c'est la pile qui joue ce rôle et permet le passage des arguments. En C, les arguments sont passés par valeur et ce sera donc les valeurs des arguments qui seront placées sur la pile. A titre d'exemple, considérons une procédure simple qui prend deux arguments entiers.

```
int g=0;
int h=2;
void init(int i, int j) {
 g=i;
 h=j;
}

int main(int argc, char *argv[]) {
  init(1252,1);
  return(EXIT_SUCCESS);
}
```

Le passage des arguments de la fonction `init` depuis la fonction `main` se fait en les plaçant sur la pile avec les instructions `movl $1252, (%esp)` et `movl $1, 4(%esp)` qui précèdent l'instruction `call init`. Le premier argument est placé au sommet de la pile et le second juste au-dessus. La fonction `main` sauvegarde d'autres registres sur la pile avant l'appel à `init`. Ces sauvegardes sont nécessaires car la fonction `main` ne sait pas quels registres seront modifiés par la fonction qu'elle appelle. En pratique l'architecture [IA32] définit des conventions d'utilisation des registres. Les registres `%eax`, `%edx` et `%ecx` sont des registres qui sont sous la responsabilité de la procédure appellante (dans ce cas `main`). Une procédure appelée (dans ce cas-ci `init`) peut modifier sans restrictions les valeurs de ces registres. Si la fonction appellante souhaite pouvoir utiliser les valeurs stockées dans ces registres après l'appel à la procédure, elle doit les sauvegarder elle-même sur la pile. C'est ce que fait la fonction `main` pour `%eax`, `%edx` et `%ecx`. Inversement, les registres `%ebx`, `%edi` et `%esi` sont des registres qui doivent être sauvés par la procédure appelée si celle-ci les modifie. La procédure `init` n'utilisant

pas ces registres, elle ne les sauvegarde pas. Par contre, la fonction `main` débute en sauvegardant le registre `%esi` sur la pile.

```
init:
      subl    $8, %esp          ; réservation d'espace sur la pile
      movl    16(%esp), %eax    ; récupération du second argument
      movl    12(%esp), %ecx    ; récupération du premier argument
      movl    %ecx, 4(%esp)     ; sauvegarde sur la pile
      movl    %eax, (%esp)      ; sauvegarde sur la pile
      movl    4(%esp), %eax     ; chargement de i
      movl    %eax, g           ; g=i
      movl    (%esp), %eax      ; chargement de j
      movl    %eax, h           ; h=j
      addl    $8, %esp          ; libération de l'espace réservé
      ret
main:
      pushl   %esi
      subl    $40, %esp
      movl    52(%esp), %eax
      movl    48(%esp), %ecx
      movl    $1252, %edx
      movl    $1, %esi
      movl    $0, 36(%esp)
      movl    %ecx, 32(%esp)
      movl    %eax, 24(%esp)
      movl    $1252, (%esp)     ; premier argument sur la pile
      movl    $1, 4(%esp)       ; deuxième argument sur la pile
      movl    %esi, 20(%esp)
      movl    %edx, 16(%esp)
      calll   init              ; appel à init
      movl    $0, %eax
      addl    $40, %esp
      popl    %esi
      ret
```

La différence entre une procédure et une fonction est qu'une fonction retourne un résultat. Considérons le programme suivant et les fonctions triviales `int init()` et `int sum(int, int)`. Pour que de telles fonctions puissent s'exécuter et retourner un résultat, il faut que la procédure appelante puisse savoir où aller chercher le résultat après exécution de l'instruction `ret`.

```
int g=0;
int h=2;

int init() {
  return 1252;
}

int sum(int a, int b) {
  return a+b;
}

int main(int argc, char *argv[]) {
  g=init();
  h=sum(1,2);
  return(EXIT_SUCCESS);
}
```

La compilation du programme C ci-dessus en assembleur produit le code suivant. Dans l'architecture [IA32], la valeur de retour d'une fonction est stockée par convention dans le registre `%eax`. Cette convention est particulièrement visible lorsque l'on regarde les instructions générées pour la fonction `int init()`. La fonction `sum` retourne également son résultat dans le registre `%eax`.

```
init:
    movl    $1252, %eax
    ret
sum:
    subl    $8, %esp            ; réservation d'espace sur la pile
    movl    16(%esp), %eax      ; récupération du second argument
    movl    12(%esp), %ecx      ; récupération du premier argument
    movl    %ecx, 4(%esp)
    movl    %eax, (%esp)
    movl    4(%esp), %eax       ; %eax=a
    addl    (%esp), %eax        ; %eax=a+b
    addl    $8, %esp            ; libération de l'espace réservé
    ret
main:
    subl    $28, %esp
    movl    36(%esp), %eax
    movl    32(%esp), %ecx
    movl    $0, 24(%esp)
    movl    %ecx, 20(%esp)      ; sauvegarde sur la pile
    movl    %eax, 16(%esp)      ; sauvegarde sur la pile
    calll   init
    movl    $1, %ecx
    movl    $2, %edx
    movl    %eax, g
    movl    $1, (%esp)          ; premier argument
    movl    $2, 4(%esp)         ; second argument
    movl    %ecx, 12(%esp)      ; sauvegarde sur la pile
    movl    %edx, 8(%esp)       ; sauvegarde sur la pile
    calll   sum
    movl    $0, %ecx
    movl    %eax, h
    movl    %ecx, %eax
    addl    $28, %esp
    ret
```

Pour terminer notre exploration de la compilation de fonctions C en assembleur, considérons une fonction récursive. Par simplicité, nous utilisons la fonction `sumn` qui calcule de façon récursive la somme des n premiers entiers.

```
int sumn(int n) {
  if(n<=1)
    return n;
  else
    return n+sumn(n-1);
}
```

Lorsque cette fonction récursive est compilée, on obtient le code ci-dessous. Celui-ci démarre par réserver une zone de 28 bytes sur la pile et récupère ensuite l'argument qui est placé dans le registre `%eax`. Cet argument est utilisé comme variable locale, il est donc sauvegardé sur la pile de la fonction `sumn` dans la zone qui vient d'être réservée. Ensuite, on compare la valeur de l'argument avec `1`. Si l'argument est inférieur ou égal à `1`, on récupère la variable locale sur la pile et on la sauve à un autre endroit en préparation à la fin du code (étiquette `.LBB1_3`) ou elle sera placée dans le registre `%eax` avant l'exécution de l'instruction `ret`. Sinon, l'appel récursif est effectué. Pour cela, il faut d'abord calculer `n-1`. Cette valeur est stockée dans le registre `%ecx` puis placée sur la pile avant l'appel récursif. Comme un appel de fonction ne préserve pas `%eax` et que cette valeur est nécessaire après l'appel récursif, elle est sauvegardée sur la pile. La première instruction qui suit l'exécution de l'appel récursif récupère la valeur de la variable `n` sur la pile et la place dans le registre `%ecx`. Le résultat de l'appel récursif étant placé dans `%eax`, l'instruction `addl %ecx, %eax` calcule bien `n+sum(n-1)`. Ce résultat est placé sur la pile puis récupéré et placé dans `%eax` avant l'exécution de `ret`. Il faut noter que les 28 bytes qui avaient étés ajoutés à la pile au début de la fonction sont retirées par l'instruction `addl $28, %esp`. C'est nécessaire pour que la pile soit bien préservée lors de l'appel à une fonction.

```
sumn:
     subl    $28, %esp          ; réservation d'espace sur la pile
     movl    32(%esp), %eax     ; récupération argument
     movl    %eax, 20(%esp)     ; sauvegarde sur pile
     cmpl    $1, 20(%esp)
     jg      .LBB1_2            ; jump si n>1
     movl    20(%esp), %eax     ; récupération n
     movl    %eax, 24(%esp)
     jmp     .LBB1_3
.LBB1_2:
     movl    20(%esp), %eax
     movl    20(%esp), %ecx
     subl    $1, %ecx           ; %ecx=n-1
     movl    %ecx, (%esp)       ; argument sur pile
     movl    %eax, 16(%esp)
recursion:
     calll   sumn
     movl    16(%esp), %ecx     ; récupération de n
     addl    %ecx, %eax         ; %eax=%eax+n
     movl    %eax, 24(%esp)
.LBB1_3:
     movl    24(%esp), %eax
     addl    $28, %esp             ; libération de l'espace réservé sur la pile
     ret
```

Ce code illustre la complexité de supporter des appels récursifs en C et le coût au niveau de la gestion de la pile notamment. Ces appels récursifs doivent être réservés à des fonctions où l'appel récursif apporte une plus value claire.

CHAPITRE 5

Organisation d'un système d'exploitation

5.1 Structure du système d'exploitation

L'objectif de ce chapitre est de présenter les grands principes de mise en œuvre d'un système d'exploitation (SE), que nous avons déjà survolé dans le chapitre d'introduction. On commencera par présenter les différent types de services fournis par un système d'exploitation, pour ensuite détailler l'interface offerte par le noyau aux programmes et librairies oeuvrant en espace utilisateur. Cette interface est composée d'*appels systèmes* : nous en nous étudierons la mise en œuvre et l'utilisation. Nous verrons ensuite différentes manières de structurer logiquement le *noyau* du système d'exploitation, et discuterons des avantages et inconvénients associés à ces méthodes. Nous aborderons enfin le processus de démarrage du SE.

5.1.1 Services

Le but du système d'exploitation est de rendre des services aux applications, aux développeurs et aux utilisateurs. Ces services permettent d'exploiter les ressources matérielles de manière simple au travers d'abstractions, mettant en œuvre la virtualisation des ressources comme nous l'avons vu dans notre introduction.

Services aux utilisateurs

Une première catégorie de service est l'*interface* apportée aux utilisateurs. Nous avons déjà étudié le principe de la ligne de commande. De nombreux systèmes proposent une interface utilisateur graphique, comme vous pouvez l'utiliser tous les jours avec un système comme Windows ou Mac OS, ou en utilisant l'un des nombreux gestionnaires de fenêtres disponibles dans les distributions GNU/Linux, comme Gnome ou KDE.

Un troisième type d'interface existe : il s'agit de l'accès en mode *batch* (ou en mode de *traitement par lot* en français). Avec une telle interface, l'utilisateur n'utilise pas le système de manière interactive. À la place, les demandes d'exécution de programmes (incluant le nom du programme, ses arguments, et les données à utiliser en entrée) sont envoyées à un gestionnaire de traitement par lot, qui se chargera de l'exécuter plus tard, typiquement lorsque les ressources matérielles nécessaires seront disponibles et en suivant une politique FIFO.

Les interfaces de traitement par lots étaient courantes sur les premiers ordinateurs ne mettant pas en œuvre le principe de temps partagé. L'entrée du programme et de ses données se faisant en effet manuellement. Le rôle du gestionnaire de traitement par lot était alors tenu par un humain. Ce mode d'utilisation reste de nos jours la plus utilisé dans les centres de calcul à haute performance, par exemple pour réaliser des simulations de processus physiques ou biologiques.

Note : Planifier l'exécution de programmes dans un système interactifs

Lorsque l'utilisateur envoie une commande au *shell*, le programme correspondant est exécuté immédiatement, contrairement au mode de traitement par lot. Il peut être parfois nécessaire de prévoir à l'avance l'exécution d'une commande, ou bien de réaliser cette exécution de façon périodique. Par exemple, un utilitaire préparant un rapport sur l'utilisation des ressources du système peut être déclenché chaque nuit et envoyer un message à l'administrateur si un quote d'utilisation est dépassé.

La commande crontab(1) permet de planifier l'exécution périodique d'une commande. La commande at(1posix) permet quand à elle de demander l'exécution d'une commande à un temps absolu ou relatif, donné, comme dans l'exemple qui suit.

```
at -m 0730 tomorrow
sort < file >outfile
EOT
```

Services aux concepteurs d'applications

Le système d'exploitation fournit par ailleurs des services aux développeurs d'applications. Ces services permettent de faciliter la création et l'amélioration des applications (et en particulier, de leur fiabilité et de leur performance).

Le système d'exploitation fournit tout d'abord des services permettant de faciliter l'assemblage et le chargement d'un programme en mémoire. Un premier utilitaire appelé le *linker* permet d'assembler les fichiers objets générés par le compilateur ainsi les librairies statiques afin d'obtenir un programme exécutable. Nous avons couvert les principes du *linker* dans la partie du syllabus traitant des grands programmes en C. Le programme ld(1) joue ce rôle sous GNU/Linux. Un deuxième utilitaire est nécessaire lors de l'exécution du programme : le *loader*. Il réalise deux opérations : (1) la mise en place de l'espace mémoire du programme, la réservation des segments, et leur remplissage à partir du fichier exécutable et (2) le chargement dans cet espace mémoire des librairies dynamiques. Le loader sous GNU/Linux est une combinaison de fonctions réalisées par le *noyau* (pour la création de l'espace mémoire virtuel principalement) et par la librairie ld.so(8). Le loader pré-assigne les différentes sections de l'espace mémoire à partir du fichier programme (section text et sections de données initialisées et non initialisées), prépare le contenu de la section de variables d'environnement, et effectue le chargement des librairies dynamiques. À la fin de l'exécution, les ressources utilisées par le processus créé sont libérées et le système d'exploitation gère la récupération du code de retour du programme.

Lorsqu'un processus est en cours d'exécution, le système d'exploitation peut permettre d'observer voire de contrôler celui-ci pour permettre la compréhension de son comportement et faciliter sa mise au point.

Tout d'abord, une erreur peut arriver dans le programme. Le système d'exploitation permet alors de récupérer des informations sur l'erreur elle-même, ainsi que son contexte d'apparition (voir par exemple l'ensemble du contenu de la mémoire au moment de son occurence). Des exemples d'erreurs classiques sont listées ci-dessous.

– L'accès à un segment de mémoire non autorisé, si par exemple le programme essaie de lire une adresse au dessus de la limite du stack (et donc avant que celle-ci ne soit étendue avec un appel à sbrk(2)), ou bien essaie de lire une adresse d'un des segments réservés du système d'exploitation au début ou à la fin de l'espace mémoire du processus, ou encore essaie d'*écrire* à une des adresses du *segment text*.
– L'utilisation d'une opération arithmétique non supportée, comme par exemple une division par 0.
– L'utilisation en mode utilisateur d'une instruction autorisée seulement en mode protégé.

Enfin, le système d'exploitation fournit des services facilitant le déboguage des applications, au delà de la simple récolte d'information lors de l'occurence d'erreurs. Un déboggueur comme gdb(1) permet ainsi d'observer l'exécution d'un processus, de la stopper lorsqu'une adresse d'instruction spécifique est atteinte (on parle de point d'arrêt ou *breakpoint* en anglais) ou même d'exécuter les instructions pas à pas (une par une). Le déboggueur est un processus comme un autre. Il est donc isolé des autres processus. Il a pour cette raison besoin de services spécifiques fournis par le *noyau* du système d'exploitation, pour pouvoir inspecter ou modifier l'espace mémoire du processus observé. Un exemple de service nécessaire est de pouvoir faire la demande au processeur qu'une interruption logicielle (*trap*) soit générée automatiquement lors de l'atteinte d'un point d'arrêt (i.e., l'adresse d'une instruction spécifique dans le segment text) ou même après chaque instruction. La configuration du processeur à ces fins est une opération qui requiert l'utilisation d'instructions seulement autorisées en mode protégé.

Services aux applications

Le système d'exploitation fournit des services aux applications en leur permettant d'exploiter de façon efficace, aisée et portable, les ressources matérielles. Nous avons abordé en introduction les ressources virtualisées fondamentales que sont la notion de processus ou la notion de mémoire virtuelle. Nous survolons ici des exemples d'autres services. Nous verrons la mise en œuvre des plus importants d'entre eux plus tard dans ce cours.

Le système d'exploitation fournit pour commencer des services pour permettre l'utilisation d'*entrées/sorties*. Comme nous l'avons vu en introduction, les gestionnaires de périphériques (connectés à un bus d'entrée/sortie) génèrent des interruptions permettant de prévenir le processeur de la disponibilité d'une donnée à traiter. De la même manière, le processeur peut envoyer des commandes au gestionnaire de périphérique pour initier une opération d'entrée sortie. Il n'est bien évidemment pas souhaitable de laisser les applications gérer ces opérations elles-même. Les instructions correspondantes sont ainsi réservées au mode protégé du processeur. Le système d'exploitation fournit donc des services d'entrée/sortie donc la spécification dépend de la nature du système d'entrée/sortie considéré (adaptateur réseau, adaptateur graphique, etc.), au travers d'abstractions facilement manipulables par un programmeur.

Note : Les drivers de périphériques

Bien que le système d'exploitation fournisse aux applications une abstraction unique pour une même classe de périphériques, ces périphériques sont de mise en œuvre matérielle variées et ne répondent pas toujours au même jeu de commandes, même lorsqu'ils ont le même objectif. Par exemple, un adaptateur réseau d'une marque ou d'une génération donnée pourra répondre à des commandes de contrôle qu'un autre adaptateur réseau ne supportera pas. Pour pallier cette hétérogénéité, le *noyau* du système d'exploitation utilise des *drivers de périphériques*. Ces modules logiciel très bas niveau reçoivent des commandes d'entrée/sortie générique en entrée, et les traduisent en des commandes spécifique à un matériel donné. Ils sont le plus souvent développés par l'entreprise fabriquant le matériel spécifique. Leur mise en œuvre nécessite fréquemment l'utilisation du langage d'assemblage.

Partage de ressources

Les services fournis aux applications, aux développeurs et aux utilisateurs permettent l'utilisation simplifiée mais aussi *mutualisée* des resources matérielles. Plusieurs utilisateurs peuvent ainsi utiliser le même système simultanément, et chaque utilisateur peut utiliser plusieurs applications. Un rôle majeur du système d'exploitation dans ce contexte est la mise en œuvre du partage des resources, en visant plusieurs objectifs :

– On souhaite que les ressources soient utilisées de façon efficace afin de maximiser l'utilité du système. Par exemple, il n'est pas toujours souhaitable qu'un processus en attente de la fin d'une opération d'entrée/sortie occupe le processeur à exécuter une boucle d'attente active (i.e., une boucle `while` vérifiant de façon répétée qu'une donnée soit disponible pour être consommée, et ce jusqu'à ce soit le cas).
– Les resources partagées doivent l'être de manière équitable, ou tout au moins qui suive les règles de priorité qui ont été édictées pour ce système.
– Enfin, il est nécessaire d'isoler l'accès aux ressources utilisées par un processus et/ou un utilisateur des autres ressources.

Le partage des ressources nécessitent donc des services spécifiques permettant :

– L'allocation des ressources. Certaines ressources peuvent être disponibles de manière exclusive (par exemple, les entrées clavier ne doivent être visibles que par un processus précis) ou de manière partagée (par exemple, l'adaptateur réseau reçoit et envoie des données pour plusieurs processus).
– Le contrôle d'usage, afin de savoir quel processus et/ou quel utilisateur utilise quelle quantité de ressources.
– La protection d'accès, afin de contrôler si un programme ou un utilisateur a l'autorisation ou non d'utiliser une ressource.

5.1.2 Appels systèmes

Outre l'utilisation de fonctions de librairies, les programmes doivent donc interagir avec le système d'exploitation afin d'utiliser les services que celui ci fournit.

Un système d'exploitation comme Unix comprend à la fois des utilitaires comme grep(1), ls(1), ... qui sont directement exécutables depuis le shell et un noyau ou *kernel* en anglais. Le *kernel* contient les fonctions de base

du système d'exploitation qui lui permettent à la fois d'interagir avec le matériel mais aussi de gérer les processus des utilisateurs. En pratique, le kernel peut être vu comme étant un programme spécial qui est toujours présent en mémoire. Parmi l'ensemble des fonctions contenues dans le *kernel*, il y en a un petit nombre, typiquement de quelques dizaines à quelques centaines, qui sont utilisables par les processus lancés par les utilisateurs. Ce sont les appels système. Un *appel système* est une fonction du *kernel* qui peut être appelée par n'importe quel processus. Comme nous l'avons vu lorsque nous avons décrit le fonctionnement du langage d'assemblage, l'exécution d'une fonction dans un programme comprend plusieurs étapes :

1. Placer les arguments de la fonction à un endroit (la pile) où la fonction peut y accéder
2. Sauvegarder sur la pile l'adresse de retour
3. Modifier le registre `%eip` de façon à ce que la prochaine instruction à exécuter soit celle de la fonction à exécuter
4. La fonction récupère ses arguments (sur la pile) et réalise son calcul
5. La fonction sauve son résultat à un endroit (`%eax`) convenu avec la fonction appelante
6. La fonction récupère l'adresse de retour sur la pile et modifie `%eip` de façon à retourner à la fonction appelante

L'exécution d'un appel système comprend les mêmes étapes avec une différence importante c'est que le flux d'exécution des instructions doit passer du programme utilisateur au noyau du système d'exploitation. Pour comprendre le fonctionnement et l'exécution d'un appel système, il est utile d'analyser les six points mentionnés ci-dessus.

Le premier problème à résoudre pour exécuter un appel système est de pouvoir placer les arguments de l'appel système dans un endroit auquel le *kernel* pourra facilement accéder. Il existe de nombreux appels systèmes avec différents arguments. La liste complète des appels systèmes est reprise dans la page de manuel syscalls(2). La table ci-dessous illustre quelques appels systèmes et leurs arguments.

Appel système	Arguments
getpid(2)	`void`
fork(2)	`void`
read(2)	`int fildes, void *buf, size_t nbyte`
kill(2)	`pid_t pid, int sig`
brk(2)	`const void *addr`

Sous Linux, les arguments d'un appel système sont placés par convention dans des registres. Sur [IA32], le premier argument est placé dans le registre `%ebx`, le second dans `%ecx`, ... Le *kernel* peut donc facilement récupérer les arguments d'un appel système en lisant le contenu des registres.

Le second problème à résoudre est celui de l'adresse de retour. Celle-ci est automatiquement sauvegardée lors de l'exécution de l'instruction qui fait appel au kernel, tout comme l'instruction `calll` sauvegarde directement l'adresse de retour d'une fonction appelée sur la pile.

Le troisième problème à résoudre est de passer de l'exécution du processus utilisateur à l'exécution du *kernel*. Comme abordé dans l'introduction, les processeurs actuels peuvent fonctionner dans au minimum deux modes : le *mode utilisateur* et le *mode protégé*. Lorsque le processeur fonctionne en mode protégé, toutes les instructions du processeur et toutes les adresses mémoire sont utilisables. Lorsqu'il fonctionne en mode utilisateur, quelques instructions spécifiques de manipulation du matériel et certaines adresses mémoire ne sont pas utilisables. Cette division en deux modes de fonctionnement permet d'avoir une séparation claire entre le système d'exploitation et les processus lancés par les utilisateurs. Le noyau du système d'exploitation s'exécute en mode protégé et peut donc utiliser entièrement le processeur et les dispositifs matériels de l'ordinateur. Les processus utilisateurs par contre s'exécutent en mode utilisateur. Ils ne peuvent donc pas directement exécuter les instructions permettant une interaction avec des dispositifs matériels. Cette interaction doit passer par le noyau du système d'exploitation qui sert de médiateur et vérifie la validité des demandes faites par un processus utilisateur.

Les transitions entre les modes protégé et utilisateur sont importantes car elles rythment le fonctionnement du système d'exploitation. Lorsque l'ordinateur démarre, le processeur est placé en mode protégé et le *kernel* se charge. Il initialise différentes structures de données et lance init(8) le premier processus du système. Dès que init(8) a été lancé, le processeur passe en mode utilisateur et exécute les instructions de ce processus. Après cette phase de démarrage, les instructions du *kernel* seront exécutées lorsque soit une interruption matérielle surviendra ou qu'un processus utilisateur exécutera un appel système. L'interruption matérielle place automatiquement le processeur en mode protégé et le *kernel* exécute la routine de traitement d'interruption correspondant à l'interruption

qui est apparue. Un appel système démarre par l'exécution d'une instruction spéciale (parfois appelée interruption logicielle) qui place le processeur en mode protégé puis démarre l'exécution d'une instruction placée à une adresse spéciale en mémoire. Sur certains processeurs de la famille [IA32], l'instruction `int 0x80` permet ce passage du mode utilisateur au mode protégé. Sur d'autres processeurs, c'est l'instruction `syscall` qui joue ce rôle. L'exécution de cette instruction est la seule possibilité pour un programme d'exécuter des instructions du *kernel*. En pratique, cette instruction fait passer le processeur en mode protégé et démarre l'exécution d'une routine spécifique du *kernel* et qui en est l'unique point d'entrée. Cette routine commence par sauvegarder le contexte du processus qui exécute l'appel système demandé. Chaque appel système est identifié par un nombre entier. Le *kernel* contient une table avec, pour chaque appel système, l'adresse de la fonction à exécuter. En pratique, le numéro de l'appel système à exécuter est placé par le processus appelant dans le registre `%eax`.

L'appel système peut donc s'exécuter en utilisant les arguments qui se trouvent dans les différents registres. Lorsque l'appel système se termine, le résultat est placé dans le registre `%eax` et une instruction spéciale permet de retourner en mode utilisateur et d'exécuter dans le processus appelant l'instruction qui suit celle qui a provoqué l'exécution de l'appel système. Si l'appel système a échoué, le *kernel* doit aussi mettre à jour le contenu de `errno` avant de retourner au processus appelant.

Ces opérations sont importantes pour comprendre le fonctionnement d'un système informatique et la différence entre un appel système et une fonction de la librairie. En pratique, la librairie cache cette complexité au programmeur en lui permettant d'utiliser des fonctions de plus haut niveau **[#fsyscall]_** . Cependant, il faut être conscient que ces fonctions s'appuient elles-même sur des appels systèmes pour s'exécuter. Ainsi par exemple, la fonction printf(3) utilise l'appel système write(2) pour écrire sur la sortie standard. La commande strace(1) permet de tracer l'ensemble des appels systèmes faits par un processus. A titre d'exemple, voici les appels systèmes effectués par le programme "Hello world" du début de la présentation du langage C, et repris ci-dessous.

```
#include <stdio.h>
#include <stdlib.h>

int main(int argc, char *argv[])
{
   printf("Hello, world! %d\n",sizeof(int));

   return EXIT_SUCCESS;
}
```

```
$ strace ./helloworld_s
execve("./helloworld_s", ["./helloworld_s"], [/* 21 vars */]) = 0
uname({sys="Linux", node="precise32", ...}) = 0
brk(0)                                  = 0x9e8b000
brk(0x9e8bd40)                          = 0x9e8bd40
set_thread_area({entry_number:-1 -> 6, base_addr:0x9e8b840, limit:1048575, seg_32bit:1, contents:0
brk(0x9eacd40)                          = 0x9eacd40
brk(0x9ead000)                          = 0x9ead000
fstat64(1, {st_mode=S_IFCHR|0620, st_rdev=makedev(136, 0), ...}) = 0
mmap2(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0xb778a000
write(1, "Hello, world! 4\n", 16Hello, world! 4
)       = 16
exit_group(0)                           = ?
```

Il n'est pas nécessairement utile de comprendre l'intégralité de ces lignes, mais on peut y déceler les points d'intérêt suivants :

– Le premier appel système execve(2) prend comme argument le programme à exécuter ;
– Les appels système brk(2), set_thread_area(2) ou mmap2(2) sont utilisées par le chargeur de programme (*loader*) pour mettre en place l'espace mémoire du processus ;
– Enfin, l'appel write(2) est utilisé pour envoyer vers *STDOUT* la chaîne de caractères formatée par la fonction correspondante de la librairie standard, printf(3).

Si, dans cet exemple, on voit une correspondance assez directe entre une fonction de la librairie standard (printf(3)) et un appel système, certaines fonctions de la librairie, ou bien certains utilitaires, utilisent de très nombreux appels système pour réaliser leur fonction. Pour reprendre l'exemple cité précédemment du débogueur gdb, celui-ci va effectuer de nombreux appels systèmes au services du *noyau* permettant le contrôle d'un processus en cours d'exécution, en en particulier l'appel ptrace(2).

5.2 Architecture logicielle des systèmes d'exploitation

Nous avons vu que l'interface entre les programmes en mode utilisateur (y compris les programmes utilitaires du système d'exploitation) et le noyau de ce système d'exploitation, utilise le principe d'appel système. Nous avons par ailleurs vu que les gestionnaires de périphériques, au plus bas niveau, utilisent des composants logiciels spécifiques au matériel utilisé, les drivers de périphériques.

Nous allons nous intéresser dans cette section à la mise en œuvre du noyau lui-même et de ses fonctions associées. Il n'existe pas d'approche universelle et idéale *dans tous les cas* pour structurer un système d'exploitation. Le choix d'une architecture logicielle spécifique est dictée par plusieurs contraintes, dont certaines peuvent être contradictoires :

– Des contraintes matérielles, et en particulier le support par le processeur de mécanismes efficaces permettant des abstractions de haut niveau (comme la mémoire virtuelle) ou le support de l'isolation entre programmes (présence des modes protégé/utilisateur ou non, par exemple).
– De la performance et du coût à l'exécution des services systèmes.
– De la consommation de ressources du système, en particulier en termes d'occupation mémoire.
– De la facilité d'évolution du système d'exploitation par l'ajout de nouvelles fonctionnalités, le support de nouveau matériel, ou sa capacité à être adapté à des contextes d'utilisation différents.
– De sa fiabilité et de la facilité de sa maintenance et de son déboguage.

Nous allons illustrer quelques unes des possibilités en utilisant quelques exemples.

5.2.1 Un système simple : MS-DOS

MS-DOS a été dans les années 1980 et une bonne partie des années 1990 le système d'exploitation principalement utilisé pour les ordinateurs de type IBM-PC et compatibles. Ce système d'exploitation ne fait pas partie de la famille UNIX.

Le système MS-DOS visait une utilisation mono-utilisateur et mono-application. Il ne met donc pas en œuvre le concept de temps partagé, et n'a donc pas besoin de supporter une isolation forte entre plusieurs applications ou même entre les applications et le noyau. Les processeurs supportés par le système MS-DOS, du type Intel 8086 et compatibles (80286, 80386, 80486) n'offraient de toutes façons pas tous un support complet pour l'isolation entre un mode d'exécution protégé pour le noyau et un mode utilisateur. En revanche, les systèmes visés avaient des contraintes très fortes en termes de mémoire disponible : le système d'exploitation doit donc tenir dans le moins d'instructions possibles pour réserver le reste de la mémoire pour les applications.

Le système MS-DOS original a donc été mis en œuvre de façon monolithique, sans séparation claire des fonctionnalités et services, et sans support réel pour la modularité. Le processus unique de l'application, ainsi que le code du noyau, résident dans le même espace mémoire. L'utilisation des appels systèmes utilise le principe d'interruption avec le passage des arguments dans les registres mais l'isolation entre la mémoire de l'application et celle du noyau n'est pas assurée (par exemple, l'application peut lire les structures de données manipulées par le noyau). Les applications peuvent, par ailleurs, accéder directement aux gestionnaires de périphériques.

Note : Quand un système simple et concis devient la base d'une industrie

Le système MS-DOS a été originalement conçu pour des ordinateurs aux capacités très limités au début des années 1980. On comprend, dès lors, la volonté de rendre le code le plus petit et le plus simple possible. MS-DOS est un bon exemple de logiciel qui n'a pas été pensé à la base pour être étendu et adapté à des ordinateurs plus complexes ou avec plus de ressources, mais qui a eu une durée de vie importante pour des raisons commerciales, et ce bien au delà des intentions initiales. Ce manque de structuration et d'isolation originel a eu des conséquences importantes sur la complexité et l'évolution des systèmes informatiques de type PC. Par exemple, lors de la conception de MS-DOS, l'espace mémoire disponible a été fixé à une capacité maximale de 640 Kilo-octets. L'utilisation de mémoire supplémentaire a été rendu possible par la suite grâce à un mécanisme dit de *mémoire étendue* dont l'utilisation n'est pas transparente pour l'application, ce qui rend la programmation inutilement complexe. L'absence d'une structure claire et de propriétés d'isolation a aussi été la source d'un grand nombre de vulnérabilités et de problèmes de sécurité dans MS-DOS et les systèmes s'y appuyant, comme les premiers systèmes Microsoft Windows.

5.2.2 Les systèmes monolithiques multi-utilisateurs : UNIX

Les premières version du système d'exploitation UNIX visaient une utilisation en partage de temps entre plusieurs applications et plusieurs utilisateurs. Le support pour l'isolation entre les applications (les processus) était donc primordial. Le matériel visé par ce système supportait déjà matériellement cette isolation, avec les deux modes d'exécution utilisateur et protégé. Contrairement à MS-DOS, l'interface entre les applications et le *noyau* était clairement définie. L'interface entre le *noyau* et le matériel s'appuie sur un ensemble de drivers de périphériques.

L'organisation du noyau des UNIX originel était ce qu'on appelle une architecture *monolithique*. L'ensemble des fonctionnalités du système était assuré par un module logiciel unique, mettant en œuvre l'ensemble des appels système, de la même façon que pour le système MS-DOS. Très rapidement, cette structure à un seul niveau s'est révélée complexe à maintenir et à faire évoluer, en particulier lorsque UNIX devait être adapté pour fonctionner sur de nouveaux modèles de mini-ordinateurs et de nouveaux processeurs. Il est donc apparu rapidement nécessaire de rendre l'organisation du système d'exploitation plus *modulaire*, c'est à dire de permettre la mise à jour ou l'évolution de différents services de manière séparée. Une modification du code de l'un de ces services ne doit, en principe, pas entraîner de changements majeurs dans les autres parties du système d'exploitation.

Structure en couches (UNIX)

Une première approche est d'organiser le système en couches : les services mis en œuvre par une couche dépendent alors uniquement des services fournis par les couches inférieures. La couche la plus basse est celle qui héberge les drivers de périphérique, et la couche la plus haute est celle qui met en œuvre la réponse aux appels systèmes. Les couches intermédiaires proposent aux couches supérieures des niveaux d'abstraction des ressources de plus en plus élevés, jusqu'à arriver au niveau d'abstraction fourni à l'espace utilisateur. Considérons un exemple simplifié d'un service de gestion de périphérique de stockage sur disque dur :

– La couche la plus basse (niveau 0) contient le driver de périphérique, qui est capable de transformer des requêtes pour des blocs de données en des commandes bas niveau pour actionner le bras de lecture du disque, lire une piste magnétique spécifique, etc.
– La couche suivante (niveau 1) construit une abstraction de volumes de données, correspondant aux disques virtuels (volumes), mais n'ayant pas connaissance de la notion de fichiers ou de répertoires.
– Enfin, la dernière couche (niveau 2) met en œuvre l'abstraction d'un système de fichier à proprement parler, en établissant une correspondance entre les blocs de données et les notions haut-niveau que sont les fichiers et les répertoires.

Une architecture en couche présente des avantages. Il est plus facile d'isoler les différentes fonctionnalités et de porter le système d'exploitation d'un environnement à un autre. Par exemple, l'utilisation d'un disque de type SSD ne demandera des changements qu'au niveau 0, et l'utilisation d'un disque distant (accédé par l'intermédiaire d'un réseau) ne demandera des changements qu'au niveau 1. Dans les deux cas, il ne sera pas nécessaire de modifier le code au niveau 2. La recherche de bugs sera aussi facilitée : on peut tester les fonctionnalités de la couche N avant de mettre en œuvre les fonctionnalités de la couche N+1.

Toutefois, cette architecture en couche présente aussi deux inconvénients. Le premier est que le service des appels systèmes doivent désormais utiliser une succession d'appels entre les couches. Chaque couche va devoir traiter un appel, mettre à jour des structures de données, et préparer un ou plusieurs appels pour les couches inférieures, ce qui peut introduire un surcoût à l'exécution par rapport à une approche monolithique. Cet inconvénient est relativement limité sur un système moderne où l'exécution du code n'est pas le facteur limitant, mais plutôt l'accès à la mémoire. La deuxième inconvénient est qu'il n'est pas aisé de structurer clairement un *noyau* de système d'exploitation de cette façon, car les services systèmes sont souvent interdépendants. Nous verrons par exemple que la gestion de la mémoire, la gestion des entrées/sorties, ou encore la gestion des processus, dépendent chacun les uns des autres pour assurer leurs fonctionnalités ou pour mettre en œuvre des optimisations. Pour cette raison, les systèmes modernes comme Linux utilisent peu de couches mais préfèrent une organisation sous forme de modules, comme nous allons le voir à présent.

Structure en modules (Linux)

La structuration en modules combine un cœur du système d'exploitation contenant les services fondamentaux du système (gestion des processus, gestion de la mémoire virtuelle) avec un certain nombre de modules mettant en œuvre les autres fonctionnalités. Cette stratégie est désormais la plus communément utilisée, par exemple par Linux, Solaris, ou les versions récentes de Windows.

Les modules peuvent être chargés dynamiquement dans l'espace mémoire du noyau, en fonction des besoins du système informatique considéré, ou lors du démarrage du système. Comme premier exemple, un module permettant l'utilisation d'une interface de périphérique sans fil Bluetooth ne sera chargée que sur un système disposant d'un contrôleur de périphérique de cette technologie. Un second exemple est le support d'un système de fichier particulier. Différents systèmes d'exploitation utilisent généralement des systèmes de fichiers différents (i.e., la manière de représenter les informations des fichiers et des répertoires sur le disque n'est pas la même). Si Linux est installé en *dual-boot* sur un ordinateur contenant aussi une copie de Windows, il sera possible d'accéder au contenu du disque Windows à partir de Linux en chargeant dans le noyau un module spécifique nommé `exFAT`.

La structuration en modules présente des avantages similaires à celle de la structuration en couches. Il est plus facile de déboguer un module dont l'interface est bien définie, que lorsque les fonctionnalités sont noyées dans un grand monolithe. La séparation en modules facilite l'évolution du système d'exploitation dans le temps et sa portabilité sur des systèmes très différents. Celle-ci explique en partie pourquoi le noyau Linux est utilisé sur des ordinateurs aussi variés qu'un smartphone Android, une télévision connectée, un ordinateur personnel, ou un super-calculateur regroupant des centaines de milliers de processeurs. Enfin, l'utilisation de modules résout le problème de l'interdépendance entre couches : les modules peuvent appeler les fonctionnalités des uns des autres sans remettre en question la séparation du code et des données correspondant aux différentes fonctionnalités.

Note : Utilisation des modules sous Linux

Sous Linux, des utilitaires systèmes permettent de charger et décharger des modules dans le noyau. Puisque ces modules vont devenir partie du code du noyau, ces opérations sont réservées aux utilisateurs avec un niveau de privilège élevé dans le système, typiquement les administrateurs. Ceux-ci peuvent par ailleurs mettre en place le chargement automatique de modules. Par exemple, le module *exFAT* pourrait n'être chargé automatiquement que lorsqu'une clé USB à ce format, en provenance d'un ordinateur Windows, est inséré dans un des ports USB de la machine.

La commande `sudo lsmod` permet de lister les modules présents. On voit un court extrait d'une sortie de cette commande ci-dessous. Le module `psmouse` permet la gestion des entrées/sorties avec une souris ou un trackpad. Les modules `soundcore` et `snd` sont dédiés à la gestion des entrées/sorties son. On peut voir qu'ils peuvent avoir des dépendances : le chargement du module `snd` requière ainsi la présence des modules `snd_intel8x0`, `snd_ac97_codec`, `snd_pcm`, et `snd_timer`.

```
$ sudo lsmod
Module                  Size  Used by
(...)
psmouse                97578  0
serio_raw              13230  0
snd                    61351  4 snd_intel8x0,snd_ac97_codec,snd_pcm,snd_timer
soundcore              12600  1 snd
nfsd                  255559  2
(...)
```

Les commandes `modprobe` et `modinfo` permettent d'installer/désinstaller des modules et d'obtenir de l'information sur un module, respectivement. Par exemple, la sortie suivante est un extrait du résultat de `sudo modinfo psmouse`.

```
$ sudo modinfo psmouse
filename:       /lib/modules/3.13.0-32-generic/kernel/drivers/input/mouse/psmouse.ko
license:        GPL
description:    PS/2 mouse driver
author:         Vojtech Pavlik <vojtech@suse.cz>
(...)
depends:
(...)
```

Structure en micro-noyau (L4)

Les structures monolithiques, en couche et utilisant des modules présentées précédemment ont toutes un défaut en commun : la quantité de code exécutée en mode protégé au sein du noyau est très importante. Ceci pose un

problème de fiabilité : une fonctionnalité mal mise en œuvre dans le noyau (par exemple, qui accède à des adresses mémoires incohérentes en déréférencant un pointeur mal initialisé, ou qui utilisent une instruction de contrôle du matériel malformée) peuvent affecter l'ensemble du noyau et donc l'ensemble du système. Cela peut résulter en un crash complet de la machine voire, ce qui est encore moins souhaitable, en des corruptions des données ou en des fautes exploitables par des logiciels malicieux pour effectuer des opérations non autorisées (comme, par exemple, casser la propriété d'isolation).

Le concept de micro-noyau est une réponse à ce problème. Il consiste à réduire la taille du code du noyau (et donc les fonctionnalités supportées) au strict nécessaire, et à mettre en œuvre le reste des fonctionnalités sous forme de programmes fonctionnant en espace utilisateur.

Les fonctionnalités fondamentales mises en œuvre dans le micro-noyau sont généralement une gestion basique de la mémoire, la gestion des processus légers (ou threads, que nous verrons en détail dans la prochaine partie du cours), et la communication entre processus. Les autres fonctionnalités, y compris les drivers de périphériques, fonctionnent sous forme de processus en mode utilisateur. Ces processus jouent un rôle similaire aux modules décrits précédemment. Toutefois, puisqu'ils ne sont plus dans l'espace mémoire du noyau, ils ne peuvent plus appeler les fonctionnalités des autres services directement, en utilisant des appels de fonctions standard. Ils doivent à la place utiliser des communications inter-processus, en appelant pour cela un appel système spécifique. Le micro-noyau se charge alors d'acheminer entre les deux processus les messages, sans que ceux-ci n'aient de visibilité mémoire commune, ce qui conserve la propriété d'isolation.

Les micro-noyaux ont un avantage majeur : le code fonctionnant en mode protégé est réduit au minimum et on peut alors se concentrer sur sa qualité. Les contributions logicielles externes, comme les drivers de périphériques, peuvent contenir des erreurs ou essayer d'utiliser des instructions interdites. Cela ne mettra toutefois pas en cause l'intégrité du système : comme pour un processus utilisateur qui effectuerait une opération interdite, le processus contenant le driver fautif sera simplement terminé (et éventuellement relancé) mais le reste du système ne sera pas affecté. Le même raisonnement s'applique pour les fonctionnalités complexes, comme les systèmes de fichiers, donc la mise en œuvre peut atteindre plusieurs dizaines de milliers de lignes de code C. On comprend l'importance qu'a cette isolation lorsque l'on considère, comme le montre l'étude de Chou *et al.* en 2001 [Chou2001] ou celle de Palix *et al.* en 2011 [Palix2011] que les branches *drivers* et *fs* du noyau Linux contiennent souvent jusqu'à 7 fois plus d'erreur par millier de lignes de code que les autres branches.

La principale raison pour laquelle le concept de micro-noyau n'est pas aussi répandu est que sa mise en œuvre efficace est particulièrement délicate. En particulier, le mécanisme de passage de message *via* le noyau, qui remplace l'appel direct de fonctions entre modules, est plus coûteux que ce dernier. À la place de placer des arguments sur la pile et de rediriger le compteur de programme vers une autre adresse du noyau, comme c'est le cas dans un noyau monolithique, avec un micro-noyau il est nécessaire de redonner le contrôle au système d'exploitation, qui doit copier le message à transmettre de l'espace mémoire d'un processus à un autre, et mettre en place plusieurs changements de contexte. Cette opération, répétée de très nombreuses fois, peut grever la performance si elle n'est pas parfaitement optimisée. On peut illustrer ce phénomène avec le système d'exploitation historique Windows NT, introduit dans les années 1990. Ce système d'exploitation était le premier système Windows qui ne dépendait pas du tout de MS-DOS. Dans ses premières versions, les concepteurs de Microsoft avaient décidé d'adopter une approche micro-noyau mais ont progressivement décidé de ramener des fonctionnalités externalisées dans ce dernier, constatant la perte importante de performance. Lorsque Windows NT a finalement évolué vers le système Windows XP, ce dernier était devenu *de facto* un système à noyau monolithique. Ce n'est que quelques années plus tard, avec les premières versions de Mac OS X, et surtout avec l'amélioration des procédures d'échange de message, qu'une approche micro-noyau a pu être déployée avec succès dans un produit commercial.

De nos jours, on retrouve des systèmes d'exploitation à micro-noyaux dans les systèmes embarqués critiques avec les systèmes L4 et QNX par exemple. Mac OS ainsi qu'iOS d'Apple sont des systèmes hybrides, combinant des fonctionnalités typiques d'un micro-noyau mais incluant des fonctionnalités qui pourraient en principe être externalisées en espace utilisateur, pour des raisons de performance.

Note : Micro-noyau et logiciel formellement certifié

Un système d'exploitation est un élément critique en ce qui concerne la sécurité et la sûreté de fonctionnement d'un système informatique. Si l'on peut parfois accepter qu'un ordinateur personnel "plante" lors de l'essai d'une version non stabilisée d'un système d'exploitation, il n'en est pas de même pour un système critique dans le domaine spatial ou le transport. De la même façon, un système d'exploitation peut être utilisé dans un domaine ou la protection des données est primordiale, comme par exemple sur un serveur qui hébergerait des données

médicales. Il ne serait pas acceptable qu'un logiciel exécuté sur la même machine puisse accéder à ces données en forçant l'accès à l'espace mémoire d'un autre processus.

Un système comme Linux contient pourtant des millions de lignes de code (à titre d'exemple, le dépôt *git* de Linux contient plus de 28 millions de lignes, principalement de C, comprenant toutefois très majoritairement des drivers de périphériques). Bien que des milliers de développeurs très talentueux travaillent constamment à découvrir des erreurs dans ce code, il est très difficile de garantir qu'un logiciel de cette taille en est complètement exempt. Certaines études [Chou2001] [Palix2011] montrent ainsi que certains bugs ne sont corrigés que plusieurs années après leur première identification !

L'utilisation d'un micro-noyau peut réduire drastiquement la quantité de lignes de code à analyser et à débogguer, mais cela n'est pas toujours suffisant. Récemment, des concepteurs de systèmes d'exploitation spécialisés pour les applications critiques ont entrepris de certifier de façon formelle la qualité de leurs systèmes. Ce processus nécessite de spécifier les fonctionnalités du système d'exploitation, comme par exemple la totale isolation entre les espaces mémoires accessibles au différents processus, à l'aide d'un formalisme mathématique. Des logiciels spécialisés permettent ensuite de valider une mise en œuvre (en C) du système d'exploitation par rapport à cette spécification formelle de haut niveau. Cette opération est très complexe et coûteuse en ressources de calcul. Elle ne peut donc s'appliquer qu'à un logiciel de taille raisonnable, comme un micro-noyau. Le projet le plus avancé dans ce domaine est sans doute le système d'exploitation seL4 développé par l'université de Sidney en Australie. Si seL4 ne comporte qu'une dizaine de milliers de lignes de C et moins d'un millier de lignes d'assembleur, la preuve mathématique de sa correction représente des millions de ligne de clauses mathématiques et un travail d'une ampleur considérable. Il faudra sans doute quelques années avant que les mêmes pratiques se généralisent aux systèmes d'exploitation grand public.

5.2.3 Démarrage du système d'exploitation

Nous terminons cette présentation de la structure des systèmes d'exploitation par une description du processus permettant le démarrage d'un système. Lors de ce démarrage, plusieurs étapes sont nécessaires pour permettre de donner la main au *noyau* du système d'exploitation.

Lors du démarrage de la machine, la mémoire principale se trouve dans un état indéterminé. Un programme de démarrage (*bootstrap* en anglais) doit être exécuté pour charger le *noyau* depuis le disque et démarrer celui-ci. Ce programme de démarrage est généralement stocké dans une mémoire non volatile (souvent dénotée ROM, pour *Read-Only Memory*). Cette mémoire ROM utilise une technologie différente de la mémoire principale, et son contenu n'est pas perdu lors de la mise hors tension de la machine. En pratique, le type de mémoire utilisé n'est pas seulement en lecture seul (Read-Only) mais supporte des mises à jour occasionnelles nécessitant un programme spécial (on parle alors d'un *firmware*, et d'une mise à jour de *firmware*)

Le processeur reçoit lors du démarrage (ou du redémarrage) de la machine une interruption dite de remise à zéro. Il charge alors son compteur de programme à la première adresse de la mémoire ROM. Cette adresse contient la première instruction du programme de démarrage. Ce dernier va en général effectuer tout d'abord un certain nombre de vérifications de la machine (comme par exemple l'absence d'erreur au niveau de la mémoire principale), initialiser les registres matériels, les bus de communication, et les gestionnaires de périphériques.

Ensuite, ce programme va devoir récupérer sur le disque le code du noyau à proprement parler, pour le copier en mémoire principale et enfin brancher vers sa première instruction. Sur la plupart des systèmes, cette étape se déroule en deux temps : le programme de démarrage est seulement capable de lire le tout premier bloc d'un support de stockage (en général un disque dur ou SSD) dans lequel un programme de chargement plus complet est stocké. C'est ce dernier qui va charger le code du noyau depuis son emplacement effectif sur le disque (le noyau n'est pas stocké dans le premier bloc, mais dans le système de fichier ; sous Linux ce fichier est généralement stocké dans le répertoire `/boot`, par exemple `/boot/vmlinuz-3.13.0-32-generic`). Sous Linux, le gestionnaire de démarrage GRUB joue ce rôle. Il permet par ailleurs de gérer le démarrage de plusieurs systèmes (comme Solaris, Windows, etc.) ou bien de permettre le démarrage de différents noyaux pour un même système, ce qui est parfois utile pour les développeurs. On notera que lors de l'exécution de GRUB, avant l'exécution du noyau Linux lui-même, les modules de Linux permettant d'utiliser le système de fichier ne sont pas chargés. GRUB inclue donc ses propres modules pour pouvoir utiliser les systèmes de fichiers les plus courant et y localiser le fichier contenant le code du noyau.

CHAPITRE 6

Systèmes Multiprocesseurs

6.1 Utilisation de plusieurs threads

Les performances des microprocesseurs se sont continuellement améliorées depuis les années 1960s. Cette amélioration a été possible grâce aux progrès constants de la micro-électronique qui a permis d'assembler des microprocesseurs contenant de plus en plus de transistors sur une surface de plus en plus réduite. La figure [1] ci-dessous illustre bien cette évolution puisqu'elle représente le nombre de transistors par microprocesseur en fonction du temps.

Cette évolution avait été prédite par Gordon Moore dans les années 1960s [Stokes2008]. Il a formulé en 1965 une hypothèse qui prédisait que le nombre de composants par puce continuerait à doubler tous les douze mois au cours de la prochaine décennie. Cette prédiction s'est avérée tout à fait réaliste. Elle est maintenant connue sous le nom de *Loi de Moore* et est fréquemment utilisée pour expliquer les améliorations de performance des ordinateurs.

Le fonctionnement d'un microprocesseur est régulé par une horloge. Celle-ci rythme la plupart des opérations du processeur et notamment le chargement des instructions depuis la mémoire. Pendant de nombreuses années, les performances des microprocesseurs ont fortement dépendu de leur vitesse d'horloge. Les premiers microprocesseurs avaient des fréquences d'horloge de quelques centaines de *kHz*. A titre d'exemple, le processeur intel 4004 avait une horloge à 740 kHz en 1971. Aujourd'hui, les processeurs rapides dépassent la fréquence de 3 *GHz*. La figure ci-dessous présente l'évolution de la fréquence d'horloge des microprocesseurs depuis les années 1970s [2]. On remarque une évolution rapide jusqu'aux environs du milieu de la dernière décennie. La barrière des 10 MHz a été franchie à la fin des années 1970s. Les 100 *MHz* ont étés atteints en 1994 et le GHz aux environs de l'an 2000.

Pendant près de quarante ans, l'évolution technologique a permis une amélioration continue des performances des microprocesseurs. Cette amélioration a directement profité aux applications informatiques car elles ont pu s'exécuter plus rapidement au fur et à mesure que la vitesse d'horloge des microprocesseurs augmentait.

Malheureusement, vers 2005 cette croissance continue s'est arrêtée. La barrière des 3 GHz s'est avérée être une barrière très difficile à franchir d'un point de vue technologique. Aujourd'hui, les fabricants de microprocesseurs n'envisagent plus de chercher à continuer à augmenter les fréquences d'horloge des microprocesseurs.

Si pendant longtemps la fréquence d'horloge d'un microprocesseur a été une bonne heuristique pour prédire les performances du microprocesseur, ce n'est pas un indicateur parfait de performance. Certains processeurs exécutent une instruction durant chaque cycle d'horloge. D'autres processeurs prennent quelques cycles d'horloge pour exécuter chaque instruction et enfin certains processeurs sont capables d'exécuter plusieurs instructions durant chaque cycle d'horloge.

Une autre façon de mesurer les performances d'un microprocesseur est de comptabiliser le nombre d'instructions qu'il exécute par seconde. On parle en général de Millions d'Instructions par Seconde (ou *MIPS*). Si les premiers

1. Source : http ://en.wikipedia.org/wiki/File :Transistor_Count_and_Moore%27s_Law_-_2011.svg
2. Plusieurs sites web recensent cette information, notamment http ://www.intel.com/pressroom/kits/quickreffam.htm, http ://en.wikipedia.org/wiki/List_of_Intel_microprocessors et http ://en.wikipedia.org/wiki/Instructions_per_second

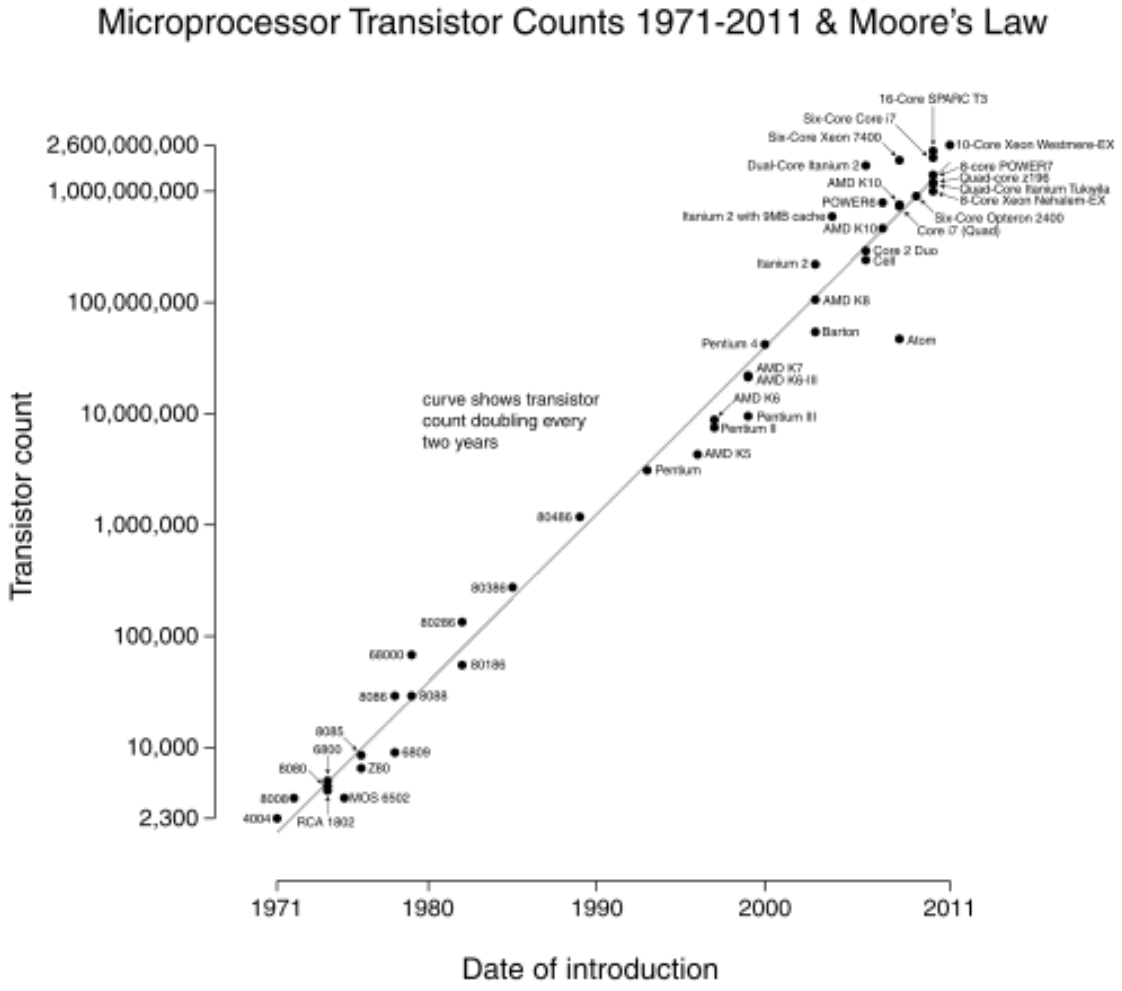


FIGURE 6.1 – Evolution du nombre de transistors par microprocesseur

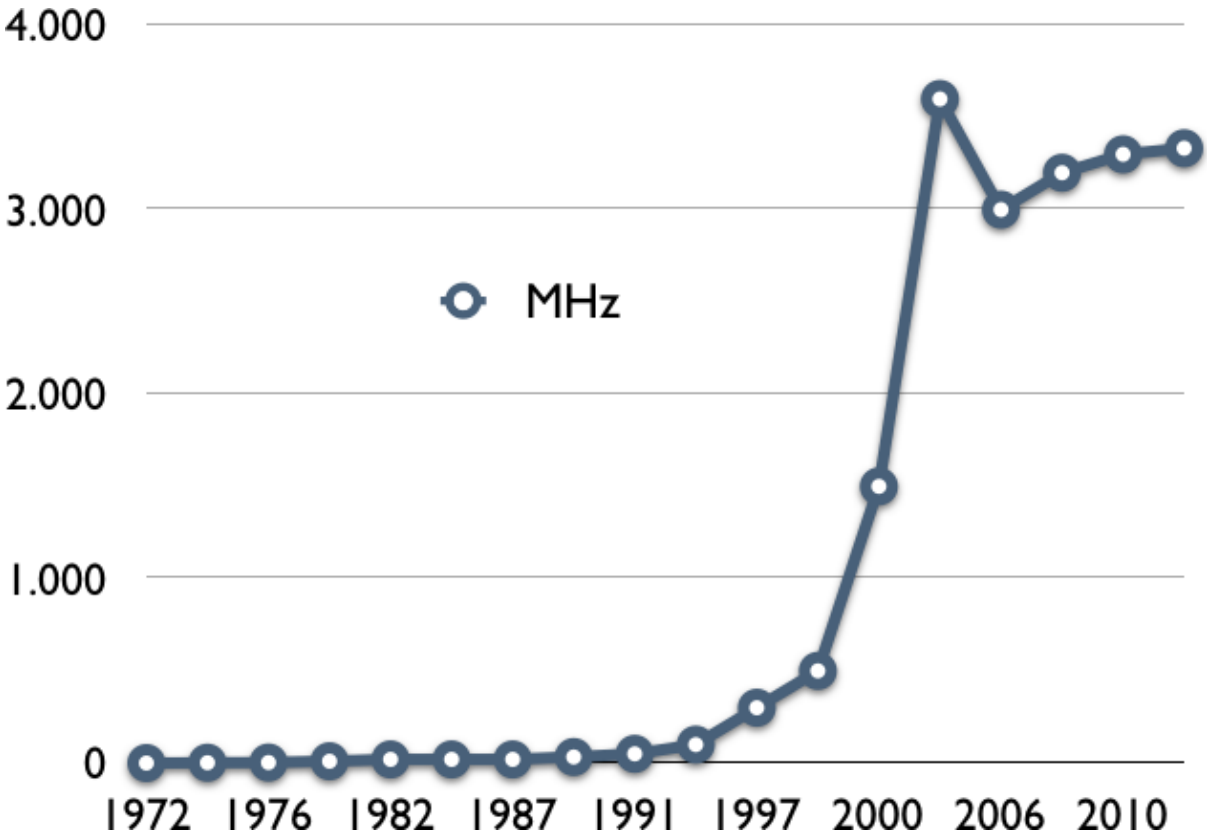


FIGURE 6.2 – Evolution de la vitesse d'horloge des microprocesseurs

microprocesseurs effectuaient moins de 100.000 instructions par seconde, la barrière du MIPS a été franchie en 1979. Mesurées en MIPS, les performances des microprocesseurs ont continué à augmenter durant les dernières années malgré la barrière des 3 GHz comme le montre la figure ci-dessous.

Note : Evaluation des performances de systèmes informatiques

La fréquence d'horloge d'un processeur et le nombre d'instructions qu'il est capable d'exécuter chaque seconde ne sont que quelques uns des paramètres qui influencent les performances d'un système informatique qui intègre ce processeur. Les performances globales d'un système informatique dépendent de nombreux autres facteurs comme la capacité de mémoire et ses performances, la vitesse des bus entre les différents composants, les performances des dispositifs de stockage ou des cartes réseaux. Les performances d'un système dépendront aussi fortement du type d'application utilisé. Un serveur web, un serveur de calcul scientifique et un serveur de bases de données n'auront pas les mêmes contraintes en termes de performance. L'évaluation complète des performances d'un système informatique se fait généralement en utilisant des benchmarks. Un *benchmark* est un ensemble de logiciels qui reproduisent le comportement de certaines classes d'applications de façon à pouvoir tester les performances de systèmes informatiques de façon reproductible. Différents organismes publient de tels benchmarks. Le plus connu est probablement Standard Performance Evaluation Corporation qui publie des benchmarks et des résultats de benchmarks pour différents types de systèmes informatiques et d'applications.

Cette progression continue des performances en MIPS a été possible grâce à l'introduction de processeurs qui sont capables d'exécuter plusieurs threads d'exécution simultanément. On parle alors de processeur *multi-coeurs* ou *multi-threadé*.

La notion de thread d'exécution est très importante dans un système informatique. Elle permet non seulement de comprendre comme un ordinateur équipé d'un seul microprocesseur peut exécuter plusieurs programmes simultanément, mais aussi comment des programmes peuvent profiter des nouveaux processeurs capables d'exécuter plusieurs threads simultanément. Pour comprendre cette notion, il est intéressant de revenir à nouveau sur l'exécution d'une fonction en langage assembleur. Considérons la fonction `f` :

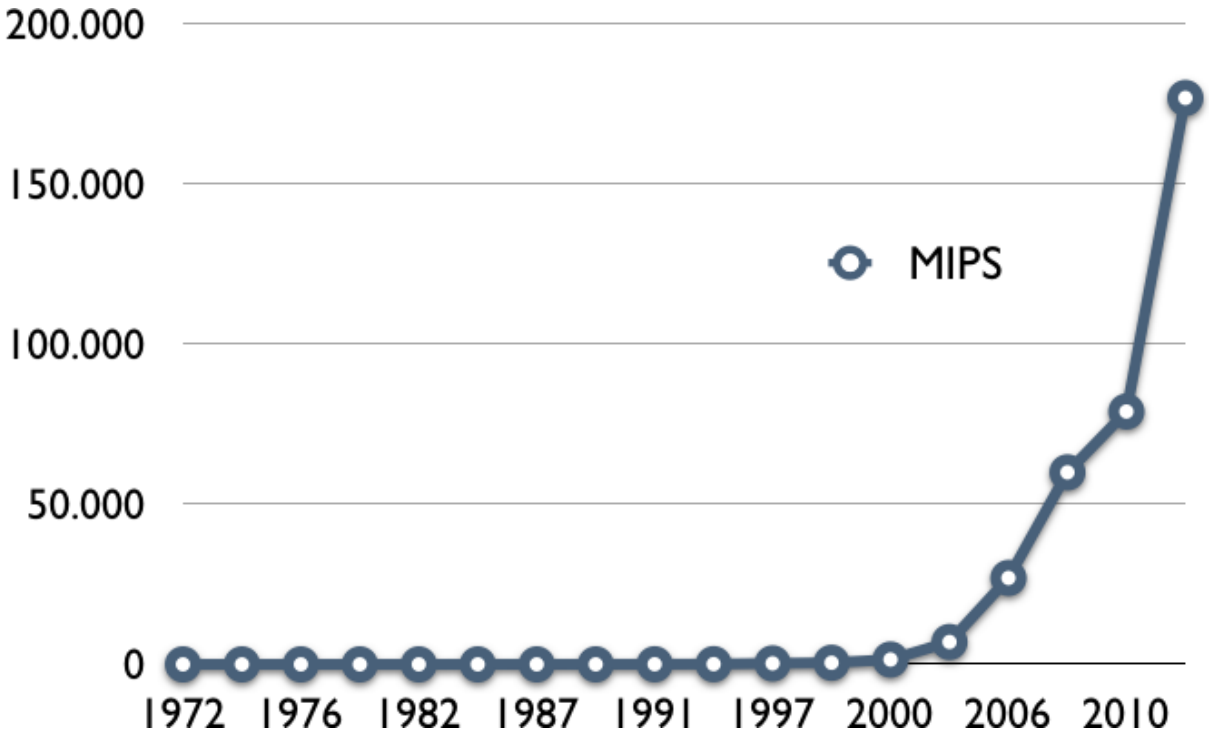


FIGURE 6.3 – Evolution des performances des microprocesseurs en MIPS

```
int f(int a, int b ) {
  int m=0;
  int c=0;
  while(c<b) {
     m+=a;
     c=c+1;
  }
  return m;
}
```

En assembleur, cette fonction se traduit en :

```
f:
    subl    $16, %esp
    movl    24(%esp), %eax
    movl    20(%esp), %ecx
    movl    %ecx, 12(%esp)
    movl    %eax, 8(%esp)
    movl    $0, 4(%esp)
    movl    $0, (%esp)
    .LBB0_1:
    movl    (%esp), %eax
    cmpl    8(%esp), %eax
    jge     .LBB0_3

    movl    12(%esp), %eax
    movl    4(%esp), %ecx
    addl    %eax, %ecx
    movl    %ecx, 4(%esp)
    movl    (%esp), %eax
    addl    $1, %eax
    movl    %eax, (%esp)
    jmp     .LBB0_1
    .LBB0_3:
```

```
movl    4(%esp), %eax
addl    $16, %esp
ret
```

Pour qu'un processeur puisse exécuter cette séquence d'instructions, il faut non seulement qu'il implémente chacune de ces instructions, mais également qu'il puisse accéder :
– à la mémoire contenant les instructions à exécuter
– à la mémoire contenant les données manipulées par cette séquence d'instruction. Pour rappel, cette mémoire est divisée en plusieurs parties :
  – la zone contenant les variables globales
  – le tas
  – la pile
– aux registres et plus particulièrement, il doit accéder :
  – aux registres de données pour stocker les résultats de chacune des instructions
  – au registre `%esp` directement ou indirectement via les instructions `push` et `pop` qui permettent de manipuler la pile
  – au registre `%eip` qui contient l'adresse de l'instruction en cours d'exécution
  – au registre `eflags` qui contient l'ensemble des drapeaux

Un processeur *multithreadé* a la capacité d'exécuter plusieurs programmes simultanément. En pratique, ce processeur disposera de plusieurs copies des registres. Chacun de ces blocs de registres pourra être utilisé pour exécuter ces programmes simultanément à raison d'un thread d'exécution par bloc de registres. Chaque thread d'exécution va correspondre à une séquence différente d'instructions qui va modifier son propre bloc de registres. C'est grâce à cette capacité d'exécuter plusieurs threads d'exécution simultanément que les performances en *MIPS* des microprocesseurs ont pu continuer à croître alors que leur fréquence d'horloge stagnait.

Cette capacité d'exécuter plusieurs threads d'exécution simultanément n'est pas limitée à un thread d'exécution par programme. Sachant qu'un thread d'exécution n'est finalement qu'une séquence d'instructions qui utilisent un bloc de registres, il est tout à fait possible à plusieurs séquences d'exécution appartenant à un même programme de s'exécuter simultanément. Si on revient à la fonction assembleur ci-dessus, il est tout à fait possible que deux invocations de cette fonction s'exécutent simultanément sur un microprocesseur. Pour démarrer une telle instance, il suffit de pouvoir initialiser le bloc de registres nécessaire à la nouvelle instance et ensuite de démarrer l'exécution à la première instruction de la fonction. En pratique, cela nécessite la coopération du système d'exploitation. Différents mécanismes ont été proposés pour permettre à un programme de lancer différents threads d'exécution. Aujourd'hui, le plus courant est connu sous le nom de threads POSIX. C'est celui que nous allons étudier en détail, mais il en existe d'autres.

Note : D'autres types de threads

À côté des threads POSIX, il existe d'autres types de threads. [Gove2011] présente comment mettre en œuvre des threads sur différents systèmes d'exploitation. Sous Linux, NTPL [DrepperMolnar2005] et LinuxThreads [Leroy] sont deux anciennes implémentations des threads POSIX. GNU PTH [GNUPTH] est une librairie qui implémente les threads sans interaction directe avec le système d'exploitation. Cela permet à la librairie d'être portable sur de nombreux systèmes d'exploitation. Malheureusement, tous les threads GNU PTH d'un programme doivent s'exécuter sur le même processeur.

6.1.1 Les threads POSIX

Les threads POSIX sont supportés par la plupart des variantes de Unix. Ils sont souvent implémentés à l'intérieur d'une librairie. Sous Linux, il s'agit de la librairie pthreads(7) qui doit être explicitement compilée avec le paramètre `-lpthread` lorsque l'on utilise gcc(1).

La librairie threads POSIX contient de nombreuses fonctions qui permettent de décomposer un programme en plusieurs threads d'exécution et de les gérer. Toutes ces fonctions nécessitent l'inclusion du fichier pthread.h. La première fonction importante est pthread_create(3) qui permet de créer un nouveau thread d'exécution. Cette fonction prend quatre arguments et retourne une valeur entière.

```
#include <pthread.h>
```

```
int
pthread_create(pthread_t *restrict thread,
               const pthread_attr_t *restrict attr,
               void *(*start_routine)(void *),
               void *restrict arg);
```

Le premier argument est un pointeur vers une structure de type `pthread_t`. Cette structure est définie dans pthread.h et contient toutes les informations nécessaires à l'exécution d'un thread. Chaque thread doit disposer de sa structure de données de type `pthread_t` qui lui est propre.

Le second argument permet de spécifier des attributs spécifiques au thread qui est créé. Ces attributs permettent de configurer différents paramètres associés à un thread. Nous y reviendrons ultérieurement. Si cet argument est mis à `NULL`, la librairie `pthreads` utilisera les attributs par défaut qui sont en général largement suffisants.

Le troisième argument contient l'adresse de la fonction par laquelle le nouveau thread va démarrer son exécution. Cette adresse est le point de départ de l'exécution du thread et peut être comparée à la fonction `main` qui est lancée par le système d'exploitation lorsqu'un programme est exécuté. Un thread doit toujours débuter son exécution par une fonction dont la signature est `void * function(void *)`, c'est-à-dire une fonction qui prend comme argument un pointeur générique (de type `void *`) et retourne un résultat du même type.

Le quatrième argument est l'argument qui est passé à la fonction qui débute le thread qui vient d'être créé. Cet argument est un pointeur générique de type `void *`, mais la fonction peut bien entendu le convertir dans un autre type.

La fonction pthread_create(3) retourne un résultat entier. Une valeur de retour non-nulle indique une erreur et `errno` est mise à jour.

Un thread s'exécute en général pendant une certaine période de temps puis il peut retourner un résultat au thread d'exécution principal. Un thread peut retourner son résultat (de type `void *`) de deux façons au thread qui l'a lancé. Tout d'abord, un thread qui a démarré par la fonction `f` se termine lorsque cette fonction exécute `return(...)`. L'autre façon de terminer un thread d'exécution est d'appeler explicitement la fonction pthread_exit(3). Celle-ci prend un argument de type `void *` et le retourne au thread qui l'avait lancé.

Pour récupérer le résultat d'un thread d'exécution, le thread principal doit utiliser la fonction pthread_join(3). Celle-ci prend deux arguments et retourne un entier.

```
#include <pthread.h>

int
pthread_join(pthread_t thread, void **value_ptr);
```

Le premier argument de pthread_join(3) est la structure `pthread_t` correspondant au thread dont le résultat est attendu. Le second argument est un pointeur vers un pointeur générique (`void **`) qui après la terminaison du thread passé comme premier argument pointera vers la valeur de retour de ce thread.

L'exemple ci-dessous illustre une utilisation simple des fonctions pthread_create(3), pthread_join(3) et pthread_exit(3).

```
#include <pthread.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <errno.h>

int global=0;

void error(int err, char *msg) {
  fprintf(stderr,"%s a retourné %d, message d'erreur : %s\n",msg,err,strerror(errno));
  exit(EXIT_FAILURE);
}

void *thread_first(void * param) {
  global++;
  return(NULL);
```

```
}

void *thread_second(void * param) {
  global++;
  pthread_exit(NULL);
}

int main (int argc, char *argv[])  {
  pthread_t first;
  pthread_t second;
  int err;

  err=pthread_create(&first,NULL,&thread_first,NULL);
  if(err!=0)
    error(err,"pthread_create");

  err=pthread_create(&second,NULL,&thread_second,NULL);
  if(err!=0)
    error(err,"pthread_create");

  for(int i=0; i<1000000000;i++) { /*...*/ }

  err=pthread_join(second,NULL);
    if(err!=0)
      error(err,"pthread_join");

  err=pthread_join(first,NULL);
    if(err!=0)
      error(err,"pthread_join");

  printf("global: %d\n",global);

  return(EXIT_SUCCESS);
}
```

Dans ce programme, la fonction `main` lance deux threads. Le premier exécute la fonction `thread_first` et le second la fonction `thread_second`. Ces deux fonctions incrémentent une variable globale et n'utilisent pas leur argument. `thread_first` se termine par `return` tandis que `thread_second` se termine par un appel à pthread_exit(3). Après avoir créé ses deux threads, la fonction `main` démarre une longue boucle puis appelle `pthread_join` pour attendre la fin des deux threads qu'elle avait lancé.

Afin d'illustrer la possibilité de passer des arguments à un thread et d'en récupérer la valeur de retour, considérons l'exemple ci-dessous.

```
#define NTHREADS 4
void *neg (void * param) {
  int *l;
  l=(int *) param;
  int *r=(int *)malloc(sizeof(int));
  *r=-*l;
  return ((void *) r);
}

int main (int argc, char *argv[])  {
  pthread_t threads[NTHREADS];
  int arg[NTHREADS];
  int err;

  for(long i=0;i<NTHREADS;i++) {
    arg[i]=i;
    err=pthread_create(&(threads[i]),NULL,&neg,(void *) &(arg[i]));
    if(err!=0)
      error(err,"pthread_create");
```

```
  }

  for(int i=0;i<NTHREADS;i++) {
    int *r;
    err=pthread_join(threads[i],(void **)&r);
    printf("Resultat[%d]=%d\n",i,*r);
    free(r);
    if(err!=0)
      error(err,"pthread_join");
  }
  return(EXIT_SUCCESS);
}
```

Ce programme lance 4 threads d'exécution en plus du thread principal. Chaque thread d'exécution exécute la fonction `neg` qui récupère un entier comme argument et retourne l'opposé de cet entier comme résultat.

Lors d'un appel à pthread_create(3), il est important de se rappeler que cette fonction crée le thread d'exécution, mais que ce thread ne s'exécute pas nécessairement immédiatement. En effet, il est très possible que le système d'exploitation ne puisse pas activer directement le nouveau thread d'exécution, par exemple parce que l'ensemble des processeurs de la machine sont actuellement utilisés. Dans ce cas, le thread d'exécution est mis en veille par le système d'exploitation et il sera démarré plus tard. Sachant que le thread peut devoir démarrer plus tard, il est important de s'assurer que la fonction lancée par pthread_create(3) aura bien accès à son argument au moment où finalement elle démarrera. Dans l'exemple ci-dessous, cela se fait en passant comme quatrième argument l'adresse d'un entier casté en `void *`. Cette valeur est copiée sur la pile de la fonction `neg`. Celle-ci pourra accéder à cet entier via ce pointeur sans problème lorsqu'elle démarrera.

Note : Un thread doit pouvoir accéder à son argument

Lorsque l'on démarre un thread via la fonction pthread_create(3), il faut s'assurer que la fonction lancée pourra bien accéder à ses arguments. Ce n'est pas toujours le cas comme le montre l'exemple ci-dessous. Dans cet exemple, c'est l'adresse de la variable locale `i` qui est passée comme quatrième argument à la fonction pthread_create(3). Cette adresse sera copiée sur la pile de la fonction `neg` pour chacun des threads créés. Malheureusement, lorsque la fonction `neg` sera exécutée, elle trouvera sur sa pile l'adresse d'une variable qui risque fort d'avoir été modifiée après l'appel à pthread_create(3) ou pire risque d'avoir disparu car la boucle `for` s'est terminée. Il est très important de bien veiller à ce que le quatrième argument passé à pthread_create(3) existe toujours au moment de l'exécution effective de la fonction qui démarre le thread lancé.

```
/// erroné !
for(long i=0;i<NTHREADS;i++) {
  err=pthread_create(&(threads[i]),NULL,&neg,(void *)&i);
  if(err!=0)
    error(err,"pthread_create");
}
```

Concernant pthread_join(3), le code ci-dessus illustre la récupération du résultat via un pointeur vers un entier. Comme la fonction `neg` retourne un résultat de type `void *` elle doit nécessairement retourner un pointeur qui peut être casté vers un pointeur de type `void *`. C'est ce que la fonction `neg` dans l'exemple réalise. Elle alloue une zone mémoire permettant de stocker un entier et place dans cette zone mémoire la valeur de retour de la fonction. Ce pointeur est ensuite casté en un pointeur de type `void *` avant d'appeler `return`. Il faut noter que l'appel à pthread_join(3) ne se termine que lorsque le thread spécifié comme premier argument se termine. Si ce thread ne se termine pas pour n'importe quelle raison, l'appel à pthread_join(3) ne se terminera pas non plus.

6.2 Communication entre threads

Lorsque un programme a été décomposé en plusieurs threads, ceux-ci ne sont en général pas complètement indépendants et ils doivent communiquer entre eux. Cette communication entre threads est un problème complexe comme nous allons le voir. Avant d'aborder ce problème, il est utile de revenir à l'organisation d'un processus et

de ses threads en mémoire. La figure ci-dessous illustre schématiquement l'organisation de la mémoire après la création d'un thread POSIX.

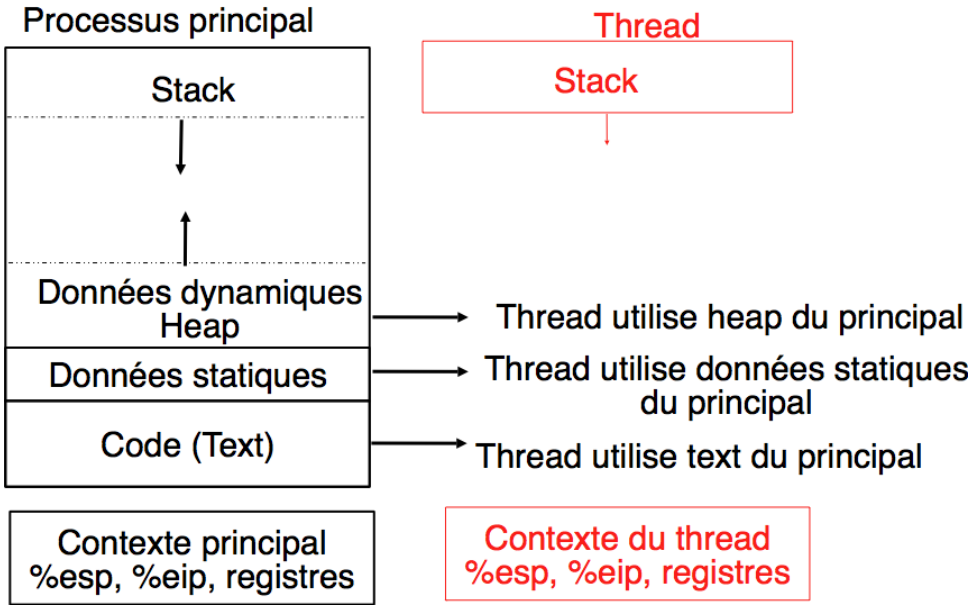


FIGURE 6.4 – Organisation de la mémoire après la création d'un thread POSIX

Le programme principal et le thread qu'il a créé partagent trois zones de la mémoire : le *segment text* qui comprend l'ensemble des instructions qui composent le programme, le *segment de données* qui comprend toutes les données statiques, initialisées ou non (c'est-à-dire les constantes, les variables globales ou encore les chaînes de caractères) et enfin le *heap*. Autant le programme principal que son thread peuvent accéder à n'importe quelle information se trouvant en mémoire dans ces zones. Par contre, le programme principal et le thread qu'il vient de créer ont chacun leur propre contexte et leur propre pile.

La première façon pour un processus de communiquer avec un thread qu'il a lancé est d'utiliser les arguments de la fonction de démarrage du thread et la valeur retournée par le thread que le processus principal peut récupérer via l'appel à pthread_join(3posix). C'est un canal de communication très limité qui ne permet pas d'échange d'information pendant l'exécution du thread.

Il est cependant assez facile pour un processus de partager de l'information avec ses threads ou même de partager de l'information entre plusieurs threads. En effet, tous les threads d'un processus ont accès aux mêmes variables globales et au même *heap*. Il est donc tout à fait possible pour n'importe quel thread de modifier la valeur d'une variable globale. Deux threads qui réalisent un calcul peuvent donc stocker des résultats intermédiaires dans une variable globale ou un tableau global. Il en va de même pour l'utilisation d'une zone de mémoire allouée par malloc(3). Chaque thread qui dispose d'un pointeur vers cette zone mémoire peut en lire le contenu ou en modifier la valeur.

Malheureusement, permettre à tous les threads de lire et d'écrire simultanément en mémoire peut être une source de problèmes. C'est une des difficultés majeures de l'utilisation de threads. Pour s'en convaincre, considérons l'exemple ci-dessous [3].

```
long global=0;
int increment(int i) {
  return i+1;
}
void *func(void * param) {
  for(int j=0;j<1000000;j++) {
    global=increment(global);
  }
  return(NULL);
}
```

3. Le programme complet est accessible via `/Threads/S5-src/pthread-test.c`

Dans cet exemple, la variable `global` est incrémentée `1000000` de fois par la fonction `func`. Après l'exécution de cette fonction, la variable `global` contient la valeur `1000000`. Sur une machine multiprocesseurs, un programmeur pourrait vouloir en accélérer les performances en lançant plusieurs threads qui exécutent chacun la fonction `func`. Cela pourrait se faire en utilisant par exemple la fonction `main` ci-dessous.

```
int main (int argc, char *argv[])  {
  pthread_t thread[NTHREADS];
  int err;
  for(int i=0;i<NTHREADS;i++) {
   err=pthread_create(&(thread[i]),NULL,&func,NULL);
   if(err!=0)
     error(err,"pthread_create");
  }
  /*...*/
  for(int i=NTHREADS-1;i>=0;i--) {
    err=pthread_join(thread[i],NULL);
    if(err!=0)
      error(err,"pthread_join");
  }
  printf("global: %ld\n",global);
  return(EXIT_SUCCESS);
}
```

Ce programme lance alors 4 threads d'exécution qui incrémentent chacun un million de fois la variable `global`. Celle-ci étant initialisée à `0`, la valeur affichée par printf(3) à la fin de l'exécution doit donc être `4000000`. L'exécution du programme ne confirme malheureusement pas cette attente.

```
$ for i in {1..5}; do ./pthread-test;  done
global: 3408577
global: 3175353
global: 1994419
global: 3051040
global: 2118713
```

Non seulement la valeur attendue (`4000000`) n'est pas atteinte, mais en plus la valeur change d'une exécution du programme à la suivante. C'est une illustration du problème majeur de la découpe d'un programme en threads. Pour bien comprendre le problème, il est utile d'analyser en détails les opérations effectuées par deux threads qui exécutent la ligne `global=increment(global);`.

La variable `global` est stockée dans une zone mémoire qui est accessibles aux deux threads. Appelons-les *T1* et *T2*. L'exécution de cette ligne par un thread nécessite l'exécution de plusieurs instructions en assembleur. Tout d'abord, il faut charger la valeur de la variable `global` depuis la mémoire vers un registre. Ensuite, il faut placer cette valeur sur la pile du thread puis appeler la fonction `increment`. Cette fonction récupère son argument sur la pile du thread, la place dans un registre, incrémente le contenu du registre et sauvegarde le résultat dans le registre `%eax`. Le résultat est retourné dans la fonction `func` et la variable globale peut enfin être mise à jour.

Malheureusement les difficultés surviennent lorsque deux threads exécutent en même temps la ligne `global=increment(global);`. Supposons qu'à cet instant, la valeur de la variable `global` est `1252`. Le premier thread charge une copie de cette variable sur sa pile. Le second fait de même. Les deux threads ont donc chacun passé la valeur `1252` comme argument à la fonction `increment`. Celle-ci s'exécute et retourne la valeur `1253` que chaque thread va récupérer dans `%eax`. Chaque thread va ensuite transférer cette valeur dans la zone mémoire correspondant à la variable `global`. Si les deux threads exécutent l'instruction assembleur correspondante exactement au même moment, les deux écritures en mémoire seront sérialisées par les processeurs sans que l'on ne puisse a priori déterminer quelle écriture se fera en premier [McKenney2005]. Alors que les deux threads ont chacun exécuté un appel à la fonction `increment`, la valeur de la variable n'a finalement été incrémentée qu'une seule fois même si cette valeur a été transférée deux fois en mémoire. Ce problème se reproduit fréquemment. C'est pour cette raison que la valeur de la variable `global` n'est pas modifiée comme attendu.

Ce problème d'accès concurrent à une zone de mémoire par plusieurs threads est un problème majeur dans le développement de programmes découpés en threads, que ceux-ci s'exécutent sur des ordinateurs monoprocesseurs ou multiprocesseurs. Dans la littérature, il est connu sous le nom de problème de la *section*

critique. La *section critique* peut être définie comme étant une séquence d'instructions qui ne peuvent *jamais* être exécutées par plusieurs threads simultanément. Dans l'exemple ci-dessus, il s'agit de la ligne `global=increment(global);`. Dans d'autres types de programmes, la section critique peut être beaucoup plus grande et comprendre par exemple la mise à jour d'une base de données. En pratique, on retrouvera une section critique chaque fois que deux threads peuvent modifier ou lire la valeur d'une même zone de la mémoire.

6.3 Coordination entre threads

L'utilisation de plusieurs threads dans un programme fonctionnant sur un seul ou plusieurs processeurs nécessite l'utilisation de mécanismes de coordination entre ces threads. Ces mécanismes ont comme objectif d'éviter que deux threads ne puissent modifier ou tester de façon non coordonnée la même zone de la mémoire.

6.3.1 Exclusion mutuelle

Le premier problème important à résoudre lorsque l'on veut coordonner plusieurs threads d'exécution d'un même processus est celui de l'*exclusion mutuelle*. Ce problème a été initialement proposé par Dijkstra en 1965 [Dijkstra1965]. Il peut être reformulé de la façon suivante pour un programme décomposé en threads.

Considérons un programme décomposé en N threads d'exécution. Supposons également que chaque thread d'exécution est cyclique, c'est-à-dire qu'il exécute régulièrement le même ensemble d'instructions, sans que la durée de ce cycle ne soit fixée ni identique pour les N threads. Chacun de ces threads peut être décomposé en deux parties distinctes. La première est la partie non-critique. C'est un ensemble d'instructions qui peuvent être exécutées par le thread sans nécessiter la moindre coordination avec un autre thread. Plus précisément, tous les threads peuvent exécuter simultanément leur partie non-critique. La seconde partie du thread est appelée sa *section critique*. Il s'agit d'un ensemble d'instructions qui ne peuvent être exécutées que par un seul et unique thread. Le problème de l'*exclusion mutuelle* est de trouver un algorithme qui permet de garantir qu'il n'y aura jamais deux threads qui simultanément exécuteront des instructions de leur section critique.

Cela revient à dire qu'il n'y aura pas de violation de la section critique. Une telle violation pourrait avoir des conséquences catastrophiques sur l'exécution du programme. Cette propriété est une propriété de *sûreté* (*safety* en anglais). Dans un programme découpé en threads, une propriété de *sûreté* est une propriété qui doit être vérifiée à tout instant de l'exécution du programme.

En outre, une solution au problème de l'*exclusion mutuelle* doit satisfaire les contraintes suivantes [Dijkstra1965] :

1. La solution doit considérer tous les threads de la même façon et ne peut faire aucune hypothèse sur la priorité relative des différents threads.
2. La solution ne peut faire aucune hypothèse sur la vitesse relative ou absolue d'exécution des différents threads. Elle doit rester valide quelle que soit la vitesse d'exécution (non nulle) de chaque thread.
3. La solution doit permettre à un thread de s'arrêter en dehors de sa section critique sans que cela n'invalide la contrainte d'exclusion mutuelle
4. Si un ou plusieurs threads souhaitent entamer leur section critique, aucun de ces threads ne doit pouvoir être empêché indéfiniment d'accéder à sa section critique.

La troisième contrainte implique que la terminaison ou le crash d'un des threads ne doit pas avoir d'impact sur le fonctionnement du programme complet et le respect de la contrainte d'exclusion mutuelle pour la section critique.

La quatrième contrainte est un peu plus subtile mais tout aussi importante. Toute solution au problème de l'exclusion mutuelle contient nécessairement un mécanisme qui permet de bloquer l'exécution d'un thread pendant qu'un autre exécute sa section critique. Il est important qu'un thread puisse accéder à sa section critique si il le souhaite. C'est un exemple de propriété de *vivacité* (*liveness* en anglais). Une propriété de *vivacité* est une propriété qui ne peut pas être éternellement invalidée. Dans notre exemple, un thread ne pourra jamais être empêché d'accéder à sa section critique.

Exclusion mutuelle sur monoprocesseurs

Même si les threads sont très utiles dans des ordinateurs multiprocesseurs, ils ont été inventés et utilisés d'abord sur des processeurs capables d'exécuter un seul thread d'exécution à la fois. Sur un tel processeur, les threads

d'exécution sont entrelacés plutôt que d'être exécutés réellement simultanément. Cet entrelacement est réalisé par le système d'exploitation.

Les systèmes d'exploitation de la famille Unix permettent d'exécuter plusieurs programmes *en même temps* sur un ordinateur, même si il est équipé d'un processeur qui n'est capable que d'exécuter un thread à la fois. Cette fonctionnalité est souvent appelée la *multiprogrammation* ou le *multitâche* (ou *multitasking* en anglais). Cette exécution simultanée de plusieurs programmes n'est en pratique qu'une illusion puisque le processeur ne peut qu'exécuter qu'une séquence d'instructions à la fois.

Pour donner cette illusion, un système d'exploitation multitâche tel que Unix exécute régulièrement des changements de contexte entre threads. Le *contexte* d'un thread est composé de l'ensemble des contenus des registres qui sont nécessaires à son exécution (y compris le contenu des registres spéciaux tels que `%esp`, `%eip` ou `%eflags`). Ces registres définissent l'état du thread du point de vue du processeur. Pour passer de l'exécution du thread *T1* à l'exécution du thread *T2*, le système d'exploitation doit initier un *changement de contexte*. Pour réaliser ce changement de contexte, le système d'exploitation initie le transfert du contenu des registres utilisés par le thread *T1* vers une zone mémoire lui appartenant. Il transfère ensuite depuis une autre zone mémoire lui appartenant le contexte du thread *T2*. Si ce changement de contexte est effectué fréquemment, il peut donner l'illusion à l'utilisateur que plusieurs threads ou programmes s'exécutent simultanément.

Sur un système Unix, il y a deux types d'événements qui peuvent provoquer un changement de contexte.

1. Le hardware génère une *interruption*
2. Un thread exécute un *appel système bloquant*

Ces deux types d'événements sont fréquents et il est important de bien comprendre comment ils sont traités par le système d'exploitation.

Une *interruption* est un signal électronique qui est généré par un dispositif connecté au microprocesseur. De nombreux dispositifs d'entrées-sorties comme les cartes réseau ou les contrôleurs de disque peuvent générer une interruption lorsqu'une information a été lue ou reçue et doit être traitée par le processeur. En outre, chaque ordinateur dispose d'une horloge temps réel qui génère des interruptions à une fréquence déterminée par le système d'exploitation mais qui est généralement comprise entre quelques dizaines et quelques milliers de *Hz*. Ces interruptions nécessitent un traitement rapide de la part du système d'exploitation. Pour cela, le processeur vérifie, à la fin de l'exécution de *chaque* instruction si un signal d'interruption [4] est présent. Si c'est le cas, le processeur passe automatiquement en mode protégé et positionne le compteur de programme vers une routine de traitement d'interruption, faisant partie du *kernel*. Cette routine sauvegarde tout d'abord en mémoire le contexte du thread en cours d'exécution. Ensuite, elle analyse l'interruption présente et lance les fonctions du système d'exploitation nécessaires à son traitement.

Le deuxième type d'événement est l'exécution d'un appel système bloquant. Un thread exécute un *appel système* chaque fois qu'il doit interagir avec le système d'exploitation. Ces appels peuvent être exécutés directement ou via une fonction de la librairie [5]. Il existe deux types d'appels systèmes : les appels bloquants et les appels non-bloquants. Un appel système non-bloquant est un appel système que le système d'exploitation peut exécuter immédiatement. Cet appel retourne généralement une valeur qui fait partie des données du système d'exploitation lui-même. L'appel gettimeofday(2) qui permet de récupérer l'heure actuelle est un exemple d'appel non-bloquant. Un appel système bloquant est un appel système dont le résultat ne peut pas toujours être fourni immédiatement. Les lectures d'information en provenance de l'entrée standard (et donc généralement du clavier) sont un bon exemple simple d'appel système bloquant. Considérons un thread qui exécute la fonction de la librairie getchar(3) qui retourne un caractère lu sur *stdin*. Cette fonction utilise l'appel système read(2) pour lire un caractère sur *stdin*. Bien entendu, le système d'exploitation est obligé d'attendre que l'utilisateur presse une touche sur le clavier pour pouvoir fournir le résultat de cet appel système à l'utilisateur. Pendant tout le temps qui s'écoule entre l'exécution de getchar(3) et la pression d'une touche sur le clavier, le thread est bloqué par le système d'exploitation. Plus aucune instruction du thread n'est exécutée tant que la fonction getchar(3) ne s'est pas terminée et le contexte du thread est mis en attente dans une zone mémoire gérée par le système d'exploitation. Il sera redémarré automatiquement par le système d'exploitation lorsque la donnée attendue sera disponible.

Ces interactions entre les threads et le système d'exploitation sont importantes. Pour bien les comprendre, il est utile de noter qu'un thread peut se trouver dans trois états différents du point de vue de son interaction avec le système d'exploitation. Ces trois états sont illustrés dans la figure ci-dessous.

4. De nombreux processeurs supportent plusieurs signaux d'interruption différents. Dans le cadre de ce cours, nous nous limiterons à l'utilisation d'un seul signal de ce type.
5. Les appels systèmes sont décrits dans la section 2 des pages de manuel tandis que la section 3 décrit les fonctions de la librairie.

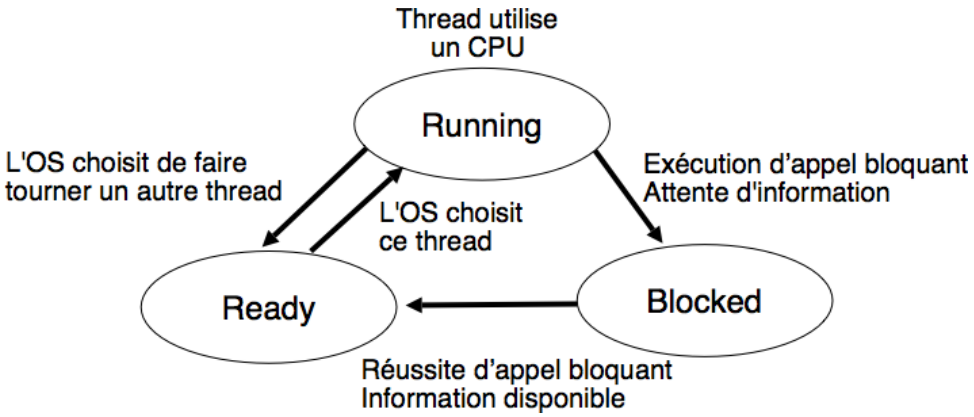


FIGURE 6.5 – Etats d'un thread d'exécution

Lorsqu'un thread est créé avec la fonction pthread_create(3), il est placé dans l'état *Ready*. Dans cet état, les instructions du thread ne s'exécutent sur aucun processeur mais il est prêt à être exécuté dès qu'un processeur se libérera. Le deuxième état pour un thread est l'état *Running*. Dans cet état, le thread est exécuté sur un des processeurs du système. Le dernier état est l'état *Blocked*. Un thread est dans l'état *Blocked* lorsqu'il a exécuté un appel système bloquant et que le système d'exploitation attend l'information permettant de retourner le résultat de l'appel système. Pendant ce temps, les instructions du thread ne s'exécutent sur aucun processeur.

Les transitions entre les différents états d'un thread sont gérées par le système d'exploitation. Lorsque plusieurs threads d'exécution sont simultanément actifs, le système d'exploitation doit arbitrer les demandes d'utilisation du CPU de chaque thread. Cet arbitrage est réalisé par l'ordonnanceur (ou *scheduler* en anglais). Le *scheduler* est un ensemble d'algorithmes qui sont utilisés par le système d'exploitation pour sélectionner le ou les threads qui peuvent utiliser un processeur à un moment donné. Il y a souvent plus de threads qui sont dans l'état *Ready* que de processeurs disponibles et le scheduler doit déterminer quels sont les threads à exécuter. Il suffit, dans le contexte de ce chapitre, de considérer que le scheduler choisit un thread à exécuter ; il s'agit d'un mécanisme. Le choix lui-même est une politique de scheduling, qui est séparée du mécanisme lui-même gérant l'état des processus. Une politique possible (bien que peu efficace) serait de choisir le thread à exécuter de manière aléatoire parmi les threads en état *Ready*. Nous verrons des politiques plus intelligentes et efficaces plus tard, dans le chapitre consacré spécifiquement au scheduling.

Il est utile d'analyser en détails quels sont les événements qui peuvent provoquer des transitions entre les états d'un thread. Certains de ces événements sont provoqués par le thread lui-même. C'est le cas de la transition entre l'état *Running* et l'état *Blocked*. Elle se produit lorsque le thread exécute un *appel système bloquant*. Dans ce cas, un processeur redevient disponible et le scheduler peut sélectionner un autre thread pour s'exécuter sur ce processeur. La transition entre l'état *Blocked* et l'état *Running* dépend elle du système d'exploitation, directement lorsque le thread a été bloqué par le système d'exploitation ou indirectement lorsque le système d'exploitation attend une information venant d'un dispositif d'entrées-sorties. Les transitions entre les états *Running* et *Ready* dépendent elles entièrement du système d'exploitation. Elles se produisent lors de l'exécution du scheduler. Celui-ci est exécuté lorsque certaines interruptions surviennent. Il est exécuté à chaque interruption d'horloge. Cela permet de garantir l'exécution régulière du scheduler même si les seuls threads actifs exécutent une boucle infinie telle que `while(true);`. A l'occasion de cette interruption, le scheduler mesure le temps d'exécution de chaque thread et si un thread a consommé beaucoup de temps CPU alors que d'autres threads sont dans l'état *Ready*, le scheduler forcera un changement de contexte pour permettre à un autre thread de s'exécuter. De la même façon, une interruption relative à un dispositif d'entrées-sorties peut faire transiter un thread de l'état *Blocked* à l'état *Ready*. Cette modification du nombre de threads dans l'état *Ready* peut amener le scheduler à devoir effectuer un changement de contexte pour permettre à ce thread de poursuivre son exécution.

Note : Un thread peut demander de passer la main.

Dans la plupart de nos exemples, les threads cherchent en permanence à exécuter des instructions. Ce n'est pas nécessairement le cas de tous les threads d'un programme. Par exemple, une application de calcul scientifique pourrait être découpée en *N+1* threads. Les *N* premiers threads réalisent le calcul tandis que le dernier calcule des statistiques. Ce dernier thread ne doit pas consommer de ressources et être en compétition pour le processeur avec les autres threads. La librairie thread POSIX contient la fonction pthread_yield(3) qui peut être utilisée par

un thread pour indiquer explicitement qu'il peut être remplacé par un autre thread. Si un thread ne doit s'exécuter qu'à intervalles réguliers, il est préférable d'utiliser des appels à sleep(3) ou usleep(3). Ces fonctions de la librairie permettent de demander au système d'exploitation de bloquer le thread pendant un temps au moins égal à l'argument de la fonction.

Sur une machine monoprocesseur, tous les threads s'exécutent sur le même processeur. Une violation de section critique peut se produire lorsque le scheduler décide de réaliser un changement de contexte alors qu'un thread se trouve dans sa section critique. Si la section critique d'un thread ne contient ni d'appel système bloquant ni d'appel à pthread_yield(3), ce changement de contexte ne pourra se produire que si une interruption survient. Une solution pour résoudre le problème de l'exclusion mutuelle sur un ordinateur monoprocesseur pourrait donc être la suivante :

```
disable_interrupts();
// début section critique
// ...
// fin section critique
enable_interrupts();
```

Cette solution est possible, mais elle souffre de plusieurs inconvénients majeurs. Tout d'abord, une désactivation des interruptions perturbe le fonctionnement du système puisque sans interruptions, la plupart des opérations d'entrées-sorties et l'horloge sont inutilisables. Une telle désactivation ne peut être que très courte, par exemple pour modifier une ou quelques variables en mémoire. Ensuite, la désactivation des interruptions, comme d'autres opérations relatives au fonctionnement du matériel, est une opération privilégiée sur un microprocesseur. Elle ne peut être réalisée que par le système d'exploitation. Il faudrait donc imaginer un appel système qui permettrait à un thread de demander au système d'exploitation de désactiver les interruptions. Si un tel appel système existait, le premier programme qui exécuterait `disable_interrupts();` sans le faire suivre de `enable_interrupts();` quelques instants après pourrait rendre la machine complètement inutilisable puisque sans interruption plus aucune opération d'entrée-sortie n'est possible et qu'en plus le scheduler ne peut plus être activé par l'interruption d'horloge. Pour toutes ces raisons, la désactivation des interruptions n'est pas un mécanisme utilisable par les threads pour résoudre le problème de l'exclusion mutuelle [6].

Coordination par Mutex

Le premier mécanisme de coordination entre threads dans la librairie POSIX sont les *mutex*. Un *mutex* (abréviation de *mutual exclusion*) est une structure de données qui permet de contrôler l'accès à une ressource. Un *mutex* qui contrôle une ressource peut se trouver dans deux états :

– *libre* (ou *unlocked* en anglais). Cet état indique que la ressource est libre et peut être utilisée sans risquer de provoquer une violation d'exclusion mutuelle.
– *réservée* (ou *locked* en anglais). Cet état indique que la ressource associée est actuellement utilisée et qu'elle ne peut pas être utilisée par un autre thread.

Un *mutex* est toujours associé à une ressource. Cette ressource peut être une variable globale comme dans les exemples précédents, mais cela peut aussi être une structure de données plus complexe, une base de données, un fichier, ... Un mutex s'utilise par l'intermédiaire de deux fonctions de base. La fonction *lock* permet à un thread d'acquérir l'usage exclusif d'une ressource. Si la ressource est libre, elle est marquée comme réservée et le thread y accède directement. Si la ressource est occupée, le thread est bloqué par le système d'exploitation jusqu'à ce qu'elle ne devienne libre. A ce moment, le thread pourra poursuivre son exécution et utilisera la ressource avec la certitude qu'aucun autre thread ne pourra faire de même. Lorsque le thread a terminé d'utiliser la ressource associée au mutex, il appelle la fonction *unlock*. Cette fonction vérifie d'abord si un ou plusieurs autres threads sont en attente pour cette ressource (c'est-à-dire qu'ils ont appelé la fonction *lock* mais celle-ci n'a pas encore réussi). Si c'est le cas, un (et un seul) thread est choisi parmi les threads en attente et celui-ci accède à la ressource. Il est important de noter qu'un programme ne peut faire aucune hypothèse sur l'ordre dans lequel les threads qui sont en attente sur un *mutex* pourront accéder à la ressource partagée. Le programme doit être conçu en faisant l'hypothèse que si plusieurs threads sont bloqués sur un appel à *lock* pour un mutex, le thread qui sera libéré est choisi aléatoirement.

6. Certains systèmes d'exploitation utilisent une désactivation parfois partielle des interruptions pour résoudre des problèmes d'exclusion mutuelle qui portent sur quelques instructions à l'intérieur du système d'exploitation lui-même. Il faut cependant noter qu'une désactivation des interruptions peut être particulièrement coûteuse en termes de performances dans un environnement multiprocesseurs.

Sans entrer dans des détails de mise en œuvre qui dépassent le contexte de ce cours, on peut schématiquement représenter un *mutex* comme étant une structure de données qui contient deux informations :
– la valeur actuelle du *mutex* (*locked* ou *unlocked*)
– une queue contenant l'ensemble des threads qui sont bloqués en attente du mutex
Schématiquement, l'implémentation des fonctions `lock` et `unlock` peut être représentée par le code ci-dessous.

```
lock(mutex m) {
 if(m.val==unlocked)
 {
   m.val=locked;
 }
 else
 {
   // Place this thread in m.queue;
   // This thread is blocked;
 }
}
```

Le fonction `lock` vérifie si le *mutex* est libre. Dans ce cas, le *mutex* est marqué comme réservé et la fonction `lock` réussi. Sinon, le thread qui a appelé la fonction `lock` est placé dans la queue associée au *mutex* et passe dans l'état *Blocked* jusqu'à ce qu'un autre thread ne libère le mutex.

```
unlock(mutex m) {
 if(m.queue is empty)
 {
   m.val=unlocked;
 }
 else
 {
  // Remove one thread(T) from m.queue;
  // Mark Thread(T) as ready to run;
 }
}
```

La fonction `unlock` vérifie d'abord l'état de la queue associée au *mutex*. Si la queue est vide, cela indique qu'aucun thread n'est en attente. Dans ce cas, la valeur du *mutex* est mise à *unlocked* et la fonction se termine. Sinon, un des threads en attente dans la queue associée au *mutex* est choisi et marqué comme prêt à s'exécuter. Cela indique implicitement que l'appel à `lock` fait par ce thread réussi et qu'il peut accéder à la ressource.

Le code présenté ci-dessus n'est qu'une illustration du fonctionnement des opérations `lock` et `unlock`. Pour que ces opérations fonctionnent correctement, il faut bien entendu que les modifications aux valeurs du *mutex* et à la queue qui y est associée se fassent en garantissant qu'un seul thread exécute l'une de ces opérations sur un *mutex* à un instant donné. Nous verrons dans le détail comment garantir cela en utilisant des instructions spécifiques (dites *atomiques*) du jeu d'instruction, dans un prochain chapitre.

Les *mutex* sont fréquemment utilisés pour protéger l'accès à une zone de mémoire partagée. Ainsi, si la variable globale *g* est utilisée en écriture et en lecture par deux threads, celle-ci devra être protégée par un *mutex*. Toute modification de cette variable devra être entourée par des appels à *lock* et *unlock*.

En C, cela se fait en utilisant les fonctions pthread_mutex_lock(3posix) et pthread_mutex_unlock(3posix). Un *mutex* POSIX est représenté par une structure de données de type *pthread_mutex_t* qui est définie dans le fichier pthread.h. Avant d'être utilisé, un *mutex* doit être initialisé via la fonction pthread_mutex_init(3posix) et lorsqu'il n'est plus nécessaire, les ressources qui lui sont associées doivent être libérées avec la fonction pthread_mutex_destroy(3posix).

L'exemple ci-dessous reprend le programme dans lequel une variable globale est incrémentée par plusieurs threads.

```
#include <pthread.h>
#define NTHREADS 4

long global=0;
pthread_mutex_t mutex_global;
```

```
int increment(int i) {
  return i+1;
}

void *func(void * param) {
  int err;
  for(int j=0;j<1000000;j++) {
    err=pthread_mutex_lock(&mutex_global);
    if(err!=0)
      error(err,"pthread_mutex_lock");
    global=increment(global);
    err=pthread_mutex_unlock(&mutex_global);
    if(err!=0)
      error(err,"pthread_mutex_unlock");
  }
  return(NULL);
}

int main (int argc, char *argv[])  {
  pthread_t thread[NTHREADS];
  int err;

  err=pthread_mutex_init( &mutex_global, NULL);
  if(err!=0)
      error(err,"pthread_mutex_init");

  for(int i=0;i<NTHREADS;i++) {
    err=pthread_create(&(thread[i]),NULL,&func,NULL);
    if(err!=0)
      error(err,"pthread_create");
  }
  for(int i=0; i<1000000000;i++) { /*...*/ }

  for(int i=NTHREADS-1;i>=0;i--) {
    err=pthread_join(thread[i],NULL);
    if(err!=0)
      error(err,"pthread_join");
  }

  err=pthread_mutex_destroy(&mutex_global);
  if(err!=0)
    error(err,"pthread_mutex_destroy");

  printf("global: %ld\n",global);

  return(EXIT_SUCCESS);
}
```

Il est utile de regarder un peu plus en détails les différentes fonctions utilisées par ce programme. Tout d'abord, la ressource partagée est ici la variable `global`. Dans l'ensemble du programme, l'accès à cette variable est protégé par le *mutex* `mutex_global`. Celui-ci est représenté par une structure de données de type `pthread_mutex_t`.

Avant de pouvoir utiliser un *mutex*, il est nécessaire de l'initialiser. Cette initialisation est effectuée par la fonction pthread_mutex_init(3posix) qui prend deux arguments[7]. Le premier est un pointeur vers une structure `pthread_mutex_t` et le second un pointeur vers une structure `pthread_mutexattr_t` contenant les attributs de ce *mutex*. Tout comme lors de la création d'un thread, ces attributs permettent de spécifier des paramètres à la création du *mutex*. Ces attributs peuvent être manipulés en utilisant les fonctions pthread_mutexattr_gettype(3posix) et pthread_mutexattr_settype(3posix). Dans le cadre de ces notes, nous utilis-

7. Linux supporte également la macro `PTHREAD_MUTEX_INITIALIZER` qui permet d'initialiser directement un `pthread_mutex_t` déclaré comme variable globale. Dans cet exemple, la déclaration aurait été : `pthread_mutex_t global_mutex=PTHREAD_MUTEX_INITIALIZER;` et l'appel à pthread_mutex_init(3posix) aurait été inutile. Comme il s'agit d'une extension spécifique à Linux, il est préférable de ne pas l'utiliser pour garantir la portabilité du code.

erons exclusivement les attributs par défaut et créerons toujours un *mutex* en passant `NULL` comme second argument à la fonction pthread_mutex_init(3posix).

Lorsqu'un *mutex* POSIX est initialisé, la ressource qui lui est associée est considérée comme libre. L'accès à la ressource doit se faire en précédant tout accès à la ressource par un appel à la fonction pthread_mutex_lock(3posix). En fonction des attributs spécifiés à la création du *mutex*, il peut y avoir de très rares cas où la fonction retourne une valeur non nulle. Dans ce cas, le type d'erreur est indiqué via *errno*. Lorsque le thread n'a plus besoin de la ressource protégée par le mutex, il doit appeler la fonction pthread_mutex_unlock(3posix) pour libérer la ressource protégée.

pthread_mutex_lock(3posix) et pthread_mutex_unlock(3posix) sont toujours utilisés en couple. pthread_mutex_lock(3posix) doit toujours précéder l'accès à la ressource partagée et pthread_mutex_unlock(3posix) doit être appelé dès que l'accès exclusif à la ressource partagée n'est plus nécessaire.

L'utilisation des mutex permet de résoudre correctement le problème de l'exclusion mutuelle. Pour s'en convaincre, considérons le programme ci-dessus et les threads qui exécutent la fonction `func`. Celle-ci peut être résumée par les trois lignes suivantes :

```
pthread_mutex_lock(&mutex_global);
global=increment(global);
pthread_mutex_unlock(&mutex_global);
```

Pour montrer que cette solution répond bien au problème de l'exclusion mutuelle, il faut montrer qu'elle respecte la propriété de sûreté et la propriété de vivacité. Pour la propriété de sûreté, c'est par construction des *mutex* et parce que chaque thread exécute pthread_mutex_lock(3posix) avant d'entrer en section critique et pthread_mutex_unlock(3posix) dès qu'il en sort. Considérons le cas de deux threads qui sont en concurrence pour accéder à cette section critique. Le premier exécute pthread_mutex_lock(3posix). Il accède à sa section critique. A partir de cet instant, le second thread sera bloqué dès qu'il exécute l'appel à pthread_mutex_lock(3posix). Il restera bloqué dans l'exécution de cette fonction jusqu'à ce que le premier thread sorte de sa section critique et exécute pthread_mutex_unlock(3posix). A ce moment, le premier thread n'est plus dans sa section critique et le système peut laisser le second y entrer en terminant l'exécution de l'appel à pthread_mutex_lock(3posix). Si un troisième thread essaye à ce moment d'entrer dans la section critique, il sera bloqué sur son appel à pthread_mutex_lock(3posix).

Pour montrer que la propriété de vivacité est bien respectée, il faut montrer qu'un thread ne sera pas empêché éternellement d'entrer dans sa section critique. Un thread peut être empêché d'entrer dans sa section critique en étant bloqué sur l'appel à pthread_mutex_lock(3posix). Comme chaque thread exécute pthread_mutex_unlock(3posix) dès qu'il sort de sa section critique, le thread en attente finira par être exécuté. Pour qu'un thread utilisant le code ci-dessus ne puisse jamais entrer en section critique, il faudrait qu'il y aie en permanence plusieurs threads en attente sur pthread_mutex_unlock(3posix) et que notre thread ne soit jamais sélectionné par le système lorsque le thread précédent termine sa section critique.

6.3.2 Le problème des philosophes

Le problème de l'exclusion mutuelle considère le cas de plusieurs threads qui se partagent une ressource commune qui doit être manipulée de façon exclusive. C'est un problème fréquent qui se pose lorsque l'on développe des programmes décomposés en différents threads d'exécution, mais ce n'est pas le seul. En pratique, il est souvent nécessaire de coordonner l'accès de *plusieurs* threads à *plusieurs* ressources, chacune de ces ressources devant être utilisée de façon exclusive. Cette utilisation de plusieurs ressources simultanément peut poser des difficultés. Un des problèmes classiques qui permet d'illustrer la difficulté de coordonner l'accès à plusieurs ressources est le *problème des philosophes*. Ce problème a été proposé par Dijkstra et illustre en termes simples la difficulté de coordonner l'accès à plusieurs ressources.

Dans le *problème des philosophes*, un ensemble de philosophes doivent se partager des baguettes pour manger. Tous les philosophes se trouvent dans une même pièce qui contient une table circulaire. Chaque philosophe dispose d'une place qui lui est réservée sur cette table. La table comprend autant de baguettes que de chaises et une baguette est placée entre chaque paire de chaises. Chaque philosophe est modélisé sous la forme d'un thread qui effectue deux types d'actions : *penser* et *manger*. Pour pouvoir manger, un philosophe doit obtenir la baguette qui se trouve à sa gauche et la baguette qui se trouve à sa droite. Lorsqu'il a fini de manger, il peut retourner à

son activité philosophale. La figure ci-dessous illustre une table avec les assiettes de trois philosophes et les trois baguettes qui sont à leur disposition.

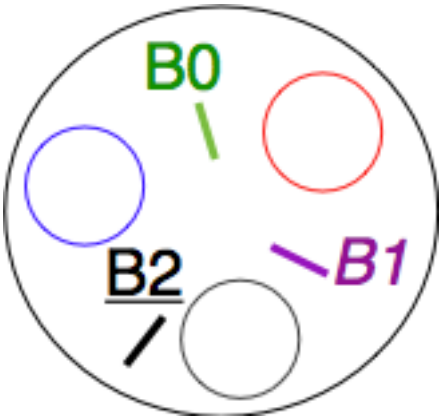


FIGURE 6.6 – Problème des philosophes

Ce problème de la vie courante peut se modéliser en utilisant un programme C avec les threads POSIX. Chaque baguette est une ressource partagée qui ne peut être utilisée que par un philosophe à la fois. Elle peut donc être modélisée par un *mutex*. Chaque philosophe est modélisé par un thread qui pense puis ensuite appelle pthread_mutex_lock(3posix) pour obtenir chacune de ses baguettes. Le philosophe peut ensuite manger puis il libère ses baguettes en appelant pthread_mutex_unlock(3posix). L'extrait [8] ci-dessous comprend les fonctions utilisées par chacun des threads.

```
#include <pthread.h>
#include <stdio.h>
#include <stdlib.h>
#include <stdbool.h>
#include <unistd.h>

#define PHILOSOPHES 3

pthread_t phil[PHILOSOPHES];
pthread_mutex_t baguette[PHILOSOPHES];

void mange(int id) {
  printf("Philosophe [%d] mange\n",id);
  for(int i=0;i< rand(); i++) {
    // philosophe mange
  }
}

void* philosophe ( void* arg )
{
  int *id=(int *) arg;
  int left = *id;
  int right = (left + 1) % PHILOSOPHES;
  while(true) {
    printf("Philosophe [%d] pense\n",*id);
    pthread_mutex_lock(&baguette[left]);
    printf("Philosophe [%d] possède baguette gauche [%d]\n",*id,left);
    pthread_mutex_lock(&baguette[right]);
    printf("Philosophe [%d] possède baguette droite [%d]\n",*id,right);
    mange(*id);
    pthread_mutex_unlock(&baguette[left]);
    printf("Philosophe [%d] a libéré baguette gauche [%d]\n",*id,left);
    pthread_mutex_unlock(&baguette[right]);
```

8. Le programme complet est `/Threads/S6-src/pthread-philo.c`

```
    printf("Philosophe [%d] a libéré baguette droite [%d]\n",*id,right);
  }
  return (NULL);
}
```

Malheureusement, cette solution ne permet pas de résoudre le problème des philosophes. En effet, la première exécution du programme (`/Threads/S6-src/pthread-philo.c`) indique à l'écran que les philosophes se partagent les baguettes puis après une fraction de seconde le programme s'arrête en affichant une sortie qui se termine :

```
Philosophe [0] a libéré baguette gauche [0]
Philosophe [0] a libéré baguette droite [1]
Philosophe [0] pense
Philosophe [0] possède baguette gauche [0]
Philosophe [2] possède baguette gauche [2]
Philosophe [1] possède baguette gauche [1]
```

Ce blocage de programme est un autre problème majeur auquel il faut faire attention lorsque l'on découpe un programme en plusieurs threads d'exécution. Il porte le nom de *deadlock* que l'on peut traduire en français pas étreinte fatale. Un programme est en situation de deadlock lorsque tous ses threads d'exécution sont bloqués et qu'aucun d'entre eux ne peux être débloqué sans exécuter d'instructions d'un des threads bloqués. Dans ce cas particulier, le programme est bloqué parce que le `philosophe[0]` a pu réserver la `baguette[0]`. Malheureusement, en même temps le `philosophe[2]` a obtenu `baguette[2]` et `philosophe[0]` est donc bloqué sur la ligne `pthread_mutex_lock(&baguette[right]);`. Entre temps, le `philosophe[1]` a pu exécuter la ligne `pthread_mutex_lock(&baguette[left]);` et a obtenu `baguette[1]`. Dans cet état, tous les threads sont bloqués sur la ligne `pthread_mutex_lock(&baguette[right]);` et plus aucun progrès n'est possible.

Ce problème de deadlock est un des problèmes les plus graves qui peuvent survenir dans un programme découpé en plusieurs threads. Si les threads entrent dans une situation de deadlock, le programme est complètement bloqué et la seule façon de sortir du deadlock est d'arrêter le programme et de le redémarrer. Ce n'est évidemment pas acceptable. Dans l'exemple des philosophes ci-dessus, le deadlock apparait assez rapidement car les threads passent la majorité de leur temps à exécuter les fonctions de la librairie threads POSIX. En pratique, un thread exécute de nombreuses lignes de code standard et utilise rarement la fonction pthread_mutex_lock(3posix). Dans de tels programmes, les tests ne permettent pas nécessairement de détecter toutes les situations qui peuvent causer un deadlock. Il est important que le programme soit conçu de façon à toujours éviter les deadlocks. Si c'est n'est pas le cas, le programme finira toujours bien à se retrouver en situation de deadlock après un temps plus ou moins long.

Le problème des philosophes est une illustration d'un problème courant dans lequel plusieurs threads doivent utiliser deux ou plusieurs ressources partagées contrôlées par un mutex. Dans la situation de deadlock, chaque thread est bloqué à l'exécution de la ligne `pthread_mutex_lock(&baguette[right]);`. On pourrait envisager de chercher à résoudre le problème en inversant les deux appels à pthread_mutex_lock(3posix) comme ci-dessous :

```
pthread_mutex_lock(&baguette[right]);
pthread_mutex_lock(&baguette[left]);
```

Malheureusement, cette solution ne résout pas le problème du deadlock. La seule différence par rapport au programme précédent est que les mutex qui sont alloués à chaque thread en deadlock ne sont pas les mêmes. Avec cette nouvelle version du programme, lorsque le deadlock survient, les mutex suivants sont alloués :

```
Philosophe [2] possède baguette droite [1]
Philosophe [0] possède baguette droite [2]
Philosophe [1] possède baguette droite [0]
```

Pour comprendre l'origine du deadlock, il faut analyser plus en détails le fonctionnement du programme et l'ordre dans lequel les appels à pthread_mutex_lock(3posix) sont effectués. Trois mutex sont utilisés dans le programme des philosophes : `baguette[0]`, `baguette[1]` et `baguette[2]`. De façon imagée, chaque philosophe s'approprie d'abord la baguette se trouvant à sa gauche et ensuite la baguette se trouvant à sa droite.

Philosophe	premier mutex	second mutex
philosophe[0]	baguette[2]	baguette[0]
philosophe[1]	baguette[0]	baguette[1]
philosophe[2]	baguette[1]	baguette[2]

L'origine du deadlock dans cette solution au problème des philosophes est l'ordre dans lequel les différents philosophes s'approprient les mutex. Le philosophe[0] s'approprie d'abord le mutex baguette[2] et ensuite essaye de s'approprier le mutex baguette[0]. Le philosophe[1] par contre peut lui directement s'approprier le mutex baguette[0]. Au vu de l'ordre dans lequel les mutex sont alloués, il est possible que chaque thread se soit approprié son premier mutex mais soit bloqué sur la réservation du second. Dans ce cas, un deadlock se produit puisque chaque thread est en attente de la libération d'un mutex sans qu'il ne puisse libérer lui-même le mutex qu'il s'est déjà approprié.

Il est cependant possible de résoudre le problème en forçant les threads à s'approprier les mutex qu'ils utilisent dans le même ordre. Considérons le tableau ci-dessous dans lequel philosophe[0] s'approprie d'abord le mutex baguette[0] et ensuite le mutex baguette[2].

Philosophe	premier mutex	second mutex
philosophe[0]	baguette[0]	baguette[2]
philosophe[1]	baguette[0]	baguette[1]
philosophe[2]	baguette[1]	baguette[2]

Avec cet ordre d'allocation des mutex, un deadlock n'est plus possible. Il y aura toujours au moins un philosophe qui pourra s'approprier les deux baguettes dont il a besoin pour manger. Pour s'en convaincre, analysons les différentes exécutions possibles. Un deadlock ne pourrait survenir que si tous les philosophes cherchent à manger simultanément. Si un des philosophes ne cherche pas à manger, ses deux baguettes sont nécessairement libres et au moins un de ses voisins philosophes pourra manger. Lorsque tous les philosophes cherchent à manger simultanément, le mutex baguette[0] ne pourra être attribué qu'à philosophe[0] ou philosophe[1]. Considérons les deux cas possibles.

1. philosophe[0] s'approprie baguette[0]. Dans ce cas, philosophe[1] est bloqué sur son premier appel à pthread_mutex_lock(3posix). Comme philosophe[1] n'a pas pu s'approprier le mutex baguette[0], il ne peut non plus s'approprier baguette[1]. philosophe[2] pourra donc s'approprier le mutex baguette[1]. philosophe[0] et philosophe[2] sont maintenant en compétition pour le mutex baguette[2]. Deux cas sont à nouveau possibles.
   (a) philosophe[0] s'approprie baguette[2]. Comme c'est le second mutex nécessaire à ce philosophe, il peut manger. Pendant ce temps, philosophe[2] est bloqué en attente de baguette[2]. Lorsque philosophe[0] aura terminé de manger, il libèrera les mutex baguette[2] et baguette[0], ce qui permettra à philosophe[2] ou philosophe[1] de manger.
   (b) philosophe[2] s'approprie baguette[2]. Comme c'est le second mutex nécessaire à ce philosophe, il peut manger. Pendant ce temps, philosophe[0] est bloqué en attente de baguette[2]. Lorsque philosophe[2] aura terminé de manger, il libèrera les mutex baguette[2] et baguette[1], ce qui permettra à philosophe[0] ou philosophe[1] de manger.
2. philosophe[1] s'approprie baguette[0]. Le même raisonnement que ci-dessus peut être suivi pour montrer qu'il n'y a pas de deadlock non plus dans ce cas.

La solution présentée empêche tout deadlock puisqu'à tout moment il n'y a au moins un philosophe qui peut manger. Malheureusement, il est possible avec cette solution qu'un philosophe mange alors que chacun des autres philosophes a pu s'approprier une baguette. Lorsque le nombre de philosophes est élevé (imaginez un congrès avec une centaine de philosophes), cela peut être une source importante d'inefficacité au niveau des performances.

Une implémentation possible de l'ordre présenté dans la table ci-dessus est reprise dans le programme /Threads/S6-src/pthread-philo2.c qui comprend la fonction philosophe suivante.

```
void* philosophe ( void* arg )
{
  int *id=(int *) arg;
  int left = *id;
  int right = (left + 1) % PHILOSOPHES;
  while(true) {
```

```
    // philosophe pense
    if(left<right) {
      pthread_mutex_lock(&baguette[left]);
      pthread_mutex_lock(&baguette[right]);
    }
    else {
      pthread_mutex_lock(&baguette[right]);
      pthread_mutex_lock(&baguette[left]);
    }
    mange(*id);
    pthread_mutex_unlock(&baguette[left]);
    pthread_mutex_unlock(&baguette[right]);
  }
  return (NULL);
}
```

Ce problème des philosophes est à l'origine d'une règle de bonne pratique essentielle pour tout programme découpé en threads dans lequel certains threads doivent acquérir plusieurs mutex. Pour éviter les deadlocks, il est nécessaire d'ordonnancer tous les mutex utilisés par le programme dans un ordre strict. Lorsqu'un thread doit réserver plusieurs mutex en même temps, il doit *toujours* effectuer ses appels à pthread_mutex_lock(3posix) dans l'ordre choisi pour les mutex. Si cet ordre n'est pas respecté par un des threads, un deadlock peut se produire.

6.4 Les sémaphores

Le problème de la coordination entre threads est un problème majeur. Outre les *mutex* que nous avons présenté, d'autres solutions à ce problème ont été développées. Historiquement, une des premières propositions de coordination sont les sémaphores [Dijkstra1965b]. Un *sémaphore* est une structure de données qui est maintenue par le système d'exploitation et contient :
– un entier qui stocke la valeur, positive ou nulle, du sémaphore.
– une queue qui contient les pointeurs vers les threads qui sont bloqués en attente sur ce sémaphore.
Tout comme pour les *mutex*, la queue associée à un sémaphore permet de bloquer les threads qui sont en attente d'une modification de la valeur du sémaphore.

Une implémentation des sémaphores se compose en général de quatre fonctions :
– une fonction d'initialisation qui permet de créer le sémaphore et de lui attribuer une valeur initiale nulle ou positive.
– une fonction permettant de détruire un sémaphore et de libérer les ressources qui lui sont associées.
– une fonction `post` qui est utilisée par les threads pour modifier la valeur du sémaphore. S'il n'y a pas de thread en attente dans la queue associée au sémaphore, sa valeur est incrémentée d'une unité. Sinon, un des threads en attente est libéré et passe à l'état *Ready*.
– une fonction `wait` qui est utilisée par les threads pour tester la valeur d'un sémaphore. Si la valeur du sémaphore est positive, elle est décrémentée d'une unité et la fonction réussit. Si le sémaphore a une valeur nulle, le thread est bloqué jusqu'à ce qu'un autre thread le débloque en appelant la fonction `post`.
Les sémaphores sont utilisés pour résoudre de nombreux problèmes de coordination [Downey2008]. Comme ils permettent de stocker une valeur entière, ils sont plus flexibles que les *mutex* qui sont utiles surtout pour les problèmes d'exclusion mutuelle.

6.4.1 Sémaphores POSIX

La librairie POSIX comprend une implémentation des sémaphores [9] qui expose plusieurs fonctions aux utilisateurs. La page de manuel sem_overview(7) présente de façon sommaire les fonctions de la librairie relatives aux sémaphores. Les quatre principales sont les suivantes :

9. Les systèmes Unix supportent également des sémaphores dits *System V* du nom de la version de Unix dans laquelle ils ont été introduits. Dans ces notes, nous nous focalisons sur les sémaphores POSIX qui ont une API un peu plus simple que les es sémaphores *System V*. Les principales fonctions pour les sémaphores *System V* sont semget(3posix), semctl(3posix) et semop(3posix).

```
#include <semaphore.h>

int sem_init(sem_t *sem, int pshared, unsigned int value);
int sem_destroy(sem_t *sem);
int sem_wait(sem_t *sem);
int sem_post(sem_t *sem);
```

Le fichier semaphore.h contient les différentes définitions de structures qui sont nécessaires au bon fonctionnement des sémaphores ainsi que les signatures des fonctions de cette API. Un sémaphore est représenté par une structure de données de type `sem_t`. Toutes les fonctions de manipulation des sémaphores prennent comme argument un pointeur vers le sémaphore concerné.

Pour pouvoir utiliser un sémaphore, il faut d'abord l'initialiser. Cela se fait en utilisant la fonction sem_init(3) qui prend comme argument un pointeur vers le sémaphore à initialiser. Nous n'utiliserons pas le second argument dans ce chapitre. Le troisième argument est la valeur initiale, positive ou nulle, du sémaphore.

La fonction sem_destroy(3) permet de libérer un sémaphore qui a été initialisé avec sem_init(3). Les sémaphores consomment des ressources qui peuvent être limitées dans certains environnements. Il est important de détruire proprement les sémaphores dès qu'ils ne sont plus nécessaires.

Les deux principales fonctions de manipulation des sémaphores sont sem_wait(3) et sem_post(3). Certains auteurs utilisent `down` ou `P` à la place de sem_wait(3) et `up` ou `V` à la place de sem_post(3) [Downey2008]. Schématiquement, l'opération `sem_wait` peut s'implémenter en utilisant le pseudo-code suivant :

```
int sem_wait(semaphore *s)
{
 s->val=s->val-1;
 if(s->val<0)
 {
  // Place this thread in s.queue;
  // This thread is blocked;
 }
}
```

La fonction sem_post(3) quant à elle peut schématiquement s'implémenter comme suit :

```
int sem_post(semaphore *s)
{
  s->val=s->val+1;
  if(s.val<=0)
  {
    // Remove one thread(T) from s.queue;
    // Mark Thread(T) as ready to run;
  }
}
```

Ces deux opérations sont bien entendu des opérations qui ne peuvent s'exécuter simultanément. Leur implémentation réelle comprend des sections critiques qui doivent être construites avec soin. Le pseudo-code ci-dessus ignore ces sections critiques. Des détails complémentaires sur l'implémentation des sémaphores peuvent être obtenus dans livre sur les systèmes d'exploitation [Stallings2011] [Tanenbaum+2009].

La meilleure façon de comprendre leur utilisation est d'analyser des problèmes classiques de coordination qui peuvent être résolus en utilisant des sémaphores.

6.4.2 Exclusion mutuelle

Les sémaphores permettent de résoudre de nombreux problèmes classiques. Le premier est celui de l'exclusion mutuelle. Lorsqu'il est initialisé à `1`, un sémaphore peut être utilisé de la même façon qu'un *mutex*. En utilisant des sémaphores, une exclusion mutuelle peut être protégée comme suit :

```
#include <semaphore.h>

//...

sem_t semaphore;

sem_init(&semaphore, 0, 1);

sem_wait(&semaphore);
// section critique
sem_post(&semaphore);

sem_destroy(&semaphore);
```

Les sémaphores peuvent être utilisés pour d'autres types de synchronisation. Par exemple, considérons une application découpée en threads dans laquelle la fonction `after` ne peut jamais être exécutée avant la fin de l'exécution de la fonction `before`. Ce problème de coordination peut facilement être résolu en utilisant un sémaphore qui est initialisé à la valeur `0`. La fonction `after` doit démarrer par un appel à sem_wait(3) sur ce sémaphore tandis que la fonction `before` doit se terminer par un appel à la fonction sem_post(3) sur ce sémaphore. De cette façon, si le thread qui exécute la fonction `after` est trop rapide, il sera bloqué sur l'appel à sem_wait(3). S'il arrive à cette fonction après la fin de la fonction `before` dans l'autre thread, il pourra passer sans être bloqué. Le programme ci-dessous illustre cette utilisation des sémaphores POSIX.

```
#define NTHREADS 2
sem_t semaphore;

void *before(void * param) {
  // do something
  for(int j=0;j<1000000;j++) {
  }
  sem_post(&semaphore);
  return(NULL);
}

void *after(void * param) {
  sem_wait(&semaphore);
  // do something
  for(int j=0;j<1000000;j++) {
  }
  return(NULL);
}

int main (int argc, char *argv[])  {
  pthread_t thread[NTHREADS];
  void * (* func[])(void *)={before, after};
  int err;

  err=sem_init(&semaphore, 0,0);
  if(err!=0) {
      error(err,"sem_init");
  }
  for(int i=0;i<NTHREADS;i++) {
    err=pthread_create(&(thread[i]),NULL,func[i],NULL);
    if(err!=0) {
      error(err,"pthread_create");
    }
  }

  for(int i=0;i<NTHREADS;i++) {
    err=pthread_join(thread[i],NULL);
    if(err!=0) {
      error(err,"pthread_join");
```

```
    }
  }
  sem_destroy(&semaphore);
  if(err!=0) {
      error(err,"sem_destroy");
  }
  return(EXIT_SUCCESS);
}
```

Si un sémaphore initialisé à la valeur `1` est généralement utilisé comme un *mutex*, il y a une différence importante entre les implémentations des sémaphores et des *mutex*. Un sémaphore est conçu pour être manipulé par différents threads et il est fort possible qu'un thread exécute sem_wait(3) et qu'un autre exécute sem_post(3). Pour les mutex, certaines implémentations supposent que le même thread exécute pthread_mutex_lock(3posix) et pthread_mutex_unlock(3posix). Lorsque ces opérations doivent être effectuées dans des threads différents, il est préférable d'utiliser des sémaphores à la place de mutex.

6.4.3 Problème du rendez-vous

Le problème du rendez-vous [Downey2008] est un problème assez courant dans les applications multithreadées. Considérons une application découpée en *N* threads. Chacun de ces threads travaille en deux phases. Durant la première phase, tous les threads sont indépendants et peuvent s'exécuter simultanément. Cependant, un thread ne peut démarrer sa seconde phase que si tous les *N* threads ont terminé leur première phase. L'organisation de chaque thread est donc :

```
premiere_phase();
// rendez-vous
seconde_phase();
```

Chaque thread doit pouvoir être bloqué à la fin de la première phase en attendant que tous les autres threads aient fini d'exécuter leur première phase. Cela peut s'implémenter en utilisant un mutex et un sémaphore.

```
sem_t rendezvous;
pthread_mutex_t mutex;
int count=0;

sem_init(&rendezvous,0,0);
```

La variable `count` permet de compter le nombre de threads qui ont atteint le point de rendez-vous. Le `mutex` protège les accès à la variable `count` qui est partagée entre les différents threads. Le sémaphore `rendezvous` est initialisé à la valeur `0`. Le rendez-vous se fera en bloquant les threads sur le sémaphore `rendezvous` tant que les `N` threads ne sont pas arrivés à cet endroit.

```
premiere_phase();

 // section critique
 pthread_mutex_lock(&mutex);
 count++;
  if(count==N) {
    // tous les threads sont arrivés
    sem_post(&rendezvous);
    }
 pthread_mutex_unlock(&mutex);
// attente à la barrière
sem_wait(&rendezvous);
// libération d'un autre thread en attente
sem_post(&rendezvous);

seconde_phase();
```

Le pseudo-code ci-dessus présente une solution permettant de résoudre ce problème du rendez-vous. Le sémaphore étant initialisé à `0`, le premier thread qui aura terminé la première phase sera bloqué sur

`sem_wait(&rendezvous);`. Les `N-1` premiers threads qui auront terminé leur première phase seront tous bloqués à cet endroit. Lorsque le dernier thread finira sa première phase, il incrémentera `count` puis exécutera `sem_post(&rendezvous);` ce qui libèrera un premier thread. Le dernier thread sera ensuite bloqué sur `sem_wait(&rendezvous);` mais il ne restera pas bloqué longtemps car chaque fois qu'un thread parvient à passer `sem_wait(&rendezvous);`, il exécute immédiatement `sem_post(&rendezvous);` ce qui permet de libérer un autre thread en cascade.

Cette solution permet de résoudre le problème du rendez-vous avec un nombre fixe de threads. Certaines implémentations de la librairie des threads POSIX comprennent une barrière qui peut s'utiliser de la même façon que la solution ci-dessus. Une barrière est une structure de données de type `pthread_barrier_t`. Elle s'initialise en utilisant la fonction pthread_barrier_init(3posix) qui prend trois arguments : un pointeur vers une barrière, des attributs optionnels et le nombre de threads qui doivent avoir atteint la barrière pour que celle-ci s'ouvre. La fonction pthread_barrier_destroy(3posix) permet de détruire une barrière. Enfin, la fonction pthread_barrier_wait(3posix) qui prend comme argument un pointeur vers une barrière bloque le thread correspondant à celle-ci tant que le nombre de threads requis pour passer la barrière n'a pas été atteint.

6.4.4 Problème des producteurs-consommateurs

Le problème des producteurs-consommateurs est un problème extrêmement fréquent et important dans les applications découpées en plusieurs threads. Il est courant de structurer une telle application, notamment si elle réalise de longs calculs, en deux types de threads :

– les *producteurs* : Ce sont des threads qui produisent des données et placent le résultat de leurs calculs dans une zone mémoire accessible aux consommateurs.
– les *consommateurs* : Ce sont des threads qui utilisent les valeurs calculées par les producteurs.

Ces deux types de threads communiquent en utilisant un buffer qui a une capacité limitée à N places comme illustré dans la figure ci-dessous.

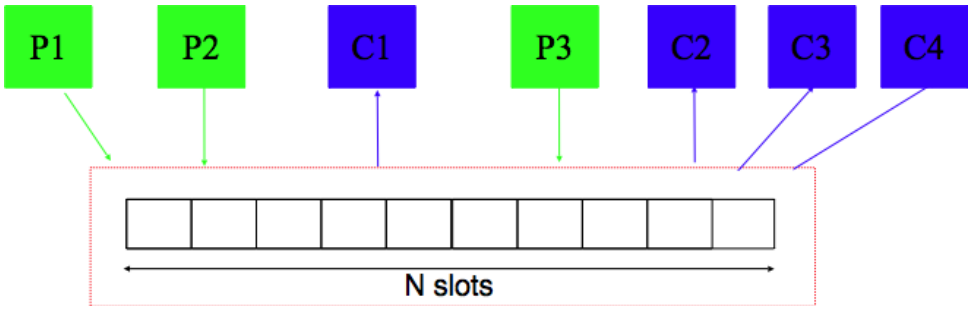


FIGURE 6.7 – Problème des producteurs-consommateurs

La difficulté du problème est de trouver une solution qui permet aux producteurs et aux consommateurs d'avancer à leur rythme sans que les producteurs ne bloquent inutilement les consommateurs et inversement. Le nombre de producteurs et de consommateurs ne doit pas nécessairement être connu à l'avance et ne doit pas être fixe. Un producteur peut arrêter de produire à n'importe quel moment.

Le buffer étant partagé entre les producteurs et les consommateurs, il doit nécessairement être protégé par un *mutex*. Les producteurs doivent pouvoir ajouter de l'information dans le buffer partagé tant qu'il y a au moins une place de libre dans le buffer. Un producteur ne doit être bloqué que si tout le buffer est rempli. Inversement, les consommateurs doivent être bloqués uniquement si le buffer est entièrement vide. Dès qu'une donnée est ajoutée dans le buffer, un consommateur doit être réveillé pour traiter cette donnée.

Ce problème peut être résolu en utilisant deux sémaphores et un mutex. L'accès au buffer, que ce soit par les consommateurs ou les producteurs est une section critique. Cet accès doit donc être protégé par l'utilisation d'un mutex. Quant aux sémaphores, le premier, baptisé `empty` dans l'exemple ci-dessous, sert à compter le nombre de places qui sont vides dans le buffer partagé. Ce sémaphore doit être initialisé à la taille du buffer puisque celui-ci est initialement vide. Le second sémaphore est baptisé `full` dans le pseudo-code ci-dessous. Sa valeur représente le nombre de places du buffer qui sont occupées. Il doit être initialisé à la valeur `0`.

```
// Initialisation
#define N 10 // places dans le buffer
pthread_mutex_t mutex;
sem_t empty;
sem_t full;

pthread_mutex_init(&mutex, NULL);
sem_init(&empty, 0 , N);  // buffer vide
sem_init(&full, 0 , 0);   // buffer vide
```

Le fonctionnement général d'un producteur est le suivant. Tout d'abord, le producteur est mis en attente sur le sémaphore `empty`. Il ne pourra passer que si il y a au moins une place du buffer qui est vide. Lorsque la ligne `sem_wait(&empty);` réussit, le producteur s'approprie le `mutex` et modifie le buffer de façon à insérer l'élément produit (dans ce cas un entier). Il libère ensuite le `mutex` pour sortir de sa section critique.

```
// Producteur
void producer(void)
{
  int item;
  while(true)
  {
    item=produce(item);
    sem_wait(&empty); // attente d'une place libre
    pthread_mutex_lock(&mutex);
     // section critique
     insert_item();
    pthread_mutex_unlock(&mutex);
    sem_post(&full); // il y a une place remplie en plus
  }
}
```

Le consommateur quant à lui essaie d'abord de prendre le sémaphore `full`. Si celui-ci est positif, cela indique la présence d'au moins un élément dans le buffer partagé. Ensuite, il entre dans la section critique protégée par le `mutex` et récupère la donnée se trouvant dans le buffer. Puis, il incrémente la valeur du sémaphore `empty` de façon à indiquer à un producteur qu'une nouvelle place est disponible dans le buffer.

```
// Consommateur
void consumer(void)
{
 int item;
 while(true)
 {
   sem_wait(&full); // attente d'une place remplie
   pthread_mutex_lock(&mutex);
    // section critique
    item=remove(item);
   pthread_mutex_unlock(&mutex);
   sem_post(&empty); // il y a une place libre en plus
 }
}
```

De nombreux programmes découpés en threads fonctionnent avec un ensemble de producteurs et un ensemble de consommateurs.

6.4.5 Problème des readers-writers

Le *problème des readers-writers* est un peu différent du précédent. Il permet de modéliser un problème qui survient lorsque des threads doivent accéder à une base de données [Courtois+1971]. Les threads sont généralement de deux types.

– les lecteurs (*readers*) sont des threads qui lisent une structure de données (ou une base de données) mais ne la modifient pas. Comme ces threads se contentent de lire de l'information en mémoire, rien ne s'oppose à ce que plusieurs *readers* s'exécutent simultanément.
– les écrivains (*writers*). Ce sont des threads qui modifient une structure de données (ou une base de données). Pendant qu'un *writer* manipule la structure de données, il ne peut y avoir aucun autre *writer* ni de *reader* qui accède à cette structure de données. Sinon, la concurrence des opérations de lecture et d'écriture donnerait un résultat incorrect.

Une première solution à ce problème est d'utiliser un mutex et un sémaphore [Courtois+1971].

```
pthread_mutex_t mutex;
sem_t db;  // accès à la db
int readcount=0; // nombre de readers

sem_init(&db, NULL, 1).
```

La solution utilise une variable partagée : `readcount`. L'accès à cette variable est protégé par `mutex`. Le sémaphore `db` sert à réguler l'accès des *writers* à la base de données. Le mutex est initialisé comme d'habitude par la fonction pthread_mutex_init(3posix). Le sémaphore `db` est initialisé à la valeur `1`. Le *writer* est assez simple :

```
void writer(void)
{
 while(true)
 {
   prepare_data();
   sem_wait(&db);
     // section critique, un seul writer à la fois
     write_database();
   sem_post(&db);
   }
 }
}
```

Le sémaphore `db` sert à assurer l'exclusion mutuelle entre les *writers* pour l'accès à la base de données. Le fonctionnement des *readers* est plus intéressant. Pour éviter un conflit entre les *writers* et les *readers*, il est nécessaire d'empêcher aux *readers* d'accéder à la base de données pendant qu'un *writer* la modifie. Cela peut se faire en utilisant l'entier *readcount* qui permet de compter le nombre de *readers* qui manipulent la base de données. Cette variable est testée et modifiée par tous les *readers*, elle doit donc être protégée par un *mutex*. Intuitivement, lorsque le premier *reader* veut accéder à la base de données (`readcount==0`), il essaye de décrémenter le sémaphore `db`. Si ce sémaphore est libre, le *reader* accède à la base de données. Sinon, il bloque sur `sem_wait(&db);` et comme il possède `mutex`, tous les autres *readers* sont bloqués sur `pthread_mutex_lock(&mutex);`. Dès que le premier *reader* est débloqué, il autorise en cascade l'accès à tous les autres *readers* qui sont en attente en libérant `pthread_mutex_unlock(&mutex);`. Lorsqu'un *reader* arrête d'utiliser la base de données, il vérifie s'il était le dernier *reader*. Si c'est le cas, il libère le sémaphore `db` de façon à permettre à un *writer* d'y accéder. Sinon, il décrémente simplement la variable `readcount` pour tenir à jour le nombre de *readers* qui sont actuellement en train d'accéder à la base de données.

```
void reader(void)
{
 while(true)
 {
   pthread_mutex_lock(&mutex);
     // section critique
     readcount++;
     if (readcount==1)
     { // arrivée du premier reader
       sem_wait(&db);
     }
   pthread_mutex_unlock(&mutex);
   read_database();
   pthread_mutex_lock(&mutex);
```

```
      // section critique
      readcount--;
      if(readcount==0)
      { // départ du dernier reader
        sem_post(&db);
      }
    pthread_mutex_unlock(&mutex);
    process_data();
  }
}
```

Cette solution fonctionne et garantit qu'il n'y aura jamais qu'un seul *writer* qui accède à la base de données. Malheureusement, elle souffre d'un inconvénient majeur lorsqu'il y a de nombreux *readers*. Dans ce cas, il est tout à fait possible qu'il y ait en permanence des *readers* qui accèdent à la base de données et que les *writers* soient toujours empêchés d'y accéder. En effet, dès que le premier *reader* a effectué `sem_wait(&db);`, aucun autre *reader* ne devra exécuter cette opération tant qu'il restera au moins un *reader* accédant à la base de données. Les *writers* par contre resteront bloqués sur l'exécution de `sem_wait(&db);`.

En utilisant des sémaphores à la place des *mutex*, il est possible de contourner ce problème. Cependant, cela nécessite d'utiliser plusieurs sémaphores. Intuitivement, l'idée de la solution est de donner priorité aux *writers* par rapport aux *readers*. Dès qu'un *writer* est prêt à accéder à la base de données, il faut empêcher de nouveaux *readers* d'y accéder tout en permettant aux *readers* présents de terminer leur lecture. Vous construirez ce mécanisme en séance de travaux dirigés.

Note : Read-Write locks

Certaines implémentations de la librairie des threads POSIX contiennent des *Read-Write locks*. Ceux-ci constituent une API de plus haut niveau qui s'appuie sur des sémaphores pour résoudre le *problème des readers-writers*. Les fonctions de création et de suppression de ces locks sont : pthread_rwlock_init(3posix), pthread_rwlock_destroy(3posix). Les fonctions pthread_rwlock_rdlock(3posix) et pthread_rwlock_unlock(3posix) sont réservées aux readers tandis que les fonctions pthread_rwlock_wrlock(3posix) et pthread_rwlock_unlock(3posix) sont utilisables par les writers. Des exemples d'utilisation de ces *Read-Write locks* peuvent être obtenus dans [Gove2011].

6.5 Compléments sur les threads POSIX

Il existe différentes implémentations des threads POSIX. Les mécanismes de coordination utilisables varient parfois d'une implémentation à l'autre. Dans les sections précédentes, nous nous sommes focalisés sur les fonctions principales qui sont en général bien implémentées. Une discussion plus détaillée des fonctions implémentées sous Linux peut se trouver dans [Kerrisk2010]. [Gove2011] présente de façon détaillée les mécanismes de coordination utilisables sous Linux, Windows et Oracle Solaris. [StevensRago2008] comprend également une description des threads POSIX mais présente des exemples sur des versions plus anciennes de Linux, FreeBSD, Solaris et MacOS.

Il reste cependant quelques concepts qu'il est utile de connaître lorsque l'on développe des programmes découpés en threads en langage C.

6.5.1 Variables `volatile`

Normalement, dans un programme C, lorsqu'une variable est définie, ses accès sont contrôlés entièrement par le compilateur. Si la variable est utilisée dans plusieurs calculs successifs, il peut être utile d'un point de vue des performances de stocker la valeur de cette variable dans un registre pendant au moins le temps correspondant à l'exécution de quelques instructions [10]. Cette optimisation peut éventuellement poser des difficultés dans certains

10. Les premiers compilateurs C permettaient au programmeur de donner des indications au compilateur en faisant précéder les déclarations de certaines variables avec le qualificatif `register` [KernighanRitchie1998]. Ce qualificatif indiquait que la variable était utilisée fréquemment et que le compilateur devrait en placer le contenu dans un registre. Les compilateurs actuels sont nettement plus performants et ils sont capables de détecter quelles sont les variables qu'il faut placer dans un registre. Il est inutile de chercher à influencer le compilateur en utilisant le qualificatif `register`. Les compilateurs actuels, dont gcc(1) supportent de nombreuses options permettant d'optimiser les performances

programmes utilisant des threads puisqu'une variable peut être potentiellement modifiée ou lue par plusieurs threads simultanément.

Les premiers compilateurs C avaient pris en compte un problème similaire. Lorsqu'un programme ou un système d'exploitation interagit avec des dispositifs d'entrée-sortie, cela se fait parfois en permettant au dispositif d'écrire directement en mémoire à une adresse connue par le système d'exploitation. La valeur présente à cette adresse peut donc être modifiée par le dispositif d'entrée-sortie sans que le programme ne soit responsable de cette modification. Face à ce problème, les inventeurs du langage C ont introduit le qualificatif `volatile`. Lorsqu'une variable est `volatile`, cela indique au compilateur qu'il doit recharger la variable de la mémoire chaque fois qu'elle est utilisée.

Pour bien comprendre l'impact de ce qualificatif, il est intéressant d'analyser le code assembleur généré par un compilateur C dans l'exemple suivant.

```
int x=1;
int v[2];

void f(void ) {
  v[0]=x;
  v[1]=x;
}
```

Dans ce cas, la fonction `f` est traduite en la séquence d'instructions suivante :

```
f:
      movl    x, %eax
      movl    %eax, v
      movl    %eax, v+4
      ret
```

Si par contre la variable `x` est déclarée comme étant `volatile`, le compilateur ajoute une instruction `movl x, %eax` qui permet de recharger la valeur de `x` dans un registre avant la seconde utilisation.

```
f:
      movl    x, %eax
      movl    %eax, v
      movl    x, %eax
      movl    %eax, v+4
      ret
```

Le qualificatif `volatile` force le compilateur à recharger la variable depuis la mémoire avant chaque utilisation. Ce qualificatif est utile lorsque le contenu stocké à une adresse mémoire peut être modifié par une autre source que le programme lui-même. C'est le cas dans les threads, mais marquer les variables partagées par des threads comme `volatile` ne suffit pas. Si ces variables sont modifiées par certains threads, il est nécessaire d'utiliser des *mutex* ou d'autres techniques de coordination pour réguler l'accès en ces variables partagées. En pratique, la documentation du programme devra spécifier quelles variables sont partagées entre les threads et la technique de coordination éventuelle qui est utilisée pour en réguler les accès. L'utilisation du qualificatif `volatile` permet de forcer le compilateur à recharger le contenu de la variable depuis la mémoire avant toute utilisation. C'est une règle de bonne pratique qu'il est utile de suivre. Il faut cependant noter que dans l'exemple ci-dessus, l'utilisation du qualificatif `volatile` augmente le nombre d'accès à la mémoire et peut donc dans certains cas réduire les performances.

6.5.2 Variables spécifiques à un thread

Dans un programme C séquentiel, on doit souvent combiner les variables globales, les variables locales et les arguments de fonctions. Lorsque le programme est découpé en threads, les variables globales restent utilisables, mais il faut faire attention aux problèmes d'accès concurrent. En pratique, il est parfois utile de pouvoir disposer dans chaque thread de variables qui tout en étant accessibles depuis toutes les fonctions du thread ne sont pas accessibles aux autres threads. Il y a différentes solutions pour résoudre ce problème.

des programmes compilés. Certaines ont comme objectif d'accélérer l'exécution du programme, d'autres visent à réduire sa taille. Pour les programmes qui consomment beaucoup de temps CPU, il est utile d'activer l'optimisation du compilateur.

Une première solution serait d'utiliser une zone mémoire qui est spécifique au thread et d'y placer par exemple une structure contenant toutes les variables auxquelles on souhaite pouvoir accéder depuis toutes les fonctions du thread. Cette zone mémoire pourrait être créée avant l'appel à pthread_create(3) et un pointeur vers cette zone pourrait être passé comme argument à la fonction qui démarre le thread. Malheureusement l'argument qui est passé à cette fonction n'est pas équivalent à une variable globale et n'est pas accessible à toutes les fonctions du thread.

Une deuxième solution serait d'avoir un tableau global qui contiendrait des pointeurs vers des zones de mémoires qui ont été allouées pour chaque thread. Chaque thread pourrait alors accéder à ce tableau sur base de son identifiant. Cette solution pourrait fonctionner si le nombre de threads est fixe et que les identifiants de threads sont des entiers croissants. Malheureusement la librairie threads POSIX ne fournit pas de tels identifiants croissants. Officiellement, la fonction pthread_self(3) retourne un identifiant unique d'un thread qui a été créé. Malheureusement cet identifiant est de type `pthread_t` et ne peut pas être utilisé comme index dans un tableau. Sous Linux, l'appel système non-standard gettid(2) retourne l'identifiant du thread, mais il ne peut pas non plus être utilisé comme index dans un tableau.

Pour résoudre ce problème, deux solutions sont possibles. La première combine une extension au langage C qui est supportée par gcc(1) avec la librairie threads POSIX. Il s'agit du qualificatif `__thread` qui peut être utilisé avant une déclaration de variable. Lorsqu'il est utilisé dans la déclaration d'une variable globale, il indique au compilateur et à la libraire POSIX qu'une copie de cette variable doit être créée pour chaque thread. Cette variable est initialisée au démarrage du thread et est utilisable uniquement à l'intérieur de ce thread. Le programme ci-dessous illustre cette utilisation du qualificatif `__thread`.

```
#define LOOP 1000000
#define NTHREADS 4

__thread int count=0;
int global_count=0;

void *f( void* param) {
  for(int i=0;i<LOOP;i++) {
    count++;
    global_count=global_count-1;
  }
  printf("Valeurs : count=%d, global_count=%d\n",count, global_count);
  return(NULL);
}

int main (int argc, char *argv[])  {
  pthread_t threads[NTHREADS];
  int err;

  for(int i=0;i<NTHREADS;i++) {
    count=i;   // local au thread du programme principal
    err=pthread_create(&(threads[i]),NULL,&f,NULL);
    if(err!=0)
      error(err,"pthread_create");
  }

  for(int i=0;i<NTHREADS;i++) {
    err=pthread_join(threads[i],NULL);
    if(err!=0)
      error(err,"pthread_create");
  }

  return(EXIT_SUCCESS);
}
```

Lors de son exécution, ce programme affiche la sortie suivante sur *stdout*. Cette sortie illustre bien que les variables dont la déclaration est précédée du qualificatif `__thread` sont utilisables uniquement à l'intérieur d'un thread.

```
Valeurs : count=1000000, global_count=-870754
Valeurs : count=1000000, global_count=-880737
Valeurs : count=1000000, global_count=-916383
Valeurs : count=1000000, global_count=-923423
```

La seconde solution proposée par la librairie POSIX est plus complexe. Elle nécessite l'utilisation des fonctions pthread_key_create(3posix), pthread_setspecific(3posix), pthread_getspecific(3posix) et pthread_key_delete(3posix). Cette API est malheureusement plus difficile à utiliser que le qualificatif `__thread`, mais elle illustre ce qu'il se passe en pratique lorsque ce qualificatif est utilisé.

Pour avoir une variable accessible depuis toutes les fonctions d'un thread, il faut tout d'abord créer une clé qui identifie cette variable. Cette clé est de type `pthread_key_t` et c'est l'adresse de cette structure en mémoire qui est utilisée comme identifiant pour la variable spécifique à chaque thread. Cette clé ne doit être créée qu'une seule fois. Cela peut se faire dans le programme qui lance les threads ou alors dans le premier thread lancé en utilisant la fonction pthread_once(3posix). Une clé est créée grâce à la fonction pthread_key_create(3posix). Cette fonction prend deux arguments. Le premier est un pointeur vers une structure de type `pthread_key_t`. Le second est la fonction optionnelle à appeler lorsque le thread utilisant la clé se termine.

Il faut noter que la fonction pthread_key_create(3posix) associe en pratique le pointeur `NULL` à la clé qui a été créée dans chaque thread. Le thread qui veut utiliser la variable correspondant à cette clé doit réserver la zone mémoire correspondante. Cela se fait en général en utilisant malloc(3) puis en appelant la fonction pthread_setspecific(3posix). Celle-ci prend deux arguments. Le premier est une clé de type `pthread_key_t` qui a été préalablement créée. Le second est un pointeur (de type `void *`) vers la zone mémoire correspondant à la variable spécifique. Une fois que le lien entre la clé et le pointeur a été fait, la fonction pthread_getspecific(3posix) peut être utilisée pour récupérer le pointeur depuis n'importe quelle fonction du thread. L'implémentation des fonctions pthread_setspecific(3posix) et pthread_getspecific(3posix) garantit que chaque thread aura sa variable qui lui est propre.

L'exemple ci-dessous illustre l'utilisation de cette API. Elle est nettement plus lourde à utiliser que le qualificatif `__thread`. Dans ce code, chaque thread démarre par la fonction `f`. Celle-ci crée une variable spécifique de type `int` qui joue le même rôle que la variable `__thread int count;` dans l'exemple précédent. La fonction `g` qui est appelée sans argument peut accéder à la zone mémoire créée en appelant `pthread_getspecific(count)`. Elle peut ensuite exécuter ses calculs en utilisant le pointeur `count_ptr`. Avant de se terminer, la fonction `f` libère la zone mémoire qui avait été allouée par malloc(3). Une alternative à l'appel explicite à free(3) aurait été de passer `free` comme second argument à pthread_key_create(3posix) lors de la création de la clé `count`. En effet, ce second argument est la fonction à appeler à la fin du thread pour libérer la mémoire correspondant à cette clé.

```
#define LOOP 1000000
#define NTHREADS 4

pthread_key_t count;
int global_count=0;

void g(void ) {
  void * data=pthread_getspecific(count);
  if(data==NULL)
    error(-1,"pthread_getspecific");
  int *count_ptr=(int *)data;
  for(int i=0;i<LOOP;i++) {
    *count_ptr=*(count_ptr)+1;
    global_count=global_count-1;
  }
}

void *f( void* param) {
  int err;
  int *int_ptr=malloc(sizeof(int));
  *int_ptr=0;
  err=pthread_setspecific(count, (void *)int_ptr);
  if(err!=0)
    error(err,"pthread_setspecific");
```

```
  g();
  printf("Valeurs : count=%d, global_count=%d\n",*int_ptr, global_count);
  free(int_ptr);
  return(NULL);
}

int main (int argc, char *argv[])  {
  pthread_t threads[NTHREADS];
  int err;

  err=pthread_key_create(&(count),NULL);
  if(err!=0)
    error(err,"pthread_key_create");

  for(int i=0;i<NTHREADS;i++) {
    err=pthread_create(&(threads[i]),NULL,&f,NULL);
    if(err!=0)
      error(err,"pthread_create");
  }

  for(int i=0;i<NTHREADS;i++) {
    err=pthread_join(threads[i],NULL);
    if(err!=0)
      error(err,"pthread_create");
  }
  err=pthread_key_delete(count);
  if(err!=0)
    error(err,"pthread_key_delete");

  return(EXIT_SUCCESS);
}
```

En pratique, on préférera évidemment d'utiliser le qualificatif `__thread` plutôt que d'utiliser une API explicite lorsque c'est possible. Cependant, il ne faut pas oublier que lorsque ce qualificatif est utilisé, le compilateur doit introduire dans le programme du code permettant de faire le même genre d'opérations que les fonctions explicites de la librairie.

6.5.3 Fonctions `thread-safe`

Dans un programme séquentiel, il n'y a qu'un thread d'exécution et de nombreux programmeurs, y compris ceux qui ont développé la librairie standard, utilisent cette hypothèse lors de l'écriture de fonctions. Lorsqu'un programme est découpé en threads, chaque fonction peut être appelée par plusieurs threads simultanément. Cette exécution simultanée d'une fonction peut poser des difficultés notamment lorsque la fonction utilise des variables globales ou des variables statiques.

Pour comprendre le problème, il est intéressant de comparer plusieurs implémentations d'une fonction simple. Considérons le problème de déterminer l'élément maximum d'une structure de données contenant des entiers. Si la structure de données est un tableau, une solution simple est de le parcourir entièrement pour déterminer l'élément maximum. C'est ce que fait la fonction `max_vector` dans le programme ci-dessous. Dans un programme purement séquentiel dans lequel le tableau peut être modifié de temps en temps, parcourir tout le tableau pour déterminer son maximum n'est pas nécessairement la solution la plus efficace. Une alternative est de mettre à jour la valeur du maximum chaque fois qu'un élément du tableau est modifié. Les fonctions `max_global` et `max_static` sont deux solutions possibles. Chacune de ces fonctions doit être appelée chaque fois qu'un élément du tableau est modifié. `max_global` stocke dans une variable globale la valeur actuelle du maximum du tableau et met à jour cette valeur à chaque appel. La fonction `max_static` fait de même en utilisant une variable statique. Ces deux solutions sont équivalentes et elles pourraient très bien être intégrées à une librairie utilisée par de nombreux programmes.

```
#include <stdint.h>
#define SIZE 10000
```

```
int g_max=INT32_MIN;
int v[SIZE];

int max_vector(int n, int *v) {
  int max=INT32_MIN;
  for(int i=0;i<n;i++) {
    if(v[i]>max)
      max=v[i];
  }
  return max;
}

int max_global(int *v) {
    if (*v>g_max) {
      g_max=*v;
    }
    return(g_max);
}

int max_static(int *v){
  static int s_max=INT32_MIN;
  if (*v>s_max) {
    s_max=*v;
  }
  return(s_max);
}
```

Considérons maintenant un programme découpé en plusieurs threads qui chacun maintient un tableau d'entiers dont il faut connaître le maximum. Ces tableaux d'entiers sont distincts et ne sont pas partagés entre les threads. La fonction `max_vector` peut être utilisée par chaque thread pour déterminer le maximum du tableau. Par contre, les fonctions `max_global` et `max_static` ne peuvent pas être utilisées. En effet, chacune de ces fonctions maintient *un* état (dans ce cas le maximum calculé) alors qu'elle peut être appelée par différents threads qui auraient chacun besoin d'un état qui leur est propre. Pour que ces fonctions soient utilisables, il faudrait que les variables `s_max` et `g_max` soient spécifiques à chaque thread.

En pratique, ce problème de l'accès concurrent à des fonctions se pose pour de nombreuses fonctions et notamment celles de la librairie standard. Lorsque l'on développe une fonction qui peut être réutilisée, il est important de s'assurer que cette fonction peut être exécutée par plusieurs threads simultanément sans que cela ne pose de problèmes à l'exécution.

Ce problème affecte certaines fonctions de la librairie standard et plusieurs d'entre elles ont dû être modifiées pour pouvoir supporter les threads. A titre d'exemple, considérons la fonction strerror(3). Cette fonction prend comme argument le numéro de l'erreur et retourne une chaîne de caractères décrivant cette erreur. Cette fonction ne peut pas être utilisée telle quelle par des threads qui pourraient l'appeler simultanément. Pour s'en convaincre, regardons une version simplifiée d'une implémentation de cette fonction [11]. Cette fonction utilise le tableau sys_errlist(3) qui contient les messages d'erreur associés aux principaux codes numériques d'erreur. Lorsque l'erreur est une erreur standard, tout se passe bien et la fonction retourne simplement un pointeur vers l'entrée du tableau `sys_errlist` correspondante. Par contre, si le code d'erreur n'est pas connu, un message est généré dans le tableau `buf[32]` qui est déclaré de façon statique. Si plusieurs threads exécutent `strerror`, ce sera le même tableau qui sera utilisé dans les différents threads. On pourrait remplacer le tableau statique par une allocation de zone mémoire faite via malloc(3), mais alors la zone mémoire créée risque de ne jamais être libérée par free(3) car l'utilisateur de strerror(3) ne doit pas libérer le pointeur qu'il a reçu, ce qui pose d'autres problèmes en pratique.

```
char * strerror (int errnoval)
{
 char * msg;
 static char buf[32];
 if ((errnoval < 0) || (errnoval >= sys_nerr))
   { // Out of range, just return NULL
```

11. Cette implémentation est adaptée de http ://opensource.apple.com/source/gcc/gcc-926/libiberty/strerror.c et est dans le domaine public.

```
    msg = NULL;
  }
 else if ((sys_errlist == NULL) || (sys_errlist[errnoval] == NULL))
  { // In range, but no sys_errlist or no entry at this index.
    sprintf (buf, "Error %d", errnoval);
    msg = buf;
  }
 else
  { // In range, and a valid message.  Just return the message.
    msg = (char *) sys_errlist[errnoval];
  }
 return (msg);
}
```

La fonction strerror_r(3) évite ce problème de tableau statique en utilisant trois arguments : le code d'erreur, un pointeur `char *` vers la zone devant stocker le message d'erreur et la taille de cette zone. Cela permet à strerror_r(3) d'utiliser une zone mémoire qui lui est passée par le thread qu'il appelle et garantit que chaque thread disposera de son message d'erreur. Voici une implémentation possible de strerror_r(3) [12].

```
strerror_r(int num, char *buf, size_t buflen)
{
  #define UPREFIX "Unknown error: %u"
  unsigned int errnum = num;
  int retval = 0;
  size_t slen;
  if (errnum < (unsigned int) sys_nerr) {
     slen = strlcpy(buf, sys_errlist[errnum], buflen);
  } else {
     slen = snprintf(buf, buflen, UPREFIX, errnum);
     retval = EINVAL;
  }
  if (slen >= buflen)
       retval = ERANGE;
  return retval;
}
```

Lorsque l'on intègre des fonctions provenant de la librairie standard ou d'une autre librairie dans un programme découpé en threads, il est important de vérifier que les fonctions utilisées sont bien *thread-safe*. La page de manuel pthreads(7) liste les fonctions qui ne sont pas *thread-safe* dans la librairie standard.

6.6 Loi de Amdahl

En découpant un programme en threads, il est théoriquement possible d'améliorer les performances du programme en lui permettant d'exécuter plusieurs threads d'exécution simultanément. Dans certains cas, la découpe d'un programme en différents threads est naturelle et relativement facile à réaliser. Dans d'autres cas, elle est nettement plus compliquée. Pour s'en convaincre, il suffit de considérer un grand tableau contenant plusieurs centaines de millions de nombres. Considérons un programme simple qui doit trouver dans ce tableau quel est l'élément du tableau pour lequel l'application d'une fonction complexe `f` donne le résultat minimal. Une implémentation purement séquentielle se contenterait de parcourir l'entièreté du tableau et d'appliquer la fonction `f` à chacun des éléments. A la fin de son exécution, le programme retournera l'élément qui donne la valeur minimale. Un tel problème est très facile à découper en threads. Il suffit de découper le tableau en `N` sous-tableaux, de lancer un thread de calcul sur chaque sous-tableau et ensuite de fusionner les résultats de chaque thread.

Un autre problème est de trier le contenu d'un tel tableau dans l'ordre croissant. De nombreux algorithmes séquentiels de tri existent pour ordonner un tableau. La découpe de ce problème en thread est nettement moins évidente que dans le problème précédent et les algorithmes de tri adaptés à une utilisation dans plusieurs threads ne sont pas une simple extension des algorithmes séquentiels.

12. Cette implémentation est adaptée de https ://www-asim.lip6.fr/trac/netbsdtsar/browser/vendor/netbsd/5/src/lib/libc/string/strerror_r.c ?rev=2 et est *Copyright (c) 1988 Regents of the University of California.*

Dans les années 1960s, à l'époque des premières réflexions sur l'utilisation de plusieurs processeurs pour résoudre un problème, Gene Amdahl [Amdahl1967] a analysé quelles étaient les gains que l'on pouvait attendre de l'utilisation de plusieurs processeurs. Dans sa réflexion, il considère un programme `P` qui peut être découpé en deux parties :

– une partie purement séquentielle. Il s'agit par exemple de l'initialisation de l'algorithme utilisé, de la collecte des résultats, ...
– une partie qui est parallélisable. Il s'agit en général du coeur de l'algorithme.

Plus les opérations réalisées à l'intérieur d'un programme sont indépendantes entre elles, plus le programme est parallélisable et inversement. Pour Amdahl, si le temps d'exécution d'un programme séquentiel est T et qu'une fraction f de ce programme est parallélisable, alors le gain qui peut être obtenu de la parallélisation est $\frac{T}{T\times((1-f)+\frac{f}{N})} = \frac{1}{(1-f)+\frac{f}{N}}$ lorsque le programme est découpé en N threads. Cette formule, connue sous le nom de la *loi de Amdahl* fixe une limite théorique sur le gain que l'on peut obtenir en parallélisant un programme. La figure ci-dessous **[#famdahl]_** illustre le gain théorique que l'on peut obtenir en parallélisant un programme en fonction du nombre de processeur et pour différentes fractions parallélisables.

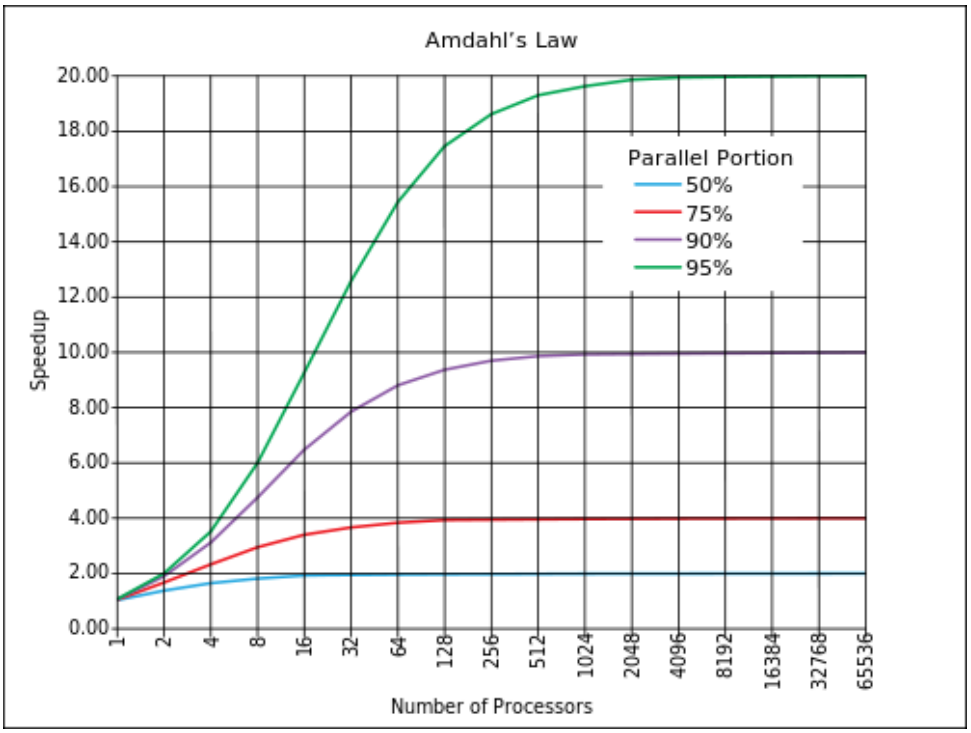


FIGURE 6.8 – Loi de Amdahl (source wikipedia)

La loi de Amdahl doit être considérée comme un maximum théorique qui est difficile d'atteindre. Elle suppose que la parallélisation est parfaite, c'est-à-dire que la création et la terminaison de threads n'ont pas de coût en terme de performance. En pratique, c'est loin d'être le cas et il peut être difficile d'estimer a priori le gain qu'une parallélisation permettra d'obtenir. En pratique, avant de découper un programme séquentiel en threads, il est important de bien identifier la partie séquentielle et la partie parallélisable du programme. Si la partie séquentielle est trop importante, le gain dû à la parallélisation risque d'être faible. Si par contre la partie purement séquentielle est faible, il est possible d'obtenir théoriquement des gains élevés. Le tout sera de trouver des solutions efficaces qui permettront aux threads de fonctionner le plus indépendamment possible.

En pratique, avant de s'attaquer à la découpe d'un programme séquentiel en threads, il est important de bien comprendre quelles sont les parties du programme qui sont les plus consommatrices de temps CPU. Ce seront souvent les boucles ou les grandes structures de données. Si le programme séquentiel existe, il est utile d'analyser son exécution avec des outils de profiling tels que gprof(1) [Graham+1982] ou oprofile. Un profiler est un logiciel qui permet d'analyser l'exécution d'un autre logiciel de façon à pouvoir déterminer notamment quelles sont les fonctions ou parties du programmes les plus exécutées. Ces parties de programme sont celles sur lesquelles l'effort de paraléllisation devra porter en pratique.

Dans un programme découpé en threads, toute utilisation de fonctions de coordination comme des sémaphores ou

des mutex, bien qu'elle soit nécessaire pour la correction du programme, risque d'avoir un impact négatif sur les performances. Pour s'en convaincre, il est intéressant de réfléchir au problème des producteurs-consommateurs. Il correspond à de nombreux programmes réels. Les performances d'une implémentation du problème des producteurs consommateurs dépendront fortement de la taille du buffer entre les producteurs et les consommateurs et de leur nombre et/ou vitesses relatives. Idéalement, il faudrait que le buffer soit en moyenne rempli à moitié. De cette façon, chaque producteur pourra déposer de l'information dans le buffer et chaque consommateur pourra en retirer. Si le buffer est souvent vide, cela indique que les consommateurs sont plus rapides que les producteurs. Ces consommateurs risquent d'être inutilement bloqués, ce qui affectera les performances. Il en va de même si le buffer était plein. Dans ce cas, les producteurs seraient souvent bloqués.

6.7 Les processus

Un système d'exploitation multitâche et multi-utilisateurs tel que Unix ou Linux permet d'exécuter de nombreux programmes simultanément. Sous Unix, les programmes sont exécutés sous la forme de *processus*. Un processus peut être défini comme étant une instance de programme qui est en train d'être exécutée sur un ou plusieurs processeurs sous le contrôle d'un système d'exploitation. Un processus comprend donc un ensemble d'instructions pour le processeur, mais aussi des données qui sont stockées en mémoire et un contexte (si le processus utilise un seul thread d'exécution, plusieurs contextes sinon). En outre, le système d'exploitation maintient un certain nombre de structures de données qui sont nécessaires au bon fonctionnement du processus. Ces structures de données sont créées au démarrage du processus, mises à jour durant la vie du processus et supprimées lorsque le processus se termine.

6.7.1 Les librairies

Lorsqu'un programme s'exécute à l'intérieur d'un processus, il exécute des instructions qui ont différentes *origines*. Il y a bien entendu les instructions qui proviennent du code source du programme qui a été converti en assembleur par le compilateur. Ces instructions correspondent au code source développé par le programmeur. Il s'agit notamment de toutes les opérations mathématiques et logiques, les boucles et les appels de fonctions internes au programme. Comme nous l'avons vu précédemment, ces instructions peuvent provenir d'un seul module ou de plusieurs modules. Dans ce dernier cas, le linker intervient pour combiner différents modules en un exécutable complet.

A côté des instructions qui correspondent aux lignes de code écrites par le développeur du programme, un processus va également exécuter de nombreuses fonctions qui font partie d'une des librairies standard du système. Tout environnement de développement comprend des librairies qui permettent de faciliter le travail des programmeurs en leur fournissant des fonctions permettant de résoudre de nombreux problèmes classiques. Un système d'exploitation tel que Unix ou Linux contient de nombreuses librairies de ce type. Nous avons déjà eu l'occasion de discuter des fonctions provenant de la librairie standard comme printf(3) ou malloc(3) et celles de la librairie pthreads(7). Ce ne sont que deux librairies parmi d'autres. Un système Linux contient plusieurs centaines de librairies utilisables par le programmeur.

A titre d'exemple, considérons la librairie math.h(7posix). Cette librairie contient de nombreuses fonctions mathématiques. Pour les utiliser dans un programme, il faut non seulement y inclure le fichier header math.h qui contient les prototypes et constantes utilisées par la librairie, mais aussi indiquer au linker que l'exécutable doit être lié avec la librairie math.h(7posix). Cela se fait en utilisant le flag `-l` de gcc(1).

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>


int main (int argc, char *argv[])  {
  double n1=1.0;
  double n2=-3.14;
  printf("Maximum : %f\n",fmax(n1,n2));

  return(EXIT_SUCCESS);
}
```

Le programme `/Threads/S8-src/math.c` ci-dessus doit être compilé en utilisant la commande `gcc -Wall -Werror math.c -o math -lm`. Le paramètre `-lm` indique au compilateur qu'il doit charger la librairie `m`. Cette librairie, est une des librairies standard du système, elle réside généralement dans le répertoire `/usr/lib`[13]. En pratique, gcc(1) charge automatiquement la librairie `C` standard lors de la compilation de tout programme. Cela revient à utiliser implicitement le paramètre `-lc`.

Lors de l'utilisation de telles librairies, on s'attendrait à ce que toutes les instructions correspondant aux fonctions de la librairie utilisée soient présentes à l'intérieur de l'exécutable. En pratique, ce n'est pas exactement le cas. Même si notre programme d'exemple utilise fmax(3) de la librairie math.h(7posix) et printf(3) de la librairie standard, son exécutable ne contient que quelques milliers d'octets.

```
$ ls -l math*
-rwxr-xr-x 1 obo stafinfo 6764 Mar 15  2012 math
-rw-r--r-- 1 obo stafinfo  373 Mar 15  2012 math.c
```

Une analyse plus détaillée de l'exécutable avec objdump(1) révèle que si l'exécutable contient bien des appels à ces fonctions, leur code n'y est pas entièrement inclus.

```
$gcc -g -lm math.c -o math
$objdump -S -d math
math:     file format elf64-x86-64
...
0000000000400468 <fmax@plt>:
400468:      ff 25 fa 04 20 00       jmpq   *0x2004fa(%rip)        # 600968 <
_GLOBAL_OFFSET_TABLE_+0x28>
40046e:      68 02 00 00 00          pushq  $0x2
400473:      e9 c0 ff ff ff          jmpq   400438 <_init+0x18>
...
 #include <stdio.h>
 #include <stdlib.h>
 #include <math.h>
 int main (int argc, char *argv[])  {
400564:      55                      push   %rbp
400565:      48 89 e5                mov    %rsp,%rbp
400568:      48 83 ec 20             sub    $0x20,%rsp
40056c:      89 7d ec                mov    %edi,-0x14(%rbp)
40056f:      48 89 75 e0             mov    %rsi,-0x20(%rbp)
 double n1=1.0;
400573:      48 b8 00 00 00 00 00    mov    $0x3ff0000000000000,%rax
40057a:      00 f0 3f
40057d:      48 89 45 f0             mov    %rax,-0x10(%rbp)
 double n2=-3.14;
400581:      48 b8 1f 85 eb 51 b8    mov    $0xc0091eb851eb851f,%rax
400588:      1e 09 c0
40058b:      48 89 45 f8             mov    %rax,-0x8(%rbp)
 printf("Maximum : %f\n",fmax(n1,n2));
40058f:      f2 0f 10 4d f8          movsd  -0x8(%rbp),%xmm1
400594:      f2 0f 10 45 f0          movsd  -0x10(%rbp),%xmm0
400599:      e8 ca fe ff ff          callq  400468 <fmax@plt>
40059e:      b8 b8 06 40 00          mov    $0x4006b8,%eax
4005a3:      48 89 c7                mov    %rax,%rdi
4005a6:      b8 01 00 00 00          mov    $0x1,%eax
4005ab:      e8 98 fe ff ff          callq  400448 <printf@plt>
 return(EXIT_SUCCESS);
4005b0:      b8 00 00 00 00          mov    $0x0,%eax
 }
```

La taille réduite des exécutables sous Linux et de nombreuses variantes de Unix s'explique par l'utilisation de librairies partagées. Un programme peut utiliser deux types de librairies : des librairies statiques et des librairies

13. Par défaut, gcc(1) cherche après les librairies spécifiées dans les répertoires de librairies standards, mais aussi dans les répertoires listés dans la variable d'environnement `LD_LIBRARY_PATH`. Il est également possible de spécifier des répertoires supplémentaires contenant les librairies avec l'argument `-L` de gcc(1).

partagées. Une *librairie statique* (ou *static library* en anglais) est une librairie de fonctions qui est intégrée directement avec le programme. Elle fait entièrement partie de l'exécutable. C'est la première solution pour intégrer des librairies dans un programme. Son avantage principal est que l'exécutable est complet et comprend toutes les instructions qui sont nécessaires au fonctionnement du programme. Malheureusement, tous les programmes qui utilisent des fonctions d'une librairie courante, comme par exemple la librairie standard, doivent inclure le code relatif à toutes les fonctions qu'ils utilisent. Sachant que chaque programme ou presque utilise des fonctions comme printf(3), cela conduit à sauvegarder de très nombreuses copies du même code. Ce problème peut être résolu en utilisant des librairies partagées [14] . Une *librairie partagée* (ou *shared library* en anglais) est un ensemble de fonctions qui peuvent être appelées par un programme mais sont stockées dans un seul fichier sur disque. Ce fichier unique est utilisé automatiquement par tous les programmes qui utilisent des fonctions de la librairie.

Il est parfois intéressant de pouvoir créer une librairie qui peut être liée de façon statique avec des programmes, par exemple lorsque ceux-ci doivent être exécutés sur d'autres ordinateurs que ceux sur lesquels ils ont été compilés. A titre d'illustration, considérons une librairie minuscule contenant une seule fonction `imax` qui calcule le maximum entre deux entiers. L'implémentation de cette fonction est très simple.

```
int imax(int i, int j) {
  return ((i>j) ? i : j);

}
```

Cette fonction est déclarée dans le fichier header `imax.h` et peut être utilisée dans un programme comme ci-dessous.

```
#include <stdio.h>
#include <stdlib.h>
#include "imax.h"

int main (int argc, char *argv[])  {
  int n1=1;
  int n2=-3;
  printf("Maximum : %d\n",imax(n1,n2));

  return(EXIT_SUCCESS);
}
```

En pratique, la construction d'une librairie se fait en deux étapes principales. Tout d'abord, il faut compiler les fichiers objet correspondant aux différents modules de la librairie. Cela peut se faire avec gcc(1) comme pour un programme C classique. Ensuite, il faut regrouper les différents modules dans une archive qui constituera la librairie qui peut être utilisée par des programmes. Par convention, toutes les librairies ont un nom qui commence par `lib` et se termine par l'extension `.a`. Sous Linux, cette opération est réalisée par l'utilitaire ar(1). La page de manuel de ar(1) décrit plus en détails son utilisation. En pratique, les opérations les plus fréquentes avec ar(1) sont :
– ajout d'un module objet à une librairie : `ar r libname.a module.o`
– suppression d'un module objet d'une librairie : `ar d libname.a module.o`
Il est aussi possible de lister le contenu de la librairie `libname.a` avec la commande `ar tv libname.a`.

L'archive contenant la librairie peut être liée en utilisant le linker à n'importe quel programme qui en utilise une ou plusieurs fonctions. Le linker de gcc(1) peut effectuer cette opération comme illustré par le `Makefile` ci-dessous. Il faut noter que l'argument `--static` permet de forcer le compilateur à inclure le code de la librairie dans l'exécutable.

```
#
# Makefile for library imax and imath
#

GCC        = gcc
AR         = ar
ARFLAGS = -cvq
```

14. Dans certains cas, on parle également de librairies dynamiques car ces librairies sont chargées dynamiquement à l'exécution du programme.

```
CFLAGS = -Wall -std=c99 -g -c
LDFLAGS = --static -g

all: imath

imax.o: imax.c
        @echo compiling imax
        $(GCC) $(CFLAGS) imax.c

libimax.a: imax.o
        @echo building libimax
        $(AR) $(ARFLAGS) libimax.a imax.o

imath.o: imath.c imax.h
        @echo compiling imath.o
        $(GCC) $(CFLAGS) imath.c

imath: imath.o libimax.a
        @echo building imath
        $(GCC) $(LDFLAGS)  -o imath libimax.a imath.o

clean:
        rm imath libimax.a imax.o imath.o
```

Ce `Makefile` est un petit peu plus long que ceux que nous avons utilisés jusque maintenant. Il illustre une structure courante pour de nombreux fichiers `Makefile`. La première partie définit des constantes qui sont utilisées dans le reste du `Makefile`. Il s'agit tout d'abord du compilateur et du programme de construction de librairies qui sont utilisés. Définir ces programmes comme des constantes dans le `Makefile` permet de facilement en changer lorsque c'est nécessaire. Ensuite, trois constantes sont définies avec les arguments de base du compilateur et de `ar`. A nouveau, définir ces constantes une fois pour toutes facilite leur modification. Ensuite, la première cible est la cible `all:`. Comme c'est la première, c'est la cible par défaut qui sera utilisée lorsque make(1) est appelé sans argument. Elle dépend de l'exécutable `imath` qui est une des cibles du `Makefile`. La cible `clean:` permet d'effacer les fichiers objet et exécutables construites par le `Makefile`. Il est utile d'avoir une telle cible lorsque l'on doit diffuser un projet en C ou le rendre dans le cadre d'un cours. Enfin, les autres cibles correspondent aux fichiers objet, à la librairie et à l'exécutable qui sont construits. La commande `@echo` affiche ses arguments sur la sortie standard. Enfin, la chaîne de caractères `$(GCC)` est remplacée par le constante définie au début du fichier. Des compléments d'information sur make(1) peuvent être obtenus dans divers documents dont make(1), [Mecklenburg+2004] ou [GNUMake].

Lorsqu'un programme est compilé de façon à utiliser une librairie dynamique, c'est le système d'exploitation qui analyse le programme lors de son chargement et intègre automatiquement les fonctions des librairies qui sont nécessaires à son exécution. L'entête de l'exécutable contient de l'information générée par le compilateur qui permet de localiser les librairies dynamiques qui doivent être intégrées de cette façon. L'utilitaire ldd(1) permet de visualiser quelles sont les librairies partagées utilisées par un programme.

```
$ ldd imath
linux-vdso.so.1 =>  (0x00007fffe41ff000)
libc.so.6 => /lib64/libc.so.6 (0x0000003eb2400000)
/lib64/ld-linux-x86-64.so.2 (0x0000003eb2000000)
```

6.7.2 Création d'un processus

Pour comprendre le fonctionnement de Unix, il est utile d'analyser plus en détails toutes les opérations qui sont effectuées à chaque fois que l'on lance un programme depuis un shell tel que bash(1). Considérons l'exécution de la commande `/bin/true` depuis le shell.

Schématiquement, l'exécution de ce programme se déroule comme suit. Le shell va d'abord localiser [15] l'exécutable `/bin/true` qui est stocké dans le système de fichiers. Ensuite, il va créer un processus et y exécuter

15. La variable d'environnement `PATH` contient la liste des répertoires que le shell parcoure afin de localiser un exécutable à lancer lorsque l'utilisateur ne fournit par le chemin complet de l'exécutable à lancer.

l'exécutable. Le shell va ensuite attendre la fin de l'exécution du programme `true` et récupérer sa valeur de retour (retournée par exit(2)) pour ensuite poursuivre son exécution.

Comme nous l'avons expliqué plus haut, le *kernel* Linux gère l'ensemble des processus qui sont utilisés à un moment. Il intervient pour toutes les opérations de création et de fin d'un processus. La création d'un processus est un événément important dans un système d'exploitation. Elle permet notamment l'exécution de programmes. Ces opérations nécessitent une interaction avec le *kernel* et se font donc en utilisant des appels systèmes. Avant d'analyser en détails comment Linux supporte précisément la création de processus, il est intéressant de réfléchir aux opérations qui doivent être effectuées lors de l'exécution d'un programme. Considérons par exemple un utilisateur qui exécute la commande `/usr/bin/expr 1 + 2` depuis un shell bash(1) interactif. Pour exécuter cette commande, il va falloir exécuter un nouveau processus contenant les instructions assembleur se trouvant dans l'exécutable `/usr/bin/expr`, lui passer les arguments `1 + 2`, l'exécuter, récupérer sa valeur de retour et la retourner au shell qui pourra l'utiliser et poursuivre son exécution.

Les designers de Unix ont choisi de construire un appel système pour chacune de ces opérations. Le premier est l'appel système fork(2). C'est l'appel système qui permet de créer un processus. Schématiquement, cet appel système crée une copie complète du processus qui l'a exécuté. Après exécution de fork(2), il y a deux copies du même processus en mémoire. Le processus qui a exécuté fork(2) est considéré comme étant le *processus père* tandis que celui qui a été créé par l'exécution de fork(2) est le *processus fils*.

```
#include <unistd.h>

pid_t fork(void);
```

L'appel système fork(2) est atypique car il est exécuté par un processus mais provoque la création d'un second processus qui est identique au premier. Après l'exécution de l'appel système fork(2), il y a donc deux séquences d'instructions qui vont s'exécuter, l'une dans le processus père et l'autre dans le processus fils. Le processus fils démarre son exécution à la récupération du résultat de l'appel système fork(2) effectué par son père. Le processus père et le processus fils récupèrent une valeur de retour différente pour cet appel système. Cette valeur de retour est d'ailleurs la seule façon de distinguer le *processus père* du *processus fils* lorsque celui-ci démarre.

– l'appel système fork(2) retourne la valeur `-1` en cas d'erreur et met à jour la variable `errno`. En cas d'erreur, aucun processus n'est créé.
– l'appel système fork(2) retourne la valeur `0` dans le processus fils.
– l'appel système fork(2) retourne une valeur positive dans le processus père. Cette valeur est l'identifiant du processus fils créé.

Pour bien comprendre le fonctionnement de fork(2), analysons l'exemple `/Threads/S8-src/fork.c` ci-dessous :

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>

int g=0; // segment données

int main (int argc, char *argv[])  {
  int l=1252; // sur la pile
  int *m;     // sur le heap
  m=(int *) malloc(sizeof(int));
  *m=-1;

  pid_t pid;

  pid=fork();

  if (pid==-1) {
    // erreur à l'exécution de fork
    perror("fork");
    exit(EXIT_FAILURE);
  }
  // pas d'erreur
  if (pid==0) {
    // processus fils
```

```
    l++;
    g++;
    *m=17;
    printf("Dans le processus fils g=%d, l=%d et *m=%d\n",g,l,*m);
    free(m);
    return(EXIT_SUCCESS);
  }
  else {
    // processus père
    sleep(2);
    printf("Dans le processus père g=%d, l=%d et *m=%d\n",g,l,*m);
    free(m);
    // ...
    return(EXIT_SUCCESS);
  }
}
```

Lors de son exécution, ce programme affiche les deux lignes suivantes sur sa sortie standard :

```
Dans le processus fils g=1, l=1253 et *m=17
Dans le processus père g=0, l=1252 et *m=-1
```

Lors de l'exécution de ce programme, deux variables sont initialisées en mémoire. La variable globale `g` est initialisée à la valeur `0` tandis que le variable locale `l` est initialisée à la valeur `1252`. malloc(3) est utilisé pour réserver une zone mémoire sur le *heap* et son contenu est initialisé à `-1`. Lorsque le processus père fait appel à fork(2) le noyau du système d'exploitation crée une copie identique à celui-ci en mémoire. Cette copie contient tous les segments du processus père (code, données, heap et stack) dans l'état exact dans lequel ils étaient au moment de l'exécution de l'appel système fork(2). Le contexte du processus père est copié et devient le contexte du processus fils. A cet instant, les deux processus sont complètement identiques à l'exception de certaines données qui sont maintenues par le système d'exploitation, comme l'identifiant de processus. Chaque processus qui s'exécute sur un système Unix a un identifiant unique et est retourné par l'appel système getpid(2). Le processus père et le processus fils ont un identifiant différent.

Les deux processus vont se différencier dès la fin de l'exécution de l'appel système fork(2). Comme tout appel système, fork(2) place sa valeur de retour dans le registre `%eax`. Comme indiqué plus haut, cette valeur sera positive dans le processus père. Celui-ci exécute `sleep(2);` et reste donc bloqué pendant deux secondes avant d'afficher un message sur sa sortie standard. Le processus fils de son côté incrémente les variables `l` et `g` et modifie la zone mémoire pointée par `*m` puis affiche leur contenu sur sa sortie standard puis se termine.

L'exécution de ce programme illustre bien que le processus fils démarre avec une copie du processus père lorsque l'appel système fork(2) se termine. Le processus fils peut modifier les variables qui ont été initialisées par le processus mais ces modifications n'ont aucun impact sur les variables utilisées dans le processus père. Même si le processus père et le processus fils sont identiques au moment de la création du processus fils, ils sont complètement indépendants par après. C'est une différence importante avec les threads. Contrairement à ce qu'il se passe avec les threads, un processus père et un processus fils ne partagent ni le segment de données, ni le heap ni le stack. Ces zones mémoires ne peuvent pas être utilisées directement pour permettre à un processus père de communiquer avec son fils.

Note : Quel est le processus qui s'exécute en premier après fork(2) ?

Après l'exécution de l'appel système fork(2) et la création du processus fils, le *kernel* se trouve face à deux processus qui sont dans l'état *Ready*. Si il y a deux processeurs libres, le *kernel* pourra les démarrer quasi simultanément. Par contre, si un seul processeur est disponible, le *kernel* devra exécuter l'un des deux processus en premier. En pratique, rien ne permet de contrôler si le *kernel* commencera d'abord l'exécution du processus père ou l'exécution du processus fils. Tout programme utilisant fork(2) doit pouvoir fonctionner correctement quel que soit l'ordre d'exécution des processus père et fils.

Le *kernel* gère les processus et attribue un identifiant à chaque processus. Le type `pid_t` est utilisé pour les identifiants de processus sous Unix. Ce type correspond à un nombre entier généralement non-signé. Le nombre maximum de processus qui peuvent être lancés sur un système Linux est un des paramètres fixés à la compilation

ou au démarrage du kernel. L'appel système getpid(2) retourne l'identifiant du processus courant tandis que l'appel système getppid(2) retourne l'identifiant du processus père.

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/types.h>


int main (int argc, char *argv[])  {
  int pid=getpid();
  int ppid=getppid();
  printf("Processus %d, parent:%d\n",pid,ppid);

  return(EXIT_SUCCESS);
}
```

Après l'exécution de fork(2) le processus père et le processus fils ont un identifiant de processus différent mais ils partagent certaines ressources qui sont gérées par le *kernel*. C'est le cas notamment des flux standard *stdin*, *stdout* et *stderr*. Lorsque le *kernel* crée un processus fils, il conserve la même sortie standard que le processus père. C'est ce qui nous permet de visualiser le résultat de l'exemple précédent. Cependant, le processus père et le processus fils sont en concurrence pour écrire sur la sortie standard. Si aucune précaution n'est prise, ces deux processus risquent d'écrire de façon désordonnée sur la sortie standard.

Pour mieux comprendre le problème, analysons l'exécution du programme ci-dessous. Il crée un processus fils puis le père et le fils écrivent sur *stdout*.

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/types.h>
#include <time.h>


void output(char c) {
  printf("Processus : %d\n",getpid());
  srand(getpid()+time(NULL));
  for(int i=0;i<60;i++) {
    putchar(c);
    int err=usleep((unsigned int) (rand()%10000));
    if(err<0) {
      perror("usleep");
      exit(EXIT_FAILURE);
    }
  }
}

int main (int argc, char *argv[])  {

  if(argc > 1) {
    setbuf(stdout,NULL);
  }

  pid_t pid;

  pid=fork();
  if (pid==-1) {
    // erreur à l'exécution de fork
    perror("fork");
    exit(EXIT_FAILURE);
  }
  // pas d'erreur
  if (pid==0) {
```

```
    // processus fils
    output('f');
    return(EXIT_SUCCESS);
  }
  else {
    // processus père
    output('P');
    printf("\n");
    sleep(1);
    // ...
    return(EXIT_SUCCESS);
  }
}
```

Lors de son exécution, il affiche les lignes suivantes sur *stdout*.

```
$ ./fork-buf
Processus : 852
Processus : 853
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
fffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

A première vue, on pourrait penser qu'il n'y a pas de problèmes d'accès concurrents à *stdout* puisque la sortie produite par ce programme semble claire. Cependant, la fonction `output` utilisée par le père et le fils pour afficher les caractères `P` et `f` devrait permettre aux deux processus d'afficher des caractères `P` et `f` avec une certaine alternance. Il n'en est rien et la sortie standard semble indiquer que le processus père a affiché tous les caractères `P` simultanément. Ce n'est pas possible si l'on regarde le code source de la fonction `output`.

L'utilitaire strace(1) qui permet de tracer tous les appels systèmes effectués par un processus nous permet d'analyser comment le processus père et le processus fils écrivent sur *stdout*. Lorsqu'ils utilisent les fonctions printf(3) et putchar(3) de la librairie standard, ils utilisent des fonctions qui finalement utilisent l'appel système write(2) qui est le seul permettant d'écrire sur un flux tel que *stdout*. L'exécution de strace(1) montre que chaque processus effectue deux appels systèmes write(2). Le premier correspond à l'appel à printf(3) et le second à tous les appels à putchar(3) fait par le processus.

```
strace -f -e trace=write ./fork-buf
Process 1398 attached
[pid  1398] write(1, "Processus : 1398\n", 17 <unfinished ...>
[pid  1397] write(1, "Processus : 1397\n", 17Processus : 1397
) = 17
Processus : 1398
[pid  1398] <... write resumed> )       = 17
[pid  1398] write(1, "fffffffffffffffffffffffffffff"..., 60fffffffff ffffffffffffffffffffff
Process 1398 detached
--- SIGCHLD (Child exited) @ 0 (0) ---
write(1, "PPPPPPPPPPPPPPPPPPPPPPPPPPPPP"..., 61PPPPPPPPPPPPPPPPPPPPP PPPPPPPPPPPPPPPPPPPPPPPP
) = 61
```

L'exécution d'un appel système tel que write(2) peut être couteux au niveau des performances. Une règle de bonne pratique est d'essayer de minimiser le nombre d'appels systèmes effectués par un programme. La librairie standard prend en compte cette règle et utilise un buffer dans lequel elle stocke temporairement les données produites par les fonctions telles que printf(3) ou putchar(3) avant de les écrire via write(2). Ce buffer peut être contrôlé via les fonctions setvbuf(3) et setbuf(3). Ces fonctions permettent de fixer la taille du buffer utilisé par la librairie standard.

Lorsque le buffer de la librairie stdio(3) est désactivé en faisant `setbuf(stdout,NULL);`, le processus père et le processus fils écrivent bien de façon entrelacée sur la sortie standard.

```
$ ./fork-buf 1
Processus : 1419
PProcessus : 1420
fffPfffPffPffPfPfffPPfPPPfPfffPfPfPPffffPfPffPPfPPPPfPPfPPfPfPPfPfPffPfPfPfffffPPfPPffPffPPfPfPfPPfPf
```

L'utilisation de *strace(1)* lors de cette exécution montre effectivement que chaque appel à la fonction putchar(3) provoque une exécution de l'appel système write(2) :

```
[pid  1420] write(1, "f", 1f)            = 1
[pid  1419] write(1, "P", 1P)            = 1
[pid  1419] write(1, "P", 1P)            = 1
[pid  1420] write(1, "f", 1f)            = 1
```

Note : Faut-il modifier le buffer de la librairie stdio(3) ?

En pratique, il est préférable de ne pas désactiver le buffer utilisé par la librairie stdio(3) car cela peut avoir des conséquences négatives sur les performances des programmes. Par contre, lorsque l'on développe des programmes qui utilisent plusieurs processus il est important de se souvenir de l'existence de ce buffer car il peut expliquer certains comportements qui pourraient apparaître comme étant bizarres lorsque l'on observe l'exécution de processus via les messages qu'ils affichent sur *stderr* ou *stdout*. Lorsque l'on soupçonne un comportement bizarre qui pourrait être expliqué par des interactions avec ce buffer, il est possible d'ajouter dans le programme des appels explicites à la fonction fflush(3) qui a pour effet de vider immédiatement le buffer de stdio(3).

Note : Génération de nombres aléatoires

Le programme `/Threads/S8-src/fork-buf.c` présenté ci-dessus est un exemple d'utilisation de nombres aléatoires. Ceux-ci sont générés avec la fonction rand(3) de la librairie standard. Cette fonction utilise un générateur de nombres aléatoires qui génère toujours la même séquence de nombres aléatoires lorsqu'elle est initialisée avec la même semence par la fonction srand(3). Souvent, les programmeurs qui utilisent des nombres aléatoires cherchent à ce que la séquence générée diffère d'une exécution du programme à l'autre. Une façon simple de procéder est d'utiliser comme semence la somme entre le temps courant retourné par time(3posix) et l'identifiant du processus obtenu via getpid(2). Une telle semence n'est cependant pas suffisante pour toutes les applications. Certaines applications cryptographiques notamment nécessitent des nombres aléatoires qui ne peuvent pas être facilement prédits. Pour ces applications, il est nécessaire d'utiliser des semences qui sont parfaitement aléatoires, comme random(4).

6.7.3 Fin d'un processus

Il y a deux événements importants dans la vie d'un processus sous Unix. Sa création avec l'appel système fork(2) et sa terminaison. Nous avons déjà vu qu'un programme C (et donc un processus) pouvait se terminer de deux façons principales [16] :
– par l'exécution de `return(...)` dans la fonction `main`
– par un appel explicite à la fonction exit(3) dans la fonction `main` ou n'importe quelle fonction du processus
Ces fonctions appellent en fait la fonction de la librairie exit(3). Cette fonction permet de faire plus que simplement terminer le processus en cours d'exécution et retourner sa valeur de retour. Il est en effet possible d'associer une ou plusieurs fonctions de terminaison à exit(3) via la fonction atexit(3). Lorsque exit(3) est appelée, elle lance d'abord les fonctions enregistrées par atexit(3) puis termine correctement le processus. Ces fonctions de terminaison d'un processus sont utilisées lorsque par exemple un processus utilise des services particuliers du système d'exploitation comme par exemple une mémoire partagée entre plusieurs processus. Ces services consomment des ressources et il est nécessaire de les libérer correctement lorsqu'un processus se termine comme nous le verrons ultérieurement.

L'exemple ci-dessous illustre brièvement l'utilisation de atexit(3).

```c
#include <stdio.h>
#include <stdlib.h>

void e1() {
  printf("Exécution de la fonction e1\n");
}
```

16. Si le processus a été découpé en threads, le processus peut aussi se terminer lorsque son dernier thread se termine en exécutant `return(...)` dans sa fonction de démarrage ou par un appel explicite à pthread_exit(3).

```
int main (int argc, char *argv[])  {

  int err;
  err=atexit(e1);
  if(err==-1) {
    perror("atexit");
    exit(EXIT_FAILURE);
  }
  return(EXIT_SUCCESS);

}
```

Après avoir exécuté les fonctions de terminaison, la fonction exit(3) appelle fflush(3) sur tous les flux existants puis les ferme proprement. Ensuite, la fonction exit(3) exécute l'appel système exit(2). Cet appel système est particulier. C'est le seul appel système qui n'a pas de valeur de retour, et pour cause ! Il ferme tous les fichiers qui étaient encore ouverts (normalement un processus devrait fermer proprement tous ses fichiers avant de s'arrêter) et libère les ressources qui étaient associées au processus.

```
#include <unistd.h>

void _exit(int status);
```

L'appel système exit(2) permet au processus qui se termine de retourner un statut à son processus père. Pour récupérer le statut de son fils, un processus père doit utiliser l'appel système waitpid(2).

```
#include <sys/types.h>
#include <sys/wait.h>

pid_t waitpid(pid_t pid, int *status, int options);
```

L'appel système waitpid(2) prend trois arguments. C'est un appel système bloquant. Le premier argument permet de spécifier quel est le processus fils dont la terminaison est attendue. Un premier argument négatif indique que waitpid(2) attend la terminaison de n'importe quel processus fils. Si le premier argument est positif, alors il contient un identifiant de processus fils et waitpid(2) attendra la terminaison de ce processus [17]. Le second argument est un pointeur vers un entier qui après le retour de waitpid(2) contiendra le statut retourné par le processus fils. Le troisième argument permet de spécifier des options à waitpid(2) que nous n'utiliserons pas. La fonction wait(2) est une simplification de waitpid(2) qui permet d'attendre n'importe quel processus fils. `wait(p)` est en pratique équivalent à `waitpid(-1,p,0)`.

Un processus qui lance un processus fils avec fork(2) *doit* attendre la terminaison de son processus fils en utilisant waitpid(2). Le programme ci-dessous illustre l'utilisation de waitpid(2).

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/wait.h>
#include <sys/types.h>


int main (int argc, char *argv[])  {
  int status;
  pid_t pid;

  pid=fork();

  if (pid==-1) {
    // erreur à l'exécution de fork
    perror("fork");
    exit(EXIT_FAILURE);
  }
  // pas d'erreur
```

17. Si le processus dont l'identifiant est passé comme argument s'est déjà terminé, alors waitpid(2) retourne en indiquant une erreur.

```
  if (pid==0) {
    sleep(8);
    return(42);
  }
  else {
    // processus père
    int fils=waitpid(pid,&status,0);
    if(fils==-1) {
      perror("wait");
      exit(EXIT_FAILURE);
    }
    if(WIFEXITED(status)) {
      printf("Le fils %d s'est terminé correctement et a retourné la valeur %d\n",fils,WEXITSTATUS
      return(EXIT_SUCCESS);
    }
    else {
      if( WIFSIGNALED(status)) {
        printf("Le fils %d a été tué par le signal %d\n",fils,WTERMSIG(status));
      }
      return(EXIT_FAILURE);
    }
  }
}
```

Dans ce programme, le processus père récupère la valeur retournée par le fils qu'il a créé. Lors de l'exécution de `waitpid(pid,&status,0);`, la valeur de retour du fils est placée dans l'entier dont l'adresse est `status`. Cet entier contient non-seulement la valeur de retour du processus fils (dans les 8 bits de poids faible), mais aussi une information permettant de déterminer si le processus fils s'est terminé correctement ou a été terminé de façon abrupte via l'utilisation de kill(1). Les macros `WEXITSTATUS` et `WTERMSIG` utilisées pour extraire la valeur de retour et la raison de la terminaison abrupte sont décrites dans waitpid(2).

Même si un processus *doit* attendre la terminaison de tout processus fils qu'il a lancé, il arrive parfois qu'un processus n'attende pas ses fils. Cela peut arriver lorsqu'un processus s'arrête suite à une erreur avant de pouvoir récupérer ses fils. Ce cas est illustré par l'exemple ci-dessous dans lequel le processus père se termine sans attendre son fils.

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/wait.h>
#include <sys/types.h>


int main (int argc, char *argv[])  {
  pid_t pid;

  pid=fork();

  if (pid==-1) {
    // erreur à l'exécution de fork
    perror("fork");
    exit(EXIT_FAILURE);
  }
  // pas d'erreur
  if (pid==0) {
    printf("Processus : %d, père : %d\n",getpid(),getppid());
    fflush(stdout);
    sleep(3);
    printf("Processus : %d, père : %d\n",getpid(),getppid());
    return(EXIT_SUCCESS);
  }
  else {
    // processus père
```

```
    sleep(1);
    printf("Fin du processus père [%d]\n",getpid());
    return(EXIT_FAILURE);
  }
}
```

Du point de vue du *kernel* cette situation est ennuyeuse car il maintient pour chaque processus non seulement son identifiant de processus mais également l'identifiant de son processus père qui est retourné par getpid(2). Lorsque le père se termine avant son fils, le processus fils est dit *orphelin* et le kernel modifie ses structures de données pour que le père de ce *processus orphelin* soit le processus dont l'identifiant est `1`. Ce processus est le processus init(8) qui est lancé au démarrage du système et n'est jamais arrêté.

```
Processus : 28750, père : 28749
Fin du processus père [28749]
Processus : 1, père : 1
```

A côté des processus orphelins dont nous venons de parler, un système Unix peut également héberger des *processus zombie*. Un *processus zombie* est un processus qui s'est terminé mais dont la valeur de retour n'a pas encore été récupérée par son père. Dans ce cas, le *kernel* libère l'ensemble des ressources associées au processus fils et ne conserve de ce processus qu'une petite structure de données contenant notamment son identifiant, l'identifiant de son processus père et sa valeur de retour. En pratique, il est préférable d'éviter les processus zombie car ils consomment quand même un peu de ressources.

6.7.4 Exécution d'un programme

fork(2) et waitpid(2) permettent respectivement de créer et de terminer des processus. Pour comprendre la façon dont les programmes sont exécutés, il nous reste à expliquer le fonctionnement de l'appel système execve(2). Cet appel système permet l'exécution d'un programme. Lors de son exécution, l'image en mémoire du processus qui effectue execve(2) est remplacée par l'image de l'exécutable passé en argument à execve(2) et son exécution démarre à sa fonction `main`.

```
#include <unistd.h>

int execve(const char *path, char *const argv[], char *const envp[]);
```

execve(2) prend trois arguments. Le premier est le nom complet du fichier exécutable qui doit être lancé. Le second est un pointeur vers un tableau de chaines de caractères contenant les arguments à passer à l'exécutable. Le troisième est un pointeur vers l'environnement qui sera nécessaire à l'exécution du programme. Comme execve(2) remplace l'image mémoire du programme en cours d'exécution, il ne retourne une valeur de retour que si l'appel système échoue. Cela peut être le cas si son premier argument n'est pas un fichier exécutable accessible par exemple.

execve(2) s'utilise souvent juste après l'exécution de fork(2), mais il est aussi possible de l'utiliser directement dans un programme. Dans ce cas, le programme qui exécute avec succès execve(2) disparaît et est remplacé par le programme appelé. Le programme ci-dessous illustre une utilisation simple de execve(2).

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>

int main (int argc, char *argv[])  {

  char *arguments[]={"expr", "1", "+", "2", NULL};
  char *environnement[]={"LANG=fr",NULL};

  printf("Exécution du processus %d\n",getpid());
  printf("Exécution de /usr/bin/expr\n");
  int err=execve("/usr/bin/expr", arguments, environnement);
  if(err!=0) {
    perror("execve");
```

```
    exit(EXIT_FAILURE);
  }
  // jamais atteint
  printf("Ce message ne sera jamais affiché\n");
  return(EXIT_SUCCESS);

}
```

Lors de son exécution, ce programme affiche sur sa sortie standard les lignes suivantes :

```
3
```

Il y a quelques points importants à noter concernant l'utilisation de execve(2). Tout d'abord, execve(2) remplace l'entièreté de l'image mémoire du processus qui exécute cet appel système, y compris les arguments, les variables d'environnement. Par contre, le *kernel* conserve certaines informations qu'il maintenait pour le processus. C'est le cas notamment de l'identifiant du processus et de l'identifiant du processus père. Si le processus qui a effectué execve(2) avait lancé des threads, ceux-ci seraient immédiatement supprimés puisque l'image du processus en cours d'exécution est remplacé lors de l'exécution de execve(2). Les flux standard (*stdin*, *stdout* et *stderr*) sont utilisables par le programme exécuté via execve(2). Il faut cependant noter que lors de l'appel à execve(2), les données qui se trouveraient éventuellement dans le buffer de la librairie *stdio* ne sont pas automatiquement envoyées vers leurs flux respectifs. Cela pourrait paraître étonnant puisque lorsqu'un processus se termine avec exit(3), exit(3) vide les buffers de *stdio* avant d'appeler exit(2). execve(2) est un appel système qui est exécuté par le kernel. Celui-ci ne peut pas savoir si il y a des données en attente d'écriture dans *stdio*. Il ne peut donc pas automatiquement vider les buffers maintenus par la librairie *stdio*. Si des données ont été écrites avec printf(3) avant l'exécution de execve(2), il est préférable de forcer leur écriture via fflush(3) avant d'appeler execve(2).

L'appel système execve(2) est très souvent exécuté dans un shell tel que bash(1). Lorsqu'un shell lance un programme externe, il doit d'abord utiliser fork(2) pour créer une copie de lui-même. Ensuite, le processus père se met en attente via waitpid(2) de la valeur de retour du processus fils créé. Le processus fils quant à lui utilise execve(2) pour exécuter le programme demandé.

La programme ci-dessous est un exemple un peu plus complexe de l'utilisation de fork(2), execve(2) et waitpid(2). Ce programme prend comme argument une liste d'exécutables et il essaye de les exécuter l'un à la suite de l'autre. Pour cela, il parcourt ses arguments et essaye pour chaque argument de créer un processus fils et d'y exécuter le programme correspondant. Si le programme a pu être exécuté, sa valeur de retour est récupérée par le processus père. Si l'appel à execve(2) a échoué, le processus fils se termine avec `127` comme valeur de retour. Comme celle-ci est stockée sur 8 bits, c'est la plus grande valeur de retour positive qu'il est possible de retourner depuis un processus fils. Cette valeur indique au processus père que le fils n'a pas réussi à exécuter execve(2).

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/wait.h>
#include <sys/types.h>
#include <libgen.h>

extern char **environ;

int main (int argc, char *argv[])  {
  int status;
  pid_t pid;

  for(int i=1;i<argc;i++) {
    // création du fils
    pid=fork();
    if (pid==-1) {
      perror("fork");
      exit(EXIT_FAILURE);
    }
    if (pid==0) {
      // fils
      printf ("Exécution de la commande %s [pid=%d]\n",argv[i],getpid());
```

```
      fflush(stdout);
      char *arguments[2];
      arguments[0]=basename(argv[i]);
      arguments[1]=NULL;
      int err=execve(argv[i], arguments, environ);
      if(err!=0)
        return(127);
    } // fils
    else {
      // processus père
      int fils=waitpid(pid,&status,0);
      if(fils==-1) {
        perror("wait");
        exit(EXIT_FAILURE);
      }
      if(WIFEXITED(status)) {
        if(WEXITSTATUS(status)==0)
          printf("La commande %s [%d] s'est terminée correctement\n",argv[i],fils);
        else
          if (WEXITSTATUS(status)==127)
            printf("La commande %s n'a pu être exécutée\n",argv[i]);
          else
            printf("La commande %s [%d] a retourné %d\n",argv[i],fils,WEXITSTATUS(status));
      }
      else {
        if( WIFSIGNALED(status))
          printf("La commande %s [%d] ne s'est pas terminée correctement\n",argv[i],fils);
      }
      fflush(stdout);
    } // père
  } // for loop
  return(EXIT_SUCCESS);
}
```

Lors de son exécution, ce programme affiche sur sa sortie standard les lignes suivantes :

```
$./fork-manyexec /bin/true /bin/false /bin/none
Exécution de la commande /bin/true [pid=14217]
La commande /bin/true [14217] s'est terminée correctement
Exécution de la commande /bin/false [pid=14218]
La commande /bin/false [14218] a retourné 1
Exécution de la commande /bin/none [pid=14219]
La commande /bin/none n'a pu être exécutée
```

En pratique, il existe plusieurs fonctions de la librairie standard qui apportent de petites variations à execve(2). Il s'agit de execl(3), execlp(3), execle(3), execv(3posix) et execv(3). Ces fonctions utilisent toutes l'appel système execve(2). Elles permettent de spécifier de différentes façons le programme à exécuter ou les variables d'environnement. Enfin, la fonction system(3) de la librairie permet d'exécuter une commande du shell directement depuis un programme.

Outre les exécutables compilés, Unix et Linux supportent également l'exécution de programmes interprétés. Contrairement aux programmes compilés que nous avons manipulé jusque maintenant, un programme interprété est un programme écrit dans un langage qui doit être utilisé via un *interpréteur*. Un *interpréteur* est un programme qui lit des commandes sous la forme de texte et exécute directement les instructions correspondant à ces commandes. Unix supporte de nombreux interpréteurs et comme nous allons le voir il est très facile de rajouter de nouveaux interpréteurs de commande. L'interpréteur le plus connu est bash(1) et ses nombreuses variantes. En voici quelques autres :

– awk(1) est un langage de programmation interprété qui permet de facilement manipuler des textes
– perl(1) est un langage de programmation complet qui a été initialement développé pour la manipulation de textes, mais est utilisé dans de nombreuses autres applications
– python(1) est un langage de programmation complet

Pour comprendre la façon dont Unix interagit avec les interpréteurs de commande, il est bon de voir en détails comment execve(2) reconnaît qu'un fichier contient un programme qui peut être exécuté. Tout d'abord, le système de fichiers contient pour chaque fichier des métadonnées qui fournissent de l'information sur le possesseur du fichier, sa date de création, ... Une de ces métadonnées est un bit [18] qui indique si le fichier est exécutable ou non. Ce bit peut être manipulé en utilisant la commande chmod(1). Lorsqu'un programme est compilé avec gcc(1), celui-ci utilise chmod(1) pour marquer le programme comme étant exécutable.

```
$ ls -l a.out
-rwxr-xr-x 1 obo stafinfo 8178 Mar 16 13:42 a.out
$ chmod -x a.out
$ ./a.out
-bash: ./a.out: Permission denied
$ chmod +x a.out
$ ./a.out
  exécution de a.out
$ ls -l a.out
-rwxr-xr-x 1 obo stafinfo 8178 Mar 16 13:42 a.out
```

Lorsqu'execve(2) est appelé, il vérifie d'abord ce bit de permission. Si il n'indique pas que le programme est exécutable, execve(2) retourne une erreur. Ensuite, execve(2) ouvre le fichier dont le nom a été passé comme premier argument. Par convention, le début du fichier contient une séquence d'octets ou de caractères qui indiquent le type de fichier dont il s'agit. La commande file(1) permet de tester le type d'un fichier inconnu.

```
$ file fork-execve.c
fork-execve.c: UTF-8 C program text
$ file script.sh
script.sh: Bourne-Again shell script text executable
$ file a.out
a.out: ELF 64-bit LSB executable, x86-64, version 1 (GNU/Linux),   dynamically linked (uses shared
```

Pour les exécutables, deux cas de figure sont possibles :

1. le fichier contient un programme compilé et directement exécutable. Sur les systèmes Linux actuels, ce fichier sera au format elf(5). Il débute par une entête qui contient une chaîne de caractères utilisée comme marqueur ou chaîne magique. L'entête fournit de l'information sur le type d'exécutable et sa structure. Voici à titre d'illustration le contenu de l'entête d'un programme compilé décortiqué par l'utilitaire readelf(1) :

   ```
   $ readelf -h a.out
   ELF Header:
    Magic:   7f 45 4c 46 02 01 01 03 00 00 00 00 00 00 00 00
    Class:                             ELF64
    Data:                              2's complement, little endian
    Version:                           1 (current)
    OS/ABI:                            UNIX - Linux
    ABI Version:                       0
    Type:                              EXEC (Executable file)
    Machine:                           Advanced Micro Devices X86-64
    Version:                           0x1
    Entry point address:               0x4006e0
    Start of program headers:          64 (bytes into file)
    Start of section headers:          3712 (bytes into file)
    Flags:                             0x0
    Size of this header:               64 (bytes)
    Size of program headers:           56 (bytes)
    Number of program headers:         8
    Size of section headers:           64 (bytes)
    Number of section headers:         30
    Section header string table index: 27
   ```

18. En pratique, il y a trois bits qui jouent ce rôle en fonction du possesseur du fichier et de l'utilisateur qui souhaite l'exécuter. Nous décrirons ces bits en détails dans un prochain chapitre.

2. Le fichier contient un programme en langage interprété. Dans ce cas, la première ligne débute par `#!` suivi du nom complet de l'interpréteur à utiliser et de ses paramètres éventuels. Le programme interprété commence sur la deuxième ligne. A titre d'exemple, voici un petit script bash(1) qui permet de tester si un fichier est interprétable ou non en testant la valeur des deux premiers caractères du fichier et ses métadonnées.

```
#!/bin/bash
# script.sh
if [ $# -ne 1 ]
then
    echo "Usage: `basename $0` fichier"
    exit 1
fi
if [ -x ${1} ]
then
    head -1 $1 | grep "^#\!" >>/dev/null
    if [ $? ]
    then
        echo "Script interprétable"
        exit 0
    else
        echo "Script non-interprétable"
        exit 1
    fi
else
    echo "Bit x non mis dans les métadonnées"
    exit 1
fi
```

Sous Unix et Linux, n'importe quel programmeur peut définir son propre interpréteur. Il suffit qu'il s'agisse d'un exécutable compilé et que le nom de cet interpréteur soit présent dans la première ligne du fichier à interpréter. Lors de l'exécution d'un programme utilisant cet interpréteur, celui-ci recevra le contenu du fichier et pourra l'interpréter. Ainsi, par exemple le programme interprété ci-dessous est tout à fait valide.

```
#!/usr/bin/tail -n +1
Hello, world
SINF1252
```

Lors de son exécution via execve(2), l'interpréteur tail(1) va être chargé avec comme arguments `-n +1` et il affichera sur *stdout* la ligne `SINF1252`.

Cette facilité d'ajouter de nouveaux interpréteurs de commande est une des forces des systèmes d'exploitation de la famille Unix.

6.7.5 Table des processus

Un système d'exploitation tel que Linux maintient certaines informations concernant chaque processus dans sa *table des processus*. Une description complète du contenu de cette table des processus sort du cadre de ce chapitre. Par contre, il est intéressant de noter que sous Linux il existe de nombreux utilitaires qui permettent de consulter le contenu de la table des processus et notamment :
– ps(1) qui est l'utilitaire de base pour accéder à la table de processus et lister les processus en cours d'exécution
– top(1) qui affiche de façon interactive les processus qui consomment actuellement du temps CPU, de la mémoire, ...
– pstree(1) qui affiche l'arbre des processus avec les relations père-fils
Tous ces utilitaires utilisent les informations contenues dans le répertoire `/proc`. Il s'agit d'un répertoire spécial qui contient de l'information à propos du système d'exploitation y compris la table de processus. Son contenu est détaillé dans la page de manuel qui lui est consacrée : proc(5).

A titre d'illustration, considérons le shell d'un utilisateur en cours. Les informations maintenues dans la table des processus pour ce processus sont accessibles depuis `/proc/pid` où `pid` est l'identifiant du processus en cours

d'exécution. Linux stocke de très nombreuses informations sur chaque processus. Celles-ci sont structurées dans des fichiers et des répertoires :

```
$ ls -l /proc/18557
total 0
dr-xr-xr-x 2 obo stafinfo 0 Mar 18 16:37 attr
-r-------- 1 obo stafinfo 0 Mar 18 16:37 auxv
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 cgroup
--w------- 1 obo stafinfo 0 Mar 18 16:37 clear_refs
-r--r--r-- 1 obo stafinfo 0 Mar 18 14:56 cmdline
-rw-r--r-- 1 obo stafinfo 0 Mar 18 16:37 coredump_filter
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 cpuset
lrwxrwxrwx 1 obo stafinfo 0 Mar 18 16:37 cwd ->  /etinfo/users2/obo/sinf1252/SINF1252/
-r-------- 1 obo stafinfo 0 Mar 18 16:37 environ
lrwxrwxrwx 1 obo stafinfo 0 Mar 18 16:37 exe -> /bin/bash
dr-x------ 2 obo stafinfo 0 Mar 18 14:56 fd
dr-x------ 2 obo stafinfo 0 Mar 18 16:37 fdinfo
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 io
-rw------- 1 obo stafinfo 0 Mar 18 16:37 limits
-rw-r--r-- 1 obo stafinfo 0 Mar 18 16:37 loginuid
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 maps
-rw------- 1 obo stafinfo 0 Mar 18 16:37 mem
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 mountinfo
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 mounts
-r-------- 1 obo stafinfo 0 Mar 18 16:37 mountstats
dr-xr-xr-x 6 obo stafinfo 0 Mar 18 16:37 net
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 numa_maps
-rw-r--r-- 1 obo stafinfo 0 Mar 18 16:37 oom_adj
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 oom_score
-r-------- 1 obo stafinfo 0 Mar 18 16:37 pagemap
-r-------- 1 obo stafinfo 0 Mar 18 16:37 personality
lrwxrwxrwx 1 obo stafinfo 0 Mar 18 16:37 root -> /
-rw-r--r-- 1 obo stafinfo 0 Mar 18 16:37 sched
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 schedstat
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 sessionid
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 smaps
-r-------- 1 obo stafinfo 0 Mar 18 16:37 stack
-r--r--r-- 1 obo stafinfo 0 Mar 18 14:56 stat
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 statm
-r--r--r-- 1 obo stafinfo 0 Mar 18 14:56 status
-r-------- 1 obo stafinfo 0 Mar 18 16:37 syscall
dr-xr-xr-x 3 obo stafinfo 0 Mar 18 15:59 task
-r--r--r-- 1 obo stafinfo 0 Mar 18 16:37 wchan
```

Certaines des entrées dans `/proc` sont des fichiers, d'autres sont des répertoires. A titre d'exemple, voici quelques unes des entrées utiles à ce stade de notre exploration de Linux.

- `cmdline` est un fichier texte contenant la ligne de commande utilisée pour lancer le processus
- `environ` est un fichier texte contenant les variables d'environnement passées au processus

  ```
  $ (cat /proc/18557/environ; echo) | tr '\000' '\n'
  USER=obo
  LOGNAME=obo
  HOME=/etinfo/users2/obo
  PATH=/usr/local/bin:/bin:/usr/bin
  MAIL=/var/mail/obo
  SHELL=/bin/bash
  ```

- `status` est une indication sur l'état actuel du processus. Les premières lignes indiquent dans quel état le processus se trouve ainsi que son identifiant, l'identifiant de son père, ...

  ```
  $ cat /proc/$$/status | head -5
  Name:     bash
  State:    S (sleeping)
  ```

```
Tgid:     18557
Pid:      18557
PPid:     18556
```

– `limits` est un fichier texte contenant les limites actuelles imposées par le système sur le processus. Ces limites peuvent être modifiées en utilisant *ulimit(1)* à l'intérieur de bash(1) ou via les appels systèmes getrlimit(2)/setrlimit(2).

```
$ cat /proc/18557/limits
Limit                     Soft Limit           Hard Limit           Units
Max cpu time              unlimited            unlimited            seconds
Max file size             unlimited            unlimited            bytes
Max data size             unlimited            unlimited            bytes
Max stack size            10485760             unlimited            bytes
Max core file size        0                    unlimited            bytes
Max resident set          unlimited            unlimited            bytes
Max processes             1024                 24064                processes
Max open files            1024                 1024                 files
Max locked memory         65536                65536                bytes
Max address space         unlimited            unlimited            bytes
Max file locks            unlimited            unlimited            locks
Max pending signals       24064                24064                signals
Max msgqueue size         819200               819200               bytes
Max nice priority         0                    0
Max realtime priority     0                    0
Max realtime timeout      unlimited            unlimited            us
```

– `task` est un répertoire qui contient pour chaque thread lancé par le processus un sous-répertoire avec toutes les informations qui sont relatives à ce thread.

Nous aurons l'occasion de présenter ultérieurement d'autres éléments utiles se trouvant dans `/proc`. Une description plus détaillée est disponible dans la page de manuel proc(5) et des livres de référence tels que [Kerrisk2010].

6.8 Mise en oeuvre de la synchronisation

Nous avons vu dans les chapitres précédents comment utiliser les constructions de synchronisation que sont les mutex et les sémaphores, fournies par les librairies POSIX. Nous allons nous intéresser dans ce chapitre à comment ces constructions sont implémentées.

Nous mettrons l'emphase plus particulièrement sur la résolution du problème de l'exclusion mutuelle, utile pour protéger l'accès aux structures de données utilisées pour la synchronisation (e.g. protéger l'accès à l'état d'un mutex ou le compteur associé à un sémaphore). Les mécanismes que nous verrons permettent de façon générale de protéger l'accès à des sections de code devant s'exécuter de manière exclusive les unes des autres.

Dans un premier temps, nous verrons des algorithmes "classiques" proposés pour résoudre le problème de l'exclusion mutuelle en utilisant uniquement des opérations de lecture et d'écriture en mémoire. Ces algorithmes ont un intérêt historique : ils ne sont pas exploitables (ou alors avec un surcoût très élevé) sur les architectures modernes. Il est toutefois intéressant de les étudier pour comprendre le problème de l'exclusion mutuelle. Par la suite, nous verrons des algorithmes utilisant des instructions spécifiques introduites dans les processeurs pour résoudre les problèmes de synchronisation, et étudierons leur performance.

6.8.1 Algorithme de Peterson

Le problème de l'exclusion mutuelle a intéressé de nombreux informaticiens depuis le début des années 1960s [Dijkstra1965] et différentes solutions à ce problème ont été proposées. Plusieurs d'entre elles sont analysées en détails dans [Alagarsamy2003]. Dans cette section, nous nous concentrerons sur une de ces solutions, proposée par G. Peterson en 1981 [Peterson1981]. Cette solution permet à plusieurs threads de coordonner leur exécution de façon à éviter une violation de section critique en utilisant uniquement des variables accessibles à tous les threads. Nous verrons tout d'abord des solutions pour coordonner l'accès à leur section critique pour deux threads. .. La

solution proposée par Peterson permet de gérer *N* threads [Peterson1981] mais nous nous limiterons à sa version permettant de coordonner deux threads.

Une première solution permettant de coordonner deux threads en utilisant des variables partagées pourrait être de s'appuyer sur une variable qui permet de déterminer quel est le thread qui peut entrer en section critique. Dans l'implémentation ci-dessous, la variable partagée `turn` est utilisée par les deux threads et permet de coordonner leur exécution. `turn` peut prendre les valeurs `0` ou `1`. Le premier thread exécute la boucle `while (turn != 0) { }`. Prise isolément, cette boucle pourrait apparaître comme une boucle inutile (`turn==0` avant son exécution) ou une boucle infinie (`turn==1` avant son exécution). Un tel raisonnement est incorrect lorsque la variable `turn` peut être modifiée par les deux threads. En effet, si `turn` vaut `1` au début de la boucle `while (turn != 0) { }`, la valeur de cette variable peut être modifiée par un autre thread pendant l'exécution de la boucle et donc provoquer son arrêt.

```
// thread 1
while (turn!=0)
{ /* loop */ }
section_critique();
turn=1;
// ...

// thread 2
while (turn!=1)
{ /* loop */ }
section_critique();
turn=0;
```

Il est intéressant d'analyser ces deux threads en détails pour déterminer si ils permettent d'éviter une violation de section critique et respectent les 4 contraintes précisées plus haut. Dans ces deux threads, pour qu'une violation de section critique puisse se produire, il faudrait que les deux threads passent en même temps la boucle `while` qui précède la section critique. Imaginons que le premier thread est entré dans sa section critique. Puisqu'il est sorti de sa boucle `while`, cela implique que la variable `turn` a la valeur `0`. Sinon, le premier thread serait toujours en train d'exécuter sa boucle `while`. Examinons maintenant le fonctionnement du second thread. Pour entrer dans sa section critique, celui-ci va exécuter la boucle `while (turn != 1){ }`. A ce moment, `turn` a la valeur `0`. La boucle dans le second thread va donc s'exécuter en permanence. Elle ne s'arrêtera que si la valeur de `turn` change. Or, le premier thread ne pourra changer la valeur de `turn` que lorsqu'il aura quitté sa section critique. Cette solution évite donc toute violation de la section critique. Malheureusement, elle ne fonctionne que si il y a une alternance stricte entre les deux threads. Le second s'exécute après le premier qui lui même s'exécute après le second, ... Cette alternance n'est évidemment pas acceptable.

Analysons une seconde solution. Celle-ci utilise un tableau `flag` contenant deux drapeaux, un par thread. Ces deux drapeaux sont initialisés à la valeur `false`. Pour plus de facilité, nous nommons les threads en utilisant la lettre `A` pour le premier et `B` pour le second. Le drapeau `flag[x]` est modifié par le thread `x` et sa valeur est testée par l'autre thread.

```
#define A 0
#define B 1
int flag[];
flag[A]=false;
flag[B]=false;
```

Le premier thread peut s'écrire comme suit. Il comprend une boucle `while` qui teste le drapeau `flag[B]` du second thread. Avant d'entrer en section critique, il met son drapeau `flag[A]` à `true` et le remet à `false` dès qu'il en est sorti.

```
// Thread A
while (flag[B]==true)
{ /* loop */ }
flag[A]=true;
section_critique();
flag[A]=false;
//...
```

Le second thread est organisé d'une façon similaire.

```
// Thread B
while (flag[A]==true)
{ /* loop */ }
flag[B]=true;
section_critique();
flag[B]=false;
// ...
```

Analysons le fonctionnement de cette solution et vérifions si elle permet d'éviter toute violation de section critique. Pour qu'une violation de section critique se produise, il faudrait que les deux threads exécutent simultanément leur section critique. La boucle `while` qui précède dans chaque thread l'entrée en section critique parait éviter les problèmes puisque si le thread `A` est dans sa section critique, il a mis `flag[A]` à la valeur `true` et donc le thread `B` exécutera en permanence sa boucle `while`. Malheureusement, la situation suivante est possible. Supposons que `flag[A]` et `flag[B]` ont la valeur `false` et que les deux threads souhaitent entrer dans leur section critique en même temps. Chaque thread va pouvoir traverser sa boucle `while` sans attente puis seulement mettre son drapeau à `true`. A cet instant il est trop tard et une violation de section critique se produira. Cette violation a été illustrée sur une machine multiprocesseur qui exécute deux threads simultanément. Elle est possible également sur une machine monoprocesseur. Dans ce cas, il suffit d'imaginer que le thread `A` passe sa boucle `while` et est interrompu par le scheduler avant d'exécuter `flag[A]=true;`. Le scheduler réalise un changement de contexte et permet au thread `B` de s'exécuter. Il peut passer sa boucle `while` puis entre en section critique alors que le thread `A` est également prêt à y entrer.

Une alternative pour éviter le problème de violation de l'exclusion mutuelle pourrait être d'inverser la boucle `while` et l'assignation du drapeau. Pour le thread `A`, cela donnerait le code ci-dessous :

```
// Thread A
flag[A]=true;
while (flag[B]==true)
{ /* loop */ }
section_critique();
flag[A]=false;
//...
```

Le thread `B` peut s'implémenter de façon similaire. Analysons le fonctionnement de cette solution sur un ordinateur monoprocesseur. Un scénario possible est le suivant. Le thread `A` exécute la ligne permettant d'assigner son drapeau, `flag[A]=true;`. Après cette assignation, le scheduler interrompt ce thread et démarre le thread `B`. Celui-ci exécute `flag[B]=true;` puis démarre sa boucle `while`. Vu le contenu du drapeau `flag[A]`, celle-ci va s'exécuter en permanence. Après quelque temps, le scheduler repasse la main au thread `A` qui va lui aussi entamer sa boucle `while`. Comme `flag[B]` a été mis à `true` par le thread `B`, le thread `A` entame également sa boucle `while`. A partir de cet instant, les deux threads vont exécuter leur boucle `while` qui protège l'accès à la section critique. Malheureusement, comme chaque thread exécute sa boucle `while` aucun des threads ne va modifier son drapeau de façon à permettre à l'autre thread de sortir de sa boucle. Cette situation perdurera indéfiniment. Dans la littérature, cette situation est baptisée un *livelock*. Un *livelock* est une situation dans laquelle plusieurs threads exécutent une séquence d'instructions (dans ce cas les instructions relatives aux boucles `while`) sans qu'aucun thread ne puisse réaliser de progrès. Un *livelock* est un problème extrêmement gênant puisque lorsqu'il survient les threads concernés continuent à utiliser le processeur mais n'exécutent aucune instruction utile. Il peut être très difficile à diagnostiquer et il est important de réfléchir à la structure du programme et aux techniques de coordination entre les threads qui sont utilisées afin de garantir qu'aucun *livelock* ne pourra se produire.

L'algorithme de Peterson [Peterson1981] combine les deux idées présentées plus tôt. Il utilise une variable `turn` qui est testée et modifiée par les deux threads comme dans la première solution et un tableau `flag[]` comme la seconde. Les drapeaux du tableau sont initialisés à `false` et la variable `turn` peut prendre la valeur `A` ou `B`.

```
#define A 0
#define B 1
int flag[];
flag[A]=false;
flag[B]=false;
```

Le thread `A` peut s'écrire comme suit.

```
// thread A
flag[A]=true;
turn=B;
while((flag[B]==true)&&(turn==B))
{ /* loop */ }
section_critique();
flag[A]=false;
// ...
```

Le thread `B` s'implémente de façon similaire.

```
// Thread B
flag[B]=true;
turn=A;
while((flag[A]==true)&&(turn==A))
{ /* loop */ }
section_critique();
flag[B]=false;
// ...
```

Pour vérifier si cette solution répond bien au problème de l'exclusion mutuelle, il nous faut d'abord vérifier qu'il ne peut y avoir de violation de la section critique. Pour qu'une violation de section critique soit possible, il faudrait que les deux threads soient sortis de leur boucle `while`. Examinons le cas où le thread `B` se trouve en section critique. Dans ce cas, `flag[B]` a la valeur `true`. Si le thread `A` veut entrer en section critique, il va d'abord devoir exécuter `flag[A]=true;` et ensuite `turn=B;`. Comme le thread `B` ne modifie ni `flag[A]` ni `turn` dans sa section critique, thread `A` va devoir exécuter sa boucle `while` jusqu'à ce que le thread `B` sorte de sa section critique et exécute `flag[B]=false;`. Il ne peut donc pas y avoir de violation de la section critique.

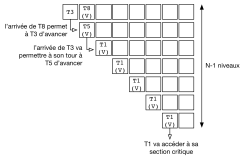
Il nous faut également montrer que l'algorithme de Peterson ne peut pas causer de *livelock*. Pour qu'un tel *livelock* soit possible, il faudrait que les boucles `while((flag[A]==true)&&(turn==A)) {};` et `while((flag[B]==true)&&(turn==B)) {};` puissent s'exécuter en permanence en même temps. Comme la variable `turn` ne peut prendre que la valeur `A` ou la valeur `B`, il est impossible que les deux conditions de boucle soient simultanément vraies.

Enfin, considérons l'impact de l'arrêt d'un des deux threads. Si thread `A` s'arrête hors de sa section critique, `flag[A]` a la valeur `false` et le thread `B` pourra toujours accéder à sa section critique.

6.8.2 Algorithme du filtre

La version de l'algorithme de Peterson que nous avons vu permet de synchroniser l'accès à la section critique de *seulement* deux threads. Il est possible d'étendre son principe pour supporter plusieurs threads, sous le principe de l'algorithme dit du filtre (Filter algorithm), lui aussi proposé par Gary L. Peterson.

Cet algorithme nécessite de connaître à l'avance le nombre de threads N qui souhaitent synchroniser l'accès à leur section critique. Le concept fondamental est celui de *niveaux*. Il y a N-1 niveaux, et chacun de ces niveaux correspond à une salle d'attente. Plus précisément, à chaque niveau, *au moins* un thread doit pouvoir passer mais, si plusieurs threads souhaitent passer le même niveau, alors au moins un d'entre eux doit y rester bloqué. Le nombre de thread pouvant passer chaque niveau décroit donc strictement de 1 à chacun d'entre eux : N-1 threads peuvent passer le premier niveau, N-2 peuvent passer le deuxième niveau, et ainsi de suite jusqu'au dernier niveau, pour lequel un seul thread peut passer et ainsi accéder à sa section critique. La figure ci-dessous illustre le principe de l'algorithme du filtre.



La mise en œuvre de chaque niveau est une généralisation du principe de l'algorithme de Peterson pour deux threads : un thread donne, pour passer un niveau, d'abord la priorité aux autres threads avant de passer lui-même

soit si (1) il n'y a pas d'autre thread en attente ou (2) un thread arrivant après lui a donné la priorité. Par exemple, le thread T1 a pu avancer dans les niveaux 3 et plus car aucun thread n'était en attente au même niveau ou à un niveau supérieur. Le thread T5 est lui en attente au deuxième niveau car il s'y est déclaré comme la victime (et donc a donné la priorité aux autres threads alors en attente sur ce niveau). L'arrivée du thread T3 à ce niveau va amener T3 à se déclarer la victime à sa place, et permettre le progrès de T5 au niveau suivant, tandis que T3 restera bloqué. De la même façon, le progrès de T3 est rendu possible par l'arrivée du thread T8 au tout premier niveau, prenant la place de T3 en tant que victime pour ce niveau.

Une mise en œuvre de l'algorithme du filtre utilise deux tableaux partagés de taille N, initialisés comme suit :

```
#define N 8
int level[N];
int victim[N];

// Initialisation
for (int j=0; j<N; j++) {
  level[j]=0;
}
```

Un thread arrivant dans un nouveau niveau commence par se déclarer comme la *victime* pour ce niveau, puis consulte les niveaux auxquels les autres threads se trouvent, en consultant les tableaux partagés. Le code ci-dessous représente l'algorithme suivi par le thread *i*.

```
// Thread i
// Parcours des niveaux 1 à n-1
for (int L = 1; L < N; L++) {
  // Annoncer l'intention de rentrer au niveau L
  level[i] = L;
  // Le thread se désigne comme la victime pour ce niveau
  victim[L] = i;
  // Attendre tant qu'il existe au moins un thread au même niveau ou à un niveau supérieur,
  // et que le thread i est la victime du niveau où il se trouve
  do {
    int t_niv_sup_egal = 0;
    for (int j=0; j< N; j++) {
      // parcours du tableau des niveaux pour déterminer si un thread
      // est au même niveau ou à un niveau supérieur
      if ((j!=i) && level[j] >=L) {
        t_niv_sup_egal = 1;
      }
    }
  } while (t_niv_sup_egal && victim[L]==i);
}

section_critique();

// Libération de threads bloqués en attente dans les niveaux inférieurs
level[i]=0;
```

Un thread *i* arrivant dans un niveau ne va progresser au niveau suivant que lorsque l'un de ces deux conditions est remplie :

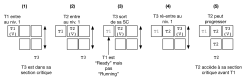
– La première condition est qu'il n'existe aucun thread en attente au même niveau ou dans un niveau supérieur. Cela est typiquement le cas lorsqu'aucun thread ne cherche à exécuter sa section critique. Le thread *i* va alors progresser dans les niveaux un à un en se déclarant comme la victime, puis en constatant que la voie est libre aux niveaux supérieurs.
– La seconde condition est qu'un autre thread soit arrivé au même niveau, permettant au premier de progresser. En effet, ce second thread aura alors positionné la case du tableau `victim[L]` à son identifiant, et devient de fait la victime, bloqué à ce niveau : au plus N-L threads pourront ainsi accéder au niveau L.

A la sortie de sa section critique un thread *i* va simplement indiquer qu'il relâche l'exclusion mutuelle en écrivant 0 dans `level[i]`, ce qui va libérer les threads en attente aux niveaux inférieurs.

Un problème d'équité

On peut observer que l'algorithme du filtre peut souffrir du problème suivant : un thread TA qui commence son parcours des niveaux avant un thread TB n'a aucune garantie qu'il pourra accéder à sa section critique avant celui-ci. Dans le pire des cas, le thread TA pourrait voir un nombre arbitraire de threads passer devant lui et accéder à leur section critique. On dit qu'un tel algorithme d'exclusion mutuelle ne respecte pas le principe d'équité.

Un exemple de la non équité de l'algorithme de filtre est donné par la figure ci-dessous pour une configuration simple où N le nombre maximal de threads est 3. Le filtre fait donc N-1=2 niveaux.



On suppose que le thread T3 est déjà dans sa section critique et que les threads T1 et T2 veulent aussi accéder à leur section critique. L'entrée dans le filtre pour T1 précède strictement l'entrée de T2. On voit que T1 se déclare comme la victime et reste bloqué au premier niveau (1). L'arrivée de T2 fait que ce dernier se déclare comme victime au niveau 1 à la place de T1 (2). T1 pourrait alors accéder au niveau 2, mais entre temps le thread est passé dans l'état Ready, i.e. le scheduler lui a dé-alloué le processeur qu'il occupait. Lorsque T3 termine sa section critique (3), T2 ne peut pas progresser car il est toujours bloqué au niveau 1 : sa condition `t_niv_sup_egal` est toujours vrai car T1 est présent au niveau 1 (et ce bien que T1 ne puisse pourtant pas s'exécuter et passer au niveau 2 tant que celui-ci n'a pas accès à un processeur). Considérons ensuite que T3 souhaite de nouveau accéder à sa section critique. T3 entre au niveau 1 et s'y déclare comme la victime (4). T2 peut ainsi passer au niveau 2, puis entrer dans sa section critique (5). Ainsi, on observe que T2 a pu accéder à sa section critique avant T1 bien que l'accès au filtre ait été fait après celui-ci.

La garantie d'équité pour l'accès à la section critique n'est pas toujours nécessaire et elle n'est pas toujours désirable d'un point de vue des performances. Ici, pour respecter l'ordre d'arrivée, il aurait été nécessaire de bloquer non seulement T2 mais aussi T3 tant que T1 n'a pas accès à un processeur pour exécuter sa section critique. Cette attente peut être significativement plus longue que le temps nécessaire à T2 et T4 pour parcourir les niveaux du filtre et exécuter plusieurs fois leur sections critiques.

6.8.3 Algorithme de la boulangerie (Bakery) de Lamport

L'algorithme de la boulangerie (Bakery algorithm) a été proposé par Leslie Lamport, un grand précurseur de l'étude formelle de la synchronisation des processus et par ailleurs auteur du logiciel Latex. Il permet de résoudre le problème de l'exclusion mutuelle avec des garanties d'équité.

Note : Définir la notion d'équité

On distingue dans un algorithme d'exclusion mutuelle tel que l'algorithme de Peterson ou le Bakery algorithm deux phases :
- Une première phase (doorway) pendant laquelle le thread configure des ressources (variables locales). Cette étape termine en un nombre de pas borné, i.e., elle ne comporte pas de boucles ;
- Une deuxième phase (waiting) pendant laquelle le thread vérifie de façon continue qu'une condition est vérifiée pour entrer dans sa section critique.

La garantie formelle d'équité stipule qu'un thread TA qui termine sa phase doorway avant le début de la phase doorway d'un thread TB a la garantie de pouvoir accéder à sa section mutuelle avant TB. Dans le cas où les deux phases doorway seraient concurrentes alors l'ordre d'accès à la section critique est arbitraire.

L'algorithme Bakery utilise un principe simple, qui est proche d'une situation de la vie courante dans un magasin (d'où son nom). Un thread souhaitant accéder à sa section critique obtient tout d'abord un numéro d'ordre, un peu comme la machine distribuant des tickets à l'entrée d'un magasin. Ensuite, ce thread attend que les threads avec un ticket de numéro plus élevé aient terminé leur section critique avant de pouvoir accéder à la sienne.

L'algorithme nécessite de connaître le nombre de threads N. Il utilise deux tableaux partagés :
- Le tableau `drapeau[]` contient des booléens (sous la forme de `int` en C, valant 0, faux ou 1, vrai). Les entrées de ce tableau indiquent la volonté de chacun des N threads d'entrer dans leur section critique ;

– Le tableau `ticket[]` contient le ticket de chaque thread intéressé dans la file d'attente, ou bien le précédent ticket lors de son dernier accès.

Les deux tableaux partagés sont définis et initialisés comme suit :

```
#define N 8
int drapeau[N];
int ticket[N];

// Initialisation
for (int j=0; j<N; j++) {
  drapeau[j]=0;
  ticket[j]=0;
}
```

Pour accéder à sa section critique, un thread va d'abord indiquer son intention en écrivant 1 (vrai) dans son entrée du tableau `drapeau[]`. Ensuite, il va lire l'ensemble des tickets des autres threads, et choisir un numéro qui est supérieur de 1 au numéro de ticket maximal. À la suite de la section critique, le thread remet simplement son drapeau à faux. Il n'est pas nécessaire de changer la valeur stockée dans `ticket[]` pour ce thread : par définition de l'équité, les threads en attente ont nécessairement un ticket de valeur plus élevée, ou, s'il n'y a pas de tel thread en attente, le thread arrivant plus tard obtiendra la valeur suivante.

```
// Thread i

// Section doorway : annoncer son intérêt et obtenir un ticket
drapeau[i]=1;
int t=0;
// Parcours des tickets
for (int j=0; j<N; j++) {
  if (ticket[j]>t) {
    t = ticket[j];
  }
}
// Prise du ticket supérieur
ticket[i]=t+1;

// Section waiting : attendre son tour ...
do {
  int mon_tour = 1;
  // Parcours des tickets des autres threads dont le drapeau est levé
  for (int j=0; j<N; j++) {
    if (drapeau[j]) {
      if (ticket[j] > ticket[i]) {
        // Il y a un autre thread actif devant dans la file ...
        mon_tour = 0;
      }
    }
  }
} while (!mon_tour);

section_critique();

// Libération de threads en attente avec les tickets suivants
drapeau[i]=0;
```

Si on analyse cet algorithme en faisant l'hypothèse que les sections doorway soient exécutés de façon non concurrente par les différents threads, celui-ci assure assez trivialement la propriété d'exclusion mutuelle ainsi que celle d'équité. Un seul thread, celui avec la valeur de ticket la plus élevée, peut exécuter sa section critique à la fois, et les threads exécutent leur section critique strictement dans l'ordre de leurs tickets. Toutefois, cette hypothèse est irréaliste : deux threads peuvent tout à fait exécuter les étapes de leur section doorway de manière concurrente. Lorsque c'est le cas, l'algorithme ci-dessus n'assure plus l'exclusion mutuelle : deux threads T1 et T2 peuvent ainsi observer les mêmes valeurs du tableau `ticket[]` et décider de prendre le même numéro de ticket, par

exemple 5. Lorsque le thread T3 avec le ticket de numéro 6 écrit `drapeau[3]=0` alors TA et TB observeront une file vide et accèderont simultanément à leur section critique !

Une solution pour pallier ce problème serait d'utiliser un mutex pour protéger l'accès au tableau ticket. Toutefois, cette solution pose un problème de poule et d'œuf (qui est arrivé le premier ?) : si il faut utiliser une primitive d'exclusion mutuelle pour mettre en œuvre un algorithme d'exclusion mutuelle c'est que ce deuxième n'apporte rien de plus ...

Une autre solution est de se fonder sur la notion d'équité, qui permet un ordre arbitraire pour les threads qui auraient effectué leur section doorway de manière concurrente. On peut alors fixer un tel ordre arbitrairement, et le plus simple est d'utiliser l'ordre des identifiants des threads. Dans notre exemple, entre T1 et T2 avec le même ticket 5, T2 est prioritaire sur T1 et ce dernier doit attendre la fin de sa section critique. On peut mettre en œuvre cette correction en remplaçant :

```
if (ticket[j] > ticket[i]) { ... }
```

Par :

```
if ((ticket[j] > ticket[i]) || ((ticket[j]==ticket[i]) && j>i)) { ... }
```

Note : Interaction entre scheduler et algorithme d'exclusion mutuelle

On voit ici les limites d'une mise en œuvre d'un algorithme d'exclusion mutuelle avec des garanties d'équité uniquement en espace utilisateur, par rapport à une mise en œuvre au niveau du noyau du système d'exploitation.

Imaginons que deux threads T1 et T2 sont en état Ready après avoir été interrompus lors de l'exécution de la section waiting (la répétition de la boucle `do {...} while();`). T1 a un ticket de valeur 24, et T2 un ticket de valeur 23. Si le scheduler passe T1 en Running en lui octroyant un processeur, T1 va s'exécuter et utiliser du temps processeur pour rien, puisqu'il sera en attente que T2 s'exécute et passe sa section critique. Le scheduler n'a pas de moyen, dans ce cas, de savoir que l'exécution de T2 est plus prioritaire que celle de T1 pour permettra à l'ensemble des threads de réaliser du progrès.

La mise en œuvre de l'exclusion mutuelle au niveau du noyau, en passant un thread souhaitant accéder à une section mutuelle bloquée en attente dans une file spécifique, permet de résoudre ce problème : les threads, comme T1, en attente d'une condition, ne sont pas en état Ready et on évite qu'ils exécutent de coûteuses boucles de vérification pour rien. Ceci est toujours bénéfique dans le cas d'un mono-processeur. Ce l'est aussi dans la majorité des cas sur un multi-processeur, mais pas toujours, comme nous le verrons plus tard dans ce chapitre.

6.8.4 Utilisation d'instructions atomiques

Sur les ordinateurs actuels, il devient difficile d'utiliser les algorithmes de Peterson, du filtre, ou de Bakery tel qu'ils ont été décrits précédemment et ce pour trois raisons.

Premièrement, ces algorithmes nécessitent de connaître le nombre de threads pouvant potentiellement accéder de façon concurrente à leur section critique. Ce nombre n'est pas toujours connu à l'avance, ce qui limite les possibilités de fournir des algorithmes génériques. Si on utilise un nombre maximal de threads comme une limite haute de la concurrence, alors le coût en mémoire (taille des tableaux partagés) et en temps d'exécution (e.g. parcours de ces tableaux pour l'algorithme Bakery ou parcours des filtres à différents niveaux pour l'algorithme du filtre) deviennent très importants.

Deuxièmement, les compilateurs C sont capables d'optimiser le code qu'ils génèrent. Pour cela, ils analysent le programme à compiler et peuvent supprimer des instructions qui leur semblent être inutiles. Dans le cas de l'algorithme de Peterson, le compilateur pourrait très bien considérer que la boucle `while` est inutile puisque les variables `turn` et `flag` ont été initialisées juste avant d'entrer dans la boucle.

La troisième raison est que sur un ordinateur multiprocesseur, chaque processeur peut réordonner les accès à la mémoire automatiquement afin d'en optimiser les performances [McKenney2005]. Cela a comme conséquence que certaines lectures et écritures en mémoires peuvent se faire dans un autre ordre que celui indiqué dans le programme sur certaines architectures de processeurs. Si dans l'algorithme de Peterson le thread `A` lit la valeur de `flag[B]` alors que l'écriture en mémoire pour `flag[A]` n'a pas encore été effectuée, une violation de la section

critique est possible. En effet, dans ce cas les deux threads peuvent tous les deux passer leur boucle `while` avant que la mise à jour de leur drapeau n'aie été faite effectivement en mémoire.

Note : Pourquoi `volatile` n'est pas suffisant

On notera que l'utilisation du mot clé `volatile` ne peut pas résoudre ces problèmes liés au réordonnancement des instructions par le compilateur ou le processeur. Le mot clé `volatile` permet d'assurer que l'accès à une variable partagé se fera toujours par une opération `mov` depuis l'adresse mémoire la contenant, et pas via un registre contenant la copie d'une lecture précédente. Il ne garantit pas, toutefois, que cet accès mémoire ne sera pas ré-ordonné par le processeur pour des raisons de performance. Un accès à une variable `volatile` peut tout à fait avoir lieu après un autre accès mémoire, ce dernier figurant pourtant avant l'accès à la variable partagée dans le programme. Il est possible de forcer le processeur à terminer les instructions d'accès à la mémoire en cours, avant de pouvoir en exécuter d'autres, désactivant de fait les optimisations utilisant le ré-ordonnancement des instructions. Cela requiert d'utiliser des opérations de barrières mémoires (memory fences) en plus de la déclaration comme `volatile` de la variable partagée. L'instruction `MFENCE` ordonne ainsi au processeur de terminer les opérations mémoires en cours, tandis que `LFENCE`, `SFENCE` permettent de terminer les opérations de lecture ou d'écriture, respectivement. L'utilisation correcte des barrières mémoires est très complexe en pratique. Elle est donc réservée pour du code de très bas niveau, par exemple dans les couches du noyau les plus proches du matériel.

La nécessité de fournir des primitives de synchronisation (entre autres, d'exclusion mutuelle) génériques et efficaces a amené les fabricants de processeurs à enrichir les jeux d'instructions avec des opérations spécifiques. Ces instructions mettent en œuvre des opérations atomiques. Une *opération atomique* est une opération qui, lorsqu'elle est exécutée sur un processeur, ne peut pas être interrompue par l'arrivée d'une interruption. Ces opérations permettent généralement de manipuler en même temps un registre et une adresse en mémoire. En plus de leur caractère ininterruptible, l'exécution de ces instructions atomiques par un ou plusieurs processeur implique une coordination des processeurs pour l'accès à la zone mémoire référencée dans l'instruction. Tous les accès à la mémoire faits par ces instructions sont ordonnés par les processeurs de façon à ce qu'ils soient toujours visibles par tous les processeurs dans le même ordre (e.g. si un processeur A voit les opérations atomiques op1, op2, op3 dans cet ordre, alors c'est le cas de tous les autres processeurs, ce qui n'est pas nécessairement le cas pour les accès mémoires traditionnels).

Plusieurs types d'instructions atomiques sont supportés par différentes architectures de processeurs. A titre d'exemple, considérons l'instruction atomique `xchg` qui est supportée par les processeurs [IA32]. Cette instruction permet d'échanger, de façon atomique, le contenu d'un registre avec une zone de la mémoire. Elle prend deux arguments, un registre et une adresse en mémoire. Ainsi, l'instruction `xchgl %eax, (var)` est équivalente aux trois instructions suivantes, en supposant le registre `%ebx` initialement vide. La première instruction sauvegarde dans `%ebx` le contenu de la mémoire à l'adresse `var`. La deuxième instruction copie le contenu du registre `%eax` à cette adresse mémoire et la dernière instruction transfère le contenu de `%ebx` dans `%eax` de façon à terminer l'échange de valeurs.

```
movl (var), %ebx
movl %eax, (var)
movl %ebx, %eax
```

Avec cette instruction atomique, il est possible de résoudre le problème de l'exclusion mutuelle en utilisant une zone mémoire, baptisée `lock` dans l'exemple. Cette zone mémoire contiendra la valeur `1` ou `0`. Elle est initialisée à `0`. Lorsqu'un thread veut accéder à sa section critique, il exécute les instructions à partir de l'étiquette `enter:`. Pour sortir de section critique, il suffit d'exécuter les instructions à partir de l'étiquette `leave:`.

```
lock:                   ; étiquette, variable
  .long    0            ; initialisée à 0

enter:
   movl    $1, %eax     ; %eax=1
   xchgl   %eax, (lock) ; instruction atomique, échange (lock) et %eax
                        ; après exécution, %eax contient la donnée qui était
                        ; dans lock et lock la valeur 1
   testl   %eax, %eax   ; met le flag ZF à vrai si %eax contient 0
   jnz     enter        ; retour à enter: si ZF n'est pas vrai
```

```
    ret

leave:
    movl    $0, %eax      ; %eax=0
    xchgl   %eax, (lock)  ; instruction atomique
    ret
```

Pour bien comprendre le fonctionnement de cette solution, il faut analyser les instructions qui composent chaque routine en assembleur. La routine `leave` est la plus simple. Elle place la valeur `0` à l'adresse `lock`. Elle utilise une instruction atomique de façon à garantir que cet accès en mémoire se fait séquentiellement [19].. Lorsque `lock` vaut `0`, cela indique qu'aucun thread ne se trouve en section critique. Si `lock` contient la valeur `1`, cela indique qu'un thread est actuellement dans sa section critique et qu'aucun autre thread ne peut y entrer.

Pour entrer en section critique, un thread doit d'abord exécuter la routine `enter`. Cette routine initialise d'abord le registre `%eax` à la valeur `1`. Ensuite, l'instruction `xchgl` est utilisée pour échanger le contenu de `%eax` avec la zone mémoire `lock`. Après l'exécution de cette instruction atomique, l'adresse `lock` contiendra nécessairement la valeur `1`. Par contre, le registre `%eax` contiendra la valeur qui se trouvait à l'adresse `lock` avant l'exécution de `xchgl`. C'est en testant cette valeur que le thread pourra déterminer si il peut entrer en section critique ou non. Deux cas sont possibles :

1. `%eax==0` après exécution de l'instruction `xchgl %eax, (lock)`. Dans ce cas, le thread peut accéder à sa section critique. En effet, cela indique qu'avant l'exécution de cette instruction l'adresse `lock` contenait la valeur `0`. Cette valeur indique que la section critique était libre avant l'exécution de l'instruction `xchgl %eax, (lock)`. En outre, cette instruction a placé la valeur `1` à l'adresse `lock`, ce qui indique qu'un thread exécute actuellement sa section critique. Si un autre thread exécute l'instruction `xchgl %eax, (lock)` à cet instant, il récupèrera la valeur `1` dans `%eax` et ne pourra donc pas entrer en section critique. Si deux threads exécutent simultanément et sur des processeurs différents l'instruction `xchgl %eax, (lock)`, la coordination des accès mémoires entre les processeurs garantit que ces accès mémoires seront séquentiels (l'un précédera l'autre). Le thread qui bénéficiera du premier accès à la mémoire sera celui qui récupèrera la valeur `0` dans `%eax` et pourra entrer dans sa section critique. Le ou les autres threads récupéreront la valeur `1` dans `%eax`.
2. `%eax==1` après exécution de l'instruction `xchgl %eax, (lock)`. Dans ce cas, le thread ne peut entrer en section critique et il entame une boucle active durant laquelle il va continuellement exécuter la boucle `enter: movl ... jnz enter`.

6.8.5 Verrous par attente active (spinlocks)

L'ensemble des algorithmes d'exclusion mutuelle que nous avons vu dans ce chapitre utilisent le principe de l'attente *active*. On les appelle des *spinlocks* en anglais, car un thread en attente, pour entrer dans sa section critique, boucle (*spin*) sur la vérification d'une condition. Par exemple, dans l'algorithme Bakery un thread boucle sur le parcours des deux tableaux partagés. Dans l'algorithme utilisant l'opération atomique `xchgl` ci-dessus, un thread bouclera sur la suite d'instruction entre l'adresse `enter` et l'instruction `jnz`. L'exclusion mutuelle par attente active est mise en œuvre seulement en mode utilisateur. Elle ne nécessite pas de support spécifique du système d'exploitation.

Les mutex et les sémaphores POSIX que nous avons vu dans les chapitres précédents sont eux, au contraire, mis en œuvre avec le concours du système d'exploitation. Un thread qui tente d'acquérir un sémaphore non disponible appèle la fonction `wait()` de la librairie. Comme le sémaphore n'est pas disponible, ceci générera une interruption qui passe le contrôle au système d'exploitation. Celui-ci passera le statut de ce thread en mode Blocked, en l'ajoutant à une file d'attente liée à ce mutex. Lorsqu'un autre thread appelle `post()` sur ce même sémaphore, un thread de la file d'attente est mis en mode Ready, mais n'obtient pas nécessairement immédiatement un processeur. C'est le scheduler qui décidera de placer plus tard ce thread sur un processeur (mode Running).

Les mutex et sémaphores POSIX sont toujours avantageux dans un contexte mono-processeur : il n'y a pas d'utilité pour un thread sur un seul processeur de boucler sur l'attente d'une condition qui ne peut se réaliser que si un autre thread obtient le processeur à sa place. Autant redonner la main immédiatement au système d'exploitation et attendre que la condition soit réalisée pour être de nouveau placé en état Ready. Leur autre avantage, que

19. En réalité, l'utilisation d'une opération atomique n'est pas strictement nécessaire ici, car on aurait pu utiliser deux barrières mémoires, une avant et une après l'écriture de 0 à l'adresse (lock) avec une instruction `mov`. L'utilisation d'une instruction atomique a l'avantage de la simplicité.

ce soit dans un système mono ou multi-processeur, est que les threads ne perdent pas de temps et de ressources processeur à effectuer leur attente active. Comme la gestion des files de threads en attente est du ressort du système d'exploitation, celui-ci peut prendre des décisions de meilleure qualité quand le scheduler choisit quel thread associer à un processeur : les threads en attente ne sont pas considérés, contrairement à une attente active où le scheduler ne peut pas savoir quel thread a une chance d'obtenir le verrou ou non (cf. encart plus haut dans ce chapitre).

Toutefois, les mutex et sémaphores POSIX mis en œuvre avec le concours du noyau ont un désavantage important qui est la latence de leurs opérations. Cette latence est due à plusieurs facteurs : (1) la combinaison entre le surcoût de la mise en attente (état Blocked) d'un thread appelant `lock()` sur un mutex, dans une file gérée par le noyau, (2) le temps qui s'écoule entre le moment où ce thread sera placé en état Ready (par un appel à `unlock()`) et celui où il sera de nouveau placé par le scheduler sur un processeur, et (3) le coût des différents changements de contexte au cours de cette procédure. Le temps total pour ces opérations peut être plus long que la durée de la section critique du thread qui détenait le mutex. Dans ce cas, il peut être plus pertinent d'effectuer une attente active courte en attenant la fin de cette section critique. L'attente active est ainsi souvent privilégiée au niveau de l'implémentation du noyau lui même, et dans la mise en œuvre des structures de données (tables de hachage, arbres, graphes, etc.) utilisées de façon concurrente par plusieurs threads.

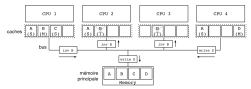
6.8.6 Attente active et performance

Bien que la mise en œuvre de l'attente active avec l'instruction atomique `xchgl` soit correcte (i.e. elle préserve la propriété d'exclusivité mutuelle) et qu'elle ne présente pas les défauts des algorithmes historiques fondés sur des lectures et écritures vus précédemment, elle n'est pourtant pas vraiment satisfaisante.

En effet, on constate rapidement que son utilisation entraîne une dégradation des performances, même pour synchroniser l'accès à quelques sections critiques dans un programme (comme par exemple l'accès à une structure de donnée globale partagée). Cette dégradation intervient à la fois pour la performance de l'accès aux sections critiques elles-même (augmentation significative de la latence) mais aussi, ce qui est encore plus ennuyeux, pour les autres threads du même programme effectuant des opérations hors de leur section critique, et même pour les autres applications présentes sur la même machine. Pour comprendre la raison de ces dégradations, nous devons revenir sur le concept de cache et en particulier sur la gestion de la cohérence de ces caches dans un système multi-processeur.

L'utilisation de caches est fondamentale pour la performance : elle permet de masquer la latence d'accès de la mémoire principale. Dans un système multi-processeur, chaque processeur (ou chaque cœur) possède son propre cache, qui conserve les données fréquemment accédées par ce processeur. La valeur stockée à une adresse donnée peut être mise à jour dans le cache d'un processeur, avant d'être propagée plus tard vers la mémoire principale (write-back). Une même adresse peut être présente dans le cache de *plusieurs* processeurs. Il est donc nécessaire de synchroniser les caches des différents processeurs pour éviter que deux processeurs puissent écrire de façon concurrente à la même adresse en mémoire. C'est le rôle d'un protocole de cohérence de caches.

Il existe de nombreux protocoles de cohérence de cache mais nous nous contenterons de présenter le plus simple d'entre eux, le protocole MSI. Chaque processeur est connecté à un bus, auquel est aussi connecté la mémoire principale. Chaque contrôleur de cache (attaché à chaque processeur), ainsi que la mémoire principale, écoute ce bus en permanence. Un seul processeur peut envoyer un message à un moment donné sur le bus, et tous les messages sont vus par tous les processeurs (et la mémoire). La figure suivante présente un exemple avec 4 mots mémoires : A, B, C et D et 4 processeurs avec un cache de 4 entrées.



Une entrée d'un cache peut prendre trois états possibles, M, S, ou I :
– Une entrée dans l'état **M** (Modified) contient une valeur qui est plus récente que celle dans la mémoire principale, et seul ce processeur a une copie de cette entrée. Dans certaines architectures, cet état est aussi appelé **E** (Exclusive).
– Une entrée dans l'état **S** (Shared) est partagée entre plusieurs caches, et la valeur stockée dans les différents caches est la même.

– Une entrée dans l'état **I** (Invalid) est invalide et ne peut plus être utilisée sans récupérer la dernière valeur associée à cette entrée.

Avant de passer une entrée en état M, il est nécessaire d'invalider les entrées pour la même adresse dans les caches des autres processeurs. Par exemple, considérons que le processeur 1 possède l'adresse B dans son cache, ce qui est le cas des processeurs 2 et 3. Ces entrées sont initialement en mode S. Avant de pouvoir écrire dans son cache, le processeur 1 doit d'abord invalider les entrées correspondantes dans les caches des autres processeurs. À cet effet, le processeur 1 envoie un message `inv B` (invalider l'entrée B) sur le bus. Ce message est intercepté par les contrôleurs des cache ayant une copie de B, en l'occurence ceux des processeurs 2 et 3, et leur copie est déclarée comme invalide (I).

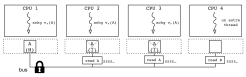
Le bus est aussi utilisé par les contrôleurs de cache pour écrire les données vers la mémoire, mais aussi pour répondre aux requêtes de lecture des autres processeurs. Si le processeur 4 émet un message pour lire la valeur à l'adresse C dans notre exemple, c'est le contrôleur de cache du processeur 1 qui répondra plutôt que la mémoire principale.

Si deux threads sur deux processeurs différents écrivent de façon répétée à la même adresse mémoire, il y a un risque qu'ils invalident en permanence les entrées de cache de leurs processeurs respectifs. Imaginons qu'un thread sur le processeur 2 modifie A très souvent, et qu'un autre thread sur le processeur 4 modifie aussi A très souvent. Les deux processeurs vont passer leur temps à attendre que les entrées de leur cache respectifs soient invalidées. Cela va générer un nombre important de messages sur le bus (dont la capacité est bien entendu limitée). On appelle ce phénomène l'effet *ping-pong* [20].

Instructions atomiques et utilisation du bus

Les instructions atomiques comme `xchgl` impliquant une adresse mémoire doivent non seulement obtenir l'accès en mode M à ce mot mémoire dans leur cache, mais aussi bloquer l'accès au bus par les autres processeurs pour empêcher son invalidation par un autre processeur pendant l'exécution de l'instruction atomique. Ce blocage du bus a un impact important sur la performance de l'ensemble des threads en cours d'exécution : les accès mémoire sont bloqués pour tous les processeurs en attendant que l'opération atomique sur un de ces processeurs soit terminée.

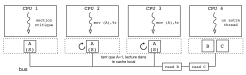
Le coût du blocage du bus explique la performance décevante du code présenté précédemment, où chaque thread souhaitant accéder à sa section critique utilise une succession d'opérations `xchgl` continue. La situation est créée est illustrée par la figure ci-dessous.



Trois des processeurs (CPU1, 2 et 3) hébergent des threads souhaitant accéder à leur section critique, en essayant d'échanger leur valeur avec l'adresse mémoire A. Le thread sur le CPU1 a accès à l'adresse mémoire en mode M dans son cache et verrouille l'utilisation du bus. Pendant ce temps, les opérations `xchgl` des threads sur les processeurs 2 et 3 sont bloquées en attendant la disponibilité du bus. Par ailleurs, le thread du processeur 4, indépendant des threads souhaitant accéder à leur section critique, voit ses opérations d'accès à la mémoire temporairement bloquées elles aussi. Une fois que le thread du processeur 1 aura gagné l'accès à sa section critique, il sera lui aussi sujet à un ralentissement car ses accès mémoire seront en concurrence avec les opérations `xchgl` continues des threads des processeurs 2 et 3. Le retard pris dans la section critique est, au final, au désavantage de ces derniers : avec les opérations atomiques en continu, ils retardent leur propre accès à leur section critique !

Une solution simple pour pallier ce problème est de tirer parti du cache. Tant que la valeur du verrou est à 1, on sait qu'un thread est dans sa section critique et il est inutile de tenter continuellement d'effectuer l'opération `xchgl` : celle ci ralentit le progrès de tous les threads et renverra toujours 1, entrainant une nouvelle boucle. À la place, il est préférable de cacher la valeur du verrou dans le cache, et de la lire continuellement tant que celle ci vaut 1. Cette situation est illustrée par la figure suivante.

20. Il existe aussi un autre phénomène ping-pong qui est dû au problème du faux partage : comme les lignes de cache ont une granularité qui est plus grande (par exemple 64 octets) que celle de la plupart des variables utilisées (int, long, etc.), deux variables utilisées par deux threads différents peuvent se retrouver sur la même ligne de cache. L'accès par l'un des threads invalidera la ligne de cache complète sur l'autre processeur, grevant les performances sans qu'il y ait de de données réellement partagées.



On appelle souvent cette solution le test-and-test-and-test, par opposition à la version de l'algorithme précédent qui est appelé test-and-set. L'idée est de ne tenter l'opération atomique `xchgl`, qui teste et assigne (set) la valeur de façon atomique, que lorsque le verrou semble libre. Cette situation apparait lorsque le thread qui effectuait sa section critique écrit 0 dans le verrou : l'adresse devient en état M dans son cache après invalidation dans le cache des autres processeurs, qui lisent alors 0 grace à une requête sur le bus. Le pseudocode de cette opération peut être :

```
while (test_and_set(verrou, 1)) {
  // on a pas obtenu le verrou car on a lu 1
  // donc on attend de lire 0 pour tenter à nouveau
  while (verrou) {}
}
```

Le test-and-test-and-set limite fortement la contention sur le but pendant l'exécution de la section critique du thread possédant le verrou. Elle n'est toutefois pas parfaite quand il y a un grand nombre de threads qui souhaitent accéder à leur section critique de façon répétée : à chaque fois qu'un de ces threads quitte sa section critique, l'invalidation de l'adresse du verrou dans les caches de tous les autres processeurs entraine une rafale de lecture de la nouvelle valeur, suivi d'une rafale d'opérations `xchgl`. De toutes ces opérations, une seule sera couronnée de succès, mais toutes devront bloquer le bus et limiter le progrès général du système. Le caractère synchrone et répété de ces pics d'occupation du bus peut réduire la performance générale du système, sans toutefois le faire autant que le test-and-set seul.

On peut encore améliorer cette solution en constatant qu'un thread qui lit la valeur 1 pour le verrou est probablement dans une situation de contention et peut se mettre en attente pour un moment avant de réessayer. On appelle cette attente un back-off en anglais, que l'on peut traduire par une mise en retrait. Un thread souhaitant entrer dans sa section critique mais observant la valeur 1 pour le verrou va effectuer une attente avant de réessayer. L'avantage de cette approche est qu'elle permet de s'adapter à la contention, et qu'elle permet de désynchroniser les demandes des différents threads, évitant ainsi les pics de requêtes de la solution test-and-test-and-set :

– Le premier avantage, l'adaptation à la contention, est obtenu en augmentant le temps d'attente à chaque essai infructueux. La première attente est typiquement assez courte, la deuxième est deux fois plus longue, la troisième quatre fois, etc. tout en bornant bien entendu l'attente à un maximum.
– Le deuxième avantage, l'évitement des pics de requêtes, est obtenu en ajoutant une part d'aléatoire dans le choix du délai d'attente. Le délai d'attente effectif est une portion aléatoire du temps maximal d'attente à chaque essai. Par exemple, si le délai minimal est de 10ns, alors la première attente pourrait être de 6ns, la deuxième (entre 0 et 20ns), de 17ns, et ainsi de suite.

La définition du pseudo-code de ce dernier algorithme, que l'on peut appeler backoff-test-and-test-and-set, est laissé en exercice.

Ordonnancement (scheduling)

7.1 Ordonnancement (Scheduling)

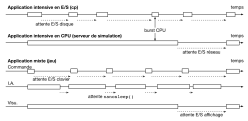
Nous avons vu dans le chapitre précédent qu'un système d'exploitation comme Linux pouvait supporter de nombreux threads (appartenant à divers processus) avec un nombre limité (ou même unique) de processeur(s). Un processeur n'exécute pourtant qu'un seul thread à la fois. Le partage des processeurs est rendu possible par un mécanisme de *partage de temps* : le système d'exploitation peut basculer de l'utilisation d'un processeur par un thread à une utilisation par un autre thread. L'enchaînement rapide de l'exécution des différents threads sur les processeurs donne l'illusion à l'utilisateur que ceux-ci s'exécutent simultanément. Le mécanisme permettant le partage de temps est le changement de contexte. Nous l'avons décrit dans les chapitres précédents.

La mise en œuvre du partage de temps illustre bien le principe de séparation entre mécanisme et politique, un objectif important poursuivi lors de la conception d'UNIX et de Linux. Le *mécanisme* de changement de contexte permet en effet d'assurer la transition entre les threads sur un ou plusieurs processeur(s), mais il ne dit pas quel thread doit être privilégié pour obtenir un processeur, ou quand un thread utilisant un processeur doit le libérer pour laisser la place à un autre thread. Ces décisions sont du ressort de la *politique* d'ordonnancement ou **scheduler**. L'avantage de séparer les décisions sur l'accès aux processeurs pour les différents threads, du mécanisme permettant d'acter ces décisions, est sa grande flexibilité. Un même mécanisme peut être utilisé avec des politiques différentes sur des systèmes aussi dissemblables qu'un super-calculateur ou une montre connectée, pour prendre des exemples de systèmes utilisant le noyau Linux.

7.1.1 Modèle d'exécution des threads et bursts CPU

Un thread exécute ses instructions par phases, alternant deux types d'opérations :
– Lorsqu'il obtient le processeur, un burst CPU (*rafale d'utilisation du processeur* en français) pendant lequel le processeur exécute des instructions du thread de façon continue ;
– Ce burst CPU se termine par l'utilisation d'une opération bloquante, comme une demande d'entrée/sortie ou une opération de synchronisation.
À la suite de l'appel bloquant, le thread ne peut pas faire de progrès tant que le résultat de l'opération n'est pas disponible.

La longueur des burst CPUs, et la fréquence des opérations bloquantes comme les entrée/sorties, peut varier fortement d'une application à l'autre. Ceci est illustré par la figure suivante.



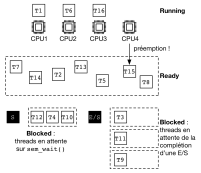
Dans cet exemple, une application de copie de fichier comme cp(1) effectue de nombreuses opérations d'entrée/sortie pour lire et écrire un fichier à copier en utilisant le système de fichiers, entremêlées de bursts CPU courts. Une application de calcul numérique présentera, au contraire, des bursts CPU très longs avec des entrée/sorties seulement au début et à la fin des calculs.

Un application peut tout à fait être composée de plusieurs threads présentant des caractéristiques différentes. Par exemple, dans un jeu, le thread chargé de prendre en compte les commandes du joueur (à l'aide du clavier ou d'une manette) présentera souvent des bursts CPU courts et de longues périodes d'attente, tandis que le thread en charge de l'intelligence artificielle du jeu pourra avoir des bursts CPU périodiques mais de durée régulière. Enfin, le thread en charge de l'affichage pourrait utiliser des bursts CPU longs pour préparer la visualisation d'une scène suivie de son envoi au dispositif d'affichage.

L'alternance entre les bursts CPU et les phases d'attente est mise en œuvre par l'alternance de chaque thread entre différents états, permis par le mécanisme de changement de contexte.

7.1.2 Évolution de l'état des threads

Le kernel maintient des structures de données à propos de tous les threads actifs. Comme nous l'avons vu précédemment, un thread peut être dans l'un des trois états suivants :
– **Running** : ce thread est *en ce moment* en train d'utiliser un des processeurs pour exécuter des instructions ;
– **Ready** : ce thread est *en attente* d'un processeur pour exécuter des instructions ;
– **Blocked** : ce thread ne peut pas s'exécuter pour l'instant car il est en attente d'une valeur de retour pour un appel système bloquant.
Un système fictif avec 16 threads et 4 processeurs est présenté dans l'illustration suivante.



Passage de l'état Running à l'état Blocked

Le passage de l'état Running à l'état Blocked advient lors de l'exécution d'un appel système bloquant appelé par le thread. Cet appel système bloquant peut être, par exemple, une demande d'entrée/sortie (écrire ou lire depuis le système de fichiers) ou un appel à une primitive de synchronisation comme par exemple pthread_mutex_lock(3posix) ou sem_wait(3posix). Les threads en état Blocked sont associés à une structure de donnée du noyau, qui joue le rôle de *salle d'attente*. Certaines de ces structures d'attente n'ont d'utilité que pour un seul thread, par exemple lorsque ce thread a demandé une lecture vers le système de fichiers. D'autres peuvent contenir plusieurs threads en attente. C'est le cas, par exemple, d'une structure d'attente pour un sémaphore. Il peut y avoir effectivement plusieurs threads ayant appelé sem_wait(3posix) (T12, T4 et T10). Un appel à sem_post(3posix) va libérer l'un de ces threads, qui passera alors en état Ready.

Note : Pas de garantie d'ordre sur le passage de l'état Blocked à l'état Ready !

De manière générale, on ne peut pas faire d'hypothèse sur l'ordre dans lequel les threads en attente dans une structure d'attente commune vont être sélectionné pour passer dans l'état Ready, lorsque la condition d'attente sera remplie. Par exemple, si plusieurs threads appelent sem_wait(3posix) sur le même sémaphore S dans un ordre donné, par exemple T12, puis T4, puis T10, il n'y a pas de garantie que lors des appels à sem_post(3posix) par d'autres threads ceux-ci passent en état Ready dans ce même ordre : le premier appel à sem_post(3posix) peut tout à fait passer T4 ou T10 en mode Ready avant T12.

Passage de l'état Running à l'état Ready

Un thread passe de l'état Running à l'état Ready lorsqu'il libère le processeur sur lequel il exécute actuellement des instructions.

On observe qu'avec uniquement les mécanismes définis précédemment, un thread qui ne génère aucun appel système pourrait rester dans l'état Running indéfiniment. C'est le cas, par exemple, d'un thread bloqué dans une boucle infinie ne comportant pas d'appel à la librairie standard. Si tous les processeurs venaient à être bloqués par des threads dans cette situation, alors la machine devient inutilisable. Par ailleurs, sans même considérer des boucles infinies, le temps d'occupation du processeur par le thread en cours d'exécution (son CPU burst) pourrait être particulièrement long, ce qui peut être problématique lorsque d'autres threads sont sujets à des contraintes de réactivité (par exemple, la réaction aux commandes utilisateurs ou la visualisation).

Les systèmes comme Linux utilisent donc une source d'interruption matérielle périodique (une horloge système) pour permettre de redonner le contrôle au système d'exploitation. À l'occasion de ces traitements d'interruption, il est possible de reprendre un processeur à un thread en état Running, en provoquant un changement de contexte. On dit alors que le thread a subit une **préemption**. C'est le cas de T15 sur notre exemple.

Passage de l'état Ready à l'état Running

La dernière transition consiste à restaurer l'état précédemment sauvegardé d'un thread en état Ready sur un processeur, et à reprendre son exécution.

7.1.3 Mise en œuvre du scheduler

La politique d'ordonnancement, que nous appellerons par la suite uniquement le *scheduler* par simplicité, est donc en charge de la prise de décision aux deux moments suivants :

– 1. Lorsqu'un processeur devient disponible, suite au passage d'un thread en mode Blocked, le scheduler doit sélectionner un thread dans l'état Ready et le promouvoir à l'état Running sur ce processeur.
– 2. Lorsqu'une interruption périodique est traité, le scheduler doit décider si un thread actuellement en état Running doit être préempté pour passer en état Ready.

Un scheduler qui prend des décisions pour les deux occasions (1) et (2) est dit préemptif (car il utilise la préemption d'un thread pour récupérer le processeur avant la fin de son CPU burst). Un scheduler qui ne prend de décision que lors de l'occasion (1) est non-préemptif. Il dépend d'appels réguliers par les threads à des appels systèmes bloquants, mais les threads ont la garantie que leurs CPU burst ne seront pas interrompus.

Objectifs

Il n'existe pas de scheduler parfait convenant à toutes les applications. Pour s'en convaincre, considérons les deux applications que sont la copie de fichier et l'application de calcul dans notre exemple précédent.

La priorité de l'application de copie de fichier est de subir le moins d'attente possible entre la disponibilité d'une valeur de retour d'un appel système vers le système de fichier, et l'envoi du prochain appel système pour continuer la copie, et éviter de ralentir l'opération de copie dans son ensemble. Pour ce thread, le délai d'attente entre sa mise en état Ready et l'obtention d'un processeur doit être la plus faible possible.

Pour l'application de calcul, le plus important est de pouvoir exécuter les instructions du long CPU burst avec le moins d'interruptions possibles. En effet, un changement de contexte est du temps perdu pour réaliser des opérations utiles (i.e., progresser dans la simulation). Par ailleurs, un thread qui est interrompu et replacé plus tard sur le processeur sera soumis à un phénomène de *cache froid* : les données qui étaient dans le cache, et donc accessibles avec un temps d'accès faible avant le changement de contexte, ont pu être remplacées par des données à des adresses différentes, utilisées par le thread qui a obtenu le processeur entre temps. Peupler de nouveau le cache avec les données nécessaire au calcul peut nécessiter de coûteux accès en mémoire principale et ralentir l'exécution.

Si l'on décide de privilégier l'application de copie, il est souhaitable d'interrompre le thread de l'application de calcul, mais cela va au détriment de ce dernier. À l'inverse, si on choisit de privilégier l'opération de calcul, alors l'opération de copie sera ralentie.

On peut définir cinq principaux critères pour mesurer la performance d'un scheduler :
– Du **point de vue du système** dans son ensemble tout d'abord :

– On veut pouvoir maximimiser l'utilisation du ou des processeur(s), c'est à dire la proportion du temps où ceux-ci exécutent des instructions des applications. Les opérations de changement de contexte ne sont évidemment pas considérées comme du travail utile pour ce critère.
– On peut vouloir maximiser le débit applicatif, c'est à dire le nombre de processus qui peuvent terminer leur exécution en une unité de temps donné (par exemple en une heure).

– D'autres critères sont applicables, cette fois-ci **du point de vue de chaque application** individuellement. On pourra par ailleurs s'intéresser à la distribution de ces métriques pour l'ensemble des applications, afin de savoir s'il existe un déséquilibre entre la métrique telle que perçue par une application et la même métrique perçue par une autre application :

– Une application peut souhaiter minimiser son temps total d'exécution, entre la création du processus et sa terminaison. Ce critère n'est pas nécessairement valide pour tous les types d'applications, par exemple il n'a que peu de sens pour une application interactive (par exemple, un shell), mais il est important pour des applications de calcul ou l'exécution d'un script par exemple.
– Ensuite, une application peut souhaiter minimiser le temps d'attente moyen, c'est à dire le temps écoulé entre la mise en état Ready (par exemple après la fin d'une entrée/sortie) et l'obtention d'un processeur. Cette métrique est particulièrement importante pour les applications interactives, comme un jeu ou une interface graphique.
– Enfin, une application voudra minimiser son temps de réponse, qui correspond à la somme entre le temps d'attente et le temps nécessaire pour terminer l'exécution de son burst CPU.

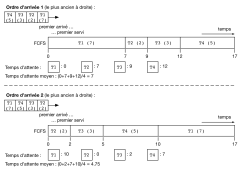
Nous allons dans la suite de ce chapitre décrire plusieurs scheduler classiques, en commençant par les scheduler non préemptifs, puis les schedulers préemptifs, et enfin les schedulers hybrides combinant plusieurs stratégies.

Note : Nous considérerons pour la présentation des schedulers uniquement le cas d'un seul processeur pour des raisons de simplicité, mais les algorithmes présentés ici peuvent être aisément étendu pour fonctionner avec plusieurs processeurs.

Le scheduler FCFS (First-Come-First-Serve)

Une première approche est d'exécuter les CPU bursts des threads dans l'ordre exact dans lequel ils ont obtenu l'état Ready (premier arrivé, premier servi). Ce scheduler n'étant pas préemptif, chaque CPU burst s'exécute intégralement avant de libérer le processeur pour un autre thread. Le temps de réponse avec un scheduler non préemptif est toujours égal au temps d'attente plus la durée du burst CPU, donc nous nous intéresserons principalement à ce premier critère.

L'exemple ci-dessous montre deux exécutions possibles pour 4 threads disponibles en état Ready simultanément, mais pour lesquels l'ordre d'ajout en état Ready a été effectué dans un ordre différent (T1, T2, T3, puis T4 dans un premier cas ; et T2, T3, T4 puis T1 dans le deuxième cas).

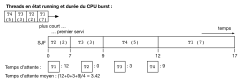


Ces figures présentent des diagrammes de Gantt, où le temps d'exécution de chaque CPU burst est représenté au cours du temps. En terme de débit applicatif et d'utilisation du processeur, cet algorithme est optimal, car il n'y a que trois changements de contexte : le temps perdu pour ces changements de contexte est donc minimal.

En revanche, si on considère le temps d'attente moyen pour chacun des threads, on observe que celui-ci diffère grandement entre le premier ordre d'arrivée et le second (de 7 unités de temps à 4.75 unités de temps). La raison est que dans la première configuration des CPU bursts courts (typiques des applications interactives ou utilisant de nombreuses entrées/sorties) se retrouvent *coincées* derrière un CPU burst long. Ce phénomène est appelé l'*effet convoi* (convoy effect en anglais). Il pénalise principalement les applications ayant des besoins d'interactivité.

Le scheduler SJF (Shortest Job First)

Le scheduler SJF (Shortest Job First) est un scheduler non préemptif qui a pour objectif de prévenir l'effet convoi. Lorsque plusieurs threads sont disponibles, le thread choisi est celui qui a le CPU burst à venir qui est le plus court. La figure ci-dessous montre le diagramme de Gantt où les threads obtiennent le processeur dans l'ordre du plus courte au plus long CPU burst.



On peut facilement montrer que le temps d'attente *moyen* avec le scheduler SJF est le meilleur possible : toute permutation d'ordre ne peut qu'augmenter ce temps d'attente moyen. Toutefois, cet algorithme n'est pas réalisable en pratique et ne peut donc servir que de mètre étalon pour analyser la performance d'autres algorithmes. Il n'est en effet pas possible de connaître à l'avance la durée d'un CPU burst, car celle-ci dépend de l'exécution du code du thread, et donc de ses boucles, conditionnelles, appels de fonctions, etc.

En revanche, il est possible de tenter d'approcher cet algorithme en partant de l'observation suivante : la durée des CPU bursts pour un thread donné a souvent tendance à être régulière dans le temps. C'est à dire qu'un thread utilisant le CPU pour de courtes périodes de temps régulièrement aura souvent tendance à répéter ce comportement (c'est le cas, par exemple, des trois threads du jeu présenté précédemment). À l'inverse, un thread utilisant régulièrement le CPU pour de longues périodes de temps sera souvent (mais pas toujours) plus susceptible d'avoir un prochain CPU burst qui sera long.

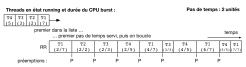
Un scheduler estimant SJF pourrait ainsi conserver dans une structure de données la durée des x derniers CPU bursts de chaque thread. En appliquant une moyenne sur cette durée, le scheduler peut alors tenter de prédire la durée du prochain CPU burst, et choisir le thread dont la durée prédite est la plus courte.

On note toutefois que, si SJF est optimal en terme de temps d'attente moyen, il n'offre que peu de propriétés d'équité. Si il existe de nombreux threads avec des CPU bursts à venir courts (ou prédits comme tels) alors un thread avec un CPU burst long (ou prédit comme tel) pourrait ne jamais avoir accès au processeur, ou bien n'y avoir accès que bien plus tard.

Le scheduler préemptif RR (Round Robin)

Un scheduler préemptif peut choisir de *préempter* un thread en cours d'exécution sur un processeur, c'est à dire de passer ce thread en état Ready pour libérer le processeur pour un autre thread. Une décision de préemption peut être prise lorsque le système d'exploitation reprend la main sur le processeur lors de l'arrivée d'une interruption. Une horloge système dédiée à cet usage génère une interruption matérielle (tick) de manière périodique.

Un premier scheduler préemptif est le scheduler RR (Round Robin), expression anglaise que l'on pourrait traduire en français par "chacun son tour" [1]. Les threads en mode Ready sont placé dans un ordre arbitraire, sur laquelle on boucle (une fois la fin de cet ordre atteint, on recommence avec le premier thread, et ainsi de suite). À chaque tick d'horloge, le scheduler décide de systématiquement préempter le thread en cours d'exécution, sauf s'il n'existe aucun autre thread en état Ready. Le thread choisi pour passer en état Running est alors le suivant dans la liste. Celui-ci peut alors exécuter une fraction de son burst CPU avant d'être lui même préempté. La figure suivante illustre ce principe avec les mêmes threads que dans les exemples précédents, et avec un tick d'horloge toutes les deux unités de temps.



On observe que le thread T1 n'exécute que deux unités de temps sur les 7 de son burst CPU avant d'être préempté pour laisser la place à T2, qui laisse la place à T3 et ainsi de suite. Le troisième accès du thread T3 au processeur

1. L'expression Round-Robin a une origine intéressante : elle est un idiotisme de l'expression française "Ruban Rond" dont vous pouvez lire l'histoire sur Wikipedia.

permet à ce thread de terminer son burst par une opération bloquante. .. On suppose dans cet exemple que le système d'exploitation remet l'horloge à 0 suite à la fin du thread en cours (ce n'est pas obligatoire).

Le scheduler RR permet à chaque thread d'accéder au processeur équitablement : même si un thread comme T1 ou T4 a un burst CPU long, les threads avec des bursts CPU courts comme T2 ou T3 auront accès au processeur de la même manière. En d'autres termes, le temps d'attente pour un thread sera toujours borné par le nombre de threads en état Ready multiplié par le durée du pas de temps.

On voit toutefois que ce scheduler n'est pas très efficace pour plusieurs raisons :
– Premièrement, il génère un grand nombre de changements de contexte (7 dans notre exemple). Comme discuté précédemment, non seulement ces changements de contexte nécessitent du temps processeur qui n'est pas utilisé pour des opérations utiles, mais ils entrainent surtout un phénomène de cache froid à chaque redémarrage d'un thread sur le processus à la suite d'un autre ayant rempli le cache avec ses propres données.
– Deuxièmement, comme le burst CPU d'un thread peut être interrompu avant sa complétion, il n'y a pas de relation directe entre le temps d'attente et le temps de réponse, et ce dernier peut devenir particulièrement long. Par exemple, bien que T3 ait un temps d'attente de 3 unités de temps, son temps de réponse (le temps entre son placement en état Ready et la fin de son burst CPU) est de 11 unités de temps.
– Enfin, il n'y a pas de distinction entre les threads ayant besoin du processeur pour des bursts courts ou ceux ayant des bursts longs, ce qui peut conjointement réduire la réactivité des threads interactifs ou effectuant de nombreuses entrées/sorties et diminuer la performance de ceux réalisant des calculs.

Note : Quelle fréquence pour l'horloge système ?

La fréquence de l'horloge système, qui génère les interruptions périodiques permettant au système d'exploitation de reprendre la main via la procédure de traitement d'interruption et (entre autres) de permettre au scheduler de préempter un processus en cours d'exécution, est un paramètre important pour la performance et la consommation d'énergie d'un système informatique. La valeur idéale dépend non seulement de l'architecture utilisée (mono- versus multi-processeur, machine alimentée par batterie ou non, etc.), de la configuration du système d'exploitation, mais aussi du type d'applications envisagées (application de type serveur, de type calcul intensif, applications interactives comme des jeux ou du traitement multimédia, etc.).

Par exemple, les versions initiales de Linux utilisaient une fréquence d'horloge de 100 Hz (100 interruptions par seconde) tandis que des versions ultérieures permettaient une fréquence plus élevée de 1.000 Hz. Une fréquence plus élevée permet de diminuer le temps d'attente moyen et augmente la réactivité du système. Elle entraîne une utilisation processeur par le système plus élevée, ce qui est particulièrement problématique pour les systèmes embarqués ou pour les ordinateurs portables alimentés par une batterie. Une fréquence élevée peut aussi augmenter le risque de pollution de caches dues aux préemptions plus important. Les versions modernes de Linux peuvent adapter la fréquence de l'horloge pour ne pas constamment réveiller un processeur lorsqu'il n'y a pas de tâche en état Ready, ou bien ne pas interrompre une tâche en état Running sur un processeur s'il n'y a pas de tâche en état Ready en attente pour le remplacer.

Schedulers à priorité

Dans un même système informatique, plusieurs applications cohabitent et toutes n'ont pas nécessairement la même priorité d'accès aux ressources. Par exemple, lors de l'utilisation d'une interface utilisateur en mode graphique, l'application actuellement utilisée par l'utilisateur local (par exemple un navigateur web) peut avoir besoin pour assurer une bonne réactivité d'accéder plus rapidement au processeur afin de limiter ses temps de réponses. À l'inverse, une opération de maintenance utilisée par le système d'exploitation, comme la mise à jour d'une base de données des fichiers pour permettre la recherche rapide par la suite, peut se contenter d'accéder au processeur uniquement lorsque celui-ci n'est pas sollicité par d'autres applications.

Un scheduler à priorité alloue à chaque thread un niveau de priorité donné. Lorsque le scheduler doit sélectionner un thread à exécuter, il commence d'abord par parcourir les threads ayant une haute priorité. En pratique, un scheduler à priorité maintiendra une liste circulaire pour chaque niveau de priorité. Lorsque le scheduler est appelé, il sélectionnera toujours le thread ayant la plus haute priorité et se trouvant dans l'état *Ready*. Si plusieurs threads ont le même niveau de priorité, un scheduler de type *round-robin* peut être utilisé dans chaque niveau de priorité. Il faut toutefois faire attention au problème de **famine** : si il existe toujours des threads de plus haute priorité qu'un thread donné, ce dernier pourrait ne jamais obtenir l'accès au processeur. Une solution simple à ce problème est de considérer une priorité de base, et une priorité courante. Au démarrage d'un cycle, les threads reçoivent

leur priorité de base. Lorsqu'ils obtiennent l'accès au processeur, leur priorité courante décroit. Ceci donne une opportunité aux threads de priorité de base plus faible de s'exécuter. Un nouveau cycle commence lorsque tous les threads en état Ready ou Running ont atteint une priorité courante de 0.

On peut combiner le principe de priorité avec celui de préemption. Un thread qui passe dans l'état Running obtient alors un crédit de temps, ou quantum. Lors de l'allocation d'un processeur à un thread, le kernel démarre une temporisation avec ce quantum, correspondant à un certain nombre de clicks de l'horloge système (la longueur du quantum doit donc être un multiple de la période de cette horloge). Si un burst CPU atteint la fin de son quantum avant de réaliser une opération bloquante, celui-ci est préempté.

Les systèmes UNIX utilisent souvent des schedulers à priorité dynamique avec un round-robin à chaque niveau de priorité, en ajoutant par ailleurs des mécanismes adaptant la priorité de base des threads pour favoriser les threads interactifs. Par exemple, un thread qui termine toujours ses quantum de temps en étant préempté est considéré comme intensif en processeur (*CPU-intensive*). Il se verra allouer une priorité de base plus grande, mais avec un quantum de temps plus long. En revanche, un thread qui termine toujours ses bursts CPU avant la fin des quantum alloués est considéré comme intensive en entrées/sorties (*interactive*). Ce thread pourra obtenir une priorité de base plus élevé, mais associée à un quantum de temps plus court.

Note : Scheduler à priorité et synchronisation des threads

L'utilisation des primitives de synchronisation comme les mutex peut aller à l'encontre des priorités utilisées par le scheduler. Considérons par exemple le cas de deux threads TA et TB. TA doit répondre à des requêtes reçues depuis le réseau en mettant à jour une structure de données partagée, par exemple un graphe. Cette opération doit terminer le plus rapidement possible et ce thread est donc assigné à une priorité élevée. TB parcours de façon périodique la structure de données commune afin d'en extraire des statistiques (par exemple, toutes les 30 secondes). TB n'a pas de contrainte forte sur son temps de réponse mais l'opération qu'il exécute peut être assez longue. On assigne donc une priorité faible à TB. TA et TB accèdent à la structure de donnée en exclusion mutuelle, en utilisant un mutex *m*. On peut alors rencontrer la situation suivante : TB verrouille le mutex *m* en appelant pthread_mutex_lock(3posix) et commence son opération de parcours de la structure de données. TA passe alors de l'état Blocked à Ready à l'occasion de la réception d'une requête depuis le réseau. Le scheduler peut alors décider de préempter TB pour donner le processeur à TA, de plus grande priorité. Celui-ci va alors appeler pthread_mutex_lock(3posix), et être placé dans la file d'attente pour le mutex *m*. Si une attente active est utilisée, la situation est encore pire : le thread TA va alors boucler pour rien en attendant que son quantum de temps soit écoulé et que TB puisse récupérer un processeur pour terminer sa section critique. Cette situation où un thread de priorité élevé est bloqué en attente d'un thread de priorité faible pour accéder à une ressource exclusive comme un mutex est appelé une **inversion de priorité**. Une solution à ce problème est que lorsqu'un thread obtient un mutex sa priorité soit automatiquement augmentée pendant le temps d'utilisation de ce mutex, limitant ainsi les risques de préemption au milieu de la section critique. Une telle priorité dite *plafond* (priority ceiling) est associée à un mutex en utilisant l'appel pthread_mutexattr_setprioceiling(3posix). Cette priorité doit être la priorité maximale accessible aux threads du processus courant, qui peut être obtenue avec l'appel sched_get_priority_max(3posix).

7.1.4 Influencer la priorité des processus sous Linux

Les processus créés sous un système Linux ont une priorité qui s'applique par défaut à l'ensemble de leurs threads. La priorité originelle d'un processus dépend de la configuration du système et des droits du processus appelant l'appel système fork(2).

Il est possible d'influer sur la priorité d'un processus en utilisant la commande nice(1) ou la fonction nice(2) définie dans *unistd.h*. La commande nice(1) prend deux paramètres : un modificateur de priorité allant de +20 à -19, et la commande à exécuter. Une valeur élevée du modificateur (0 à +20) indique une priorité de plus en plus faible (la priorité avec +20 est la plus faible possible). On peut voir la valeur de nice comme une mesure de *politesse*, qui indique à quel point les threads de ce processus vont accepter de laisser passer les threads des autres processus devant eux pour l'accès au(x) processeur(s). Tout utilisateur peut utiliser une valeur de nice positive, car cela revient à réduire la facilité d'accès au processeur et non à s'octroyer des ressources supplémentaires. Une valeur négative (de -1 à -19) permet d'augmenter la priorité du processus. Leur utilisation nécessite en général des droits spécifiques, dits de super-utilisateur, afin d'éviter que des utilisateurs allouent systématiquement une priorité élevée à leurs programmes dans un environnement partagé, au détriment des autres utilisateurs.

L'exemple suivant montre le démarrage du programme ls(1) tout d'abord avec une valeur de nice de 15 (priorité faible) puis l'essai d'utilisation d'une valeur négative (priorité élevée) dont on voit qu'il est refusé par la commande pour cause de droits insuffisants.

```
utilisateur@systeme:~$ nice -15 ls -la .bash*
-rw------- 1 utilisateur groupe   64 Nov 16 21:33 .bash_history
-rw-r--r-- 1 utilisateur groupe  220 Jun  6  2018 .bash_logout
-rw-r--r-- 1 utilisateur groupe 3536 Oct 24 15:02 .bashrc

utilisateur@systeme:~$ nice --15 ls -la .bash*
nice: cannot set niceness: Permission denied
-rw------- 1 utilisateur groupe   64 Nov 16 21:33 .bash_history
-rw-r--r-- 1 utilisateur groupe  220 Jun  6  2018 .bash_logout
-rw-r--r-- 1 utilisateur groupe 3536 Oct 24 15:02 .bashrc
```

CHAPITRE 8

Mémoire virtuelle

8.1 La mémoire virtuelle

Le modèle d'interaction entre le processeur et la mémoire que nous avons utilisé jusqu'à présent est le modèle traditionnel. Dans ce modèle, illustré sur la figure ci-dessous, la mémoire est divisée en octets. Chaque octet est identifié par une adresse encodée sur n bits. Une telle mémoire peut donc contenir au maximum 2^n octets de données. Aujourd'hui, les processeurs utilisent généralement des adresses sur 32 ou 64 bits. Avec des adresses sur 32 bits, la mémoire peut stocker $4.294.967.296$ octets (4 Go) de données. Avec des adresses sur 64 bits, la capacité de stockage de la mémoire monte à $18.446.744.073.709.551.616$ octets (16 Eo, exa-octets). Si on trouve facilement aujourd'hui des mémoires avec une capacité de $4.294.967.296$ octets, il n'en existe pas encore qui sont capables de stocker $18.446.744.073.709.551.616$ et il faudra probablement quelques années avant que de telles capacités ne soient utilisables en pratique.

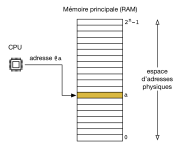


FIGURE 8.1 – Modèle simple d'interaction entre le processeur et la mémoire

Ce modèle correspond au fonctionnement de processeurs simples tels que ceux que l'on trouve sur des systèmes embarqués comme une machine à lessiver. Malheureusement, il ne permet pas d'expliquer et de comprendre le fonctionnement des ordinateurs actuels. Pour s'en convaincre, il suffit de réfléchir à quelques problèmes liés à l'utilisation de la mémoire sur un ordinateur fonctionnant sous Unix.

Le premier problème est lié à l'organisation d'un processus en mémoire. Sous Unix, le bas de la mémoire est réservé au code, le milieu au heap et le haut au stack. Le modèle simple d'organisation de la mémoire ne permet pas facilement de comprendre comment un tel processus peut pouvoir utiliser la mémoire sur un processeur 64 bits qui est placé dans un ordinateur qui ne dispose que de 4 GBytes de mémoire. Avec une telle quantité de mémoire, le sommet de la pile devrait se trouver à une adresse proche de 2^{32} et non 2^{64}.

Un deuxième problème est lié à l'utilisation de plusieurs processus simultanément en mémoire. Lorsque deux processus s'exécutent, ils utilisent nécessairement la même mémoire physique. Si un processus utilise l'adresse *x* et y place des instructions ou des données, cette adresse ne peut pas être utilisée par un autre processus. Physiquement, ces deux processus doivent utiliser des zones mémoires distinctes. Pourtant, le programme ci-dessous affiche les adresses de `argc`, de la fonction `main` et de la fonction `printf` de la librairie standard puis effectue `sleep(20);`. Lors de l'exécution de deux instances de ce programmes simultanément, on observe ceci sur la sortie standard.

```
$ ./simple &
[pid=32955] Adresse de argc : 0x7fff5fbfe18c
```

```
[pid=32955] Adresse de main : 0x100000e28
[pid-32955] Adresse de printf : 0x7fff8a524f3a
$ ./simple 2 3 4
[pid=32956] Adresse de argc : 0x7fff5fbfe18c
[pid=32956] Adresse de main : 0x100000e28
[pid-32956] Adresse de printf : 0x7fff8a524f3a
```

Manifestement, les deux programmes utilisent exactement les mêmes adresses en mémoire. Pourtant, ces deux programmes doivent nécessairement utiliser des zones mémoires différentes pour pouvoir s'exécuter correctement. Ceci est possible grâce à l'utilisation de la *mémoire virtuelle*. Avec la *mémoire virtuelle*, deux types d'adresses sont utilisées sur le système : les *adresses virtuelles* et les *adresses réelles* ou *physiques*. Une *adresse virtuelle* est une adresse qui est utilisée à l'intérieur d'un programme. Les adresses des variables ou des fonctions de notre programme d'exemple ci-dessus sont des adresses virtuelles. Une *adresse physique* est l'adresse qui est utilisée par des puces de RAM pour les opérations d'écriture et de lecture. Ce sont les adresses physiques qui sont échangées sur le bus auquel la mémoire est connectée. Pour que les programmes puissent accéder aux instructions et données qui se trouvent en mémoire, il est nécessaire de pouvoir traduire les adresses virtuelles en adresses physiques. C'est le rôle du *MMU* ou *Memory Management Unit*. Historiquement, le *MMU* était implémenté sous la forme d'un chip séparé qui était placé entre le processeur (qui utilisait alors des adresses virtuelles) et la mémoire (qui utilise elle toujours des adresses physiques). Aujourd'hui, le *MMU* est directement intégré au processeur pour des raisons de performance, mais conceptuellement son rôle reste essentiel comme nous allons le voir.

8.1.1 La mémoire virtuelle

Le rôle principal du *MMU* est de traduire toute adresse virtuelle en une adresse physique. Avant d'expliquer comment le *MMU* peut être mis en œuvre en pratique, il est utile de passer en revue plusieurs avantages de l'utilisation d'adresses virtuelles.

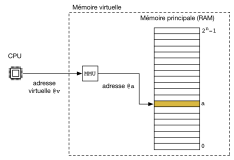


FIGURE 8.2 – *MMU* et *mémoire virtuelle*

Un premier avantage de l'utilisation de la mémoire virtuelle est qu'elle permet de découpler les adresses virtuelles des adresses physiques. Celles-ci ne doivent pas nécessairement être encodées en utilisant le même nombre de bits. La longueur des adresses dépend généralement de l'architecture du processeur et de la taille des registres qu'il utilise. Une organisation possible de la mémoire virtuelle est d'utiliser des adresses virtuelles qui sont encodées sur autant de bits que les adresses physiques, mais ce n'est pas la seule. Il est tout à fait possible d'avoir un ordinateur sur lequel les adresses virtuelles sont plus longues que les adresses physiques. C'est le cas par exemple sur les ordinateurs bon marché qui utilisent une quantité réduite de mémoire RAM. Inversement, la mémoire virtuelle permet à un serveur d'utiliser des adresses physiques qui sont plus longues que les adresses virtuelles. Cela lui permet d'utiliser une capacité de mémoire plus importante que celle autorisée par l'architecture de son processeur. Dans ce cas, un processus ne peut pas utiliser plus de mémoire que l'espace d'adressage virtuel disponible. Mais ensemble, tous les processus fonctionnant sur l'ordinateur peuvent utiliser tout l'espace d'adressage physique disponible.

Un deuxième avantage de la mémoire virtuelle est qu'elle permet, à condition de pouvoir réaliser une traduction spécifique à chaque processus, de partager efficacement la mémoire entre plusieurs processus tout en leur permettant d'utiliser les mêmes adresses virtuelles. C'est cette particularité de la mémoire virtuelle qui nous a permis dans l'exemple précédent d'avoir deux processus qui en apparence utilisent les mêmes adresses. En effectuant une traduction spécifique à chaque processus, le *MMU* permet d'autres avantages qui sont encore plus intéressants.

Le *MMU* est capable d'effectuer des traductions d'adresses virtuelles qui sont spécifiques à chaque processus. Cela implique qu'en général la traduction de l'adresse x dans le processus $P1$ ne donnera pas la même adresse physique que la traduction de l'adresse x dans le processus $P2$. Par contre, il est tout à fait possible que la traduction de l'adresse w (resp. y) dans le processus $P1$ (resp. $P2$) donne l'adresse physique z dans les deux processus.

Comme nous le verrons ultérieurement, cela permet à deux processus distincts de partager de la mémoire. Cette propriété est aussi à la base du fonctionnement des librairies partagées dans un système Unix. Dans notre exemple, la fonction `printf` qui est utilisée par les deux processus fait partie de la librairie standard. Celle-ci doit être chargée en mémoire lors de l'exécution de chacun des processus. Grâce à l'utilisation du *MMU* et de la mémoire virtuelle, une seule copie physique de la librairie standard est chargée en mémoire et tous les processus qui y font appel utilisent les instructions se trouvant dans cette copie physique. Cela permet de réduire fortement la consommation de mémoire lorsque de nombreux processus s'exécutent simultanément, ce qui est souvent le cas sur un système Unix.

Le dernier avantage de l'utilisation de la mémoire virtuelle est qu'il est possible de combiner ensemble la mémoire RAM et un ou des dispositifs de stockage tels que des disques durs ou des disques SSD pour constituer une mémoire virtuelle de plus grande capacité que la mémoire RAM disponible. Pour cela, il suffit, conceptuellement, que le *MMU* soit capable de supporter deux types d'adresses physiques : les adresses physiques en RAM et les adresses physiques qui correspondent à des données stockées sur un dispositif de stockage [1].

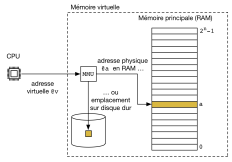


FIGURE 8.3 – Organisation de la *mémoire virtuelle*

Cette possibilité de combiner la mémoire RAM et les dispositifs de stockage offre encore plus de possibilités. Comme nous le verrons, grâce à la mémoire virtuelle, un processus pourra accéder à des fichiers via des pointeurs et des écriture/lectures en mémoire. Le chargement d'un programme pourra s'effectuer en passant par la mémoire virtuelle de façon à charger uniquement les parties du programme qui sont nécessaires en mémoire. Nous verrons également qu'il existe plusieurs appels systèmes qui permettent à des processus de contrôler leur utilisation de la mémoire virtuelle.

8.1.2 Fonctionnement de la mémoire virtuelle

Avant d'analyser comment la mémoire virtuelle peut être utilisée par les processus, il est important de bien comprendre son organisation et les principes de base de fonctionnement du *MMU*. La mémoire virtuelle combine la mémoire RAM et les dispositifs de stockage. Comme la mémoire RAM et les dispositifs de stockage ont des caractéristiques fort différentes, il n'est pas trivial de les combiner pour donner l'illusion d'une mémoire virtuelle unique.

Au niveau de l'adressage, la mémoire RAM permet d'adresser des octets et supporte des lectures et des écritures à n'importe quelle adresse. La mémoire RAM permet au processeur d'écrire et de lire des octets ou des mots à une position déterminée en mémoire.

Un dispositif de stockage (disque dur, SSD, CD/DVD, ...) quant à lui contient un ensemble de secteurs. Chaque secteur peut être identifié par une adresse, comprenant par exemple pour un disque dur le numéro du plateau, le numéro de la piste et le numéro du secteur sur la piste. Sur un tel dispositif, le secteur est l'unité de transfert de l'information. Cela implique que la moindre lecture/écriture sur un dispositif de stockage nécessite la lecture/écriture d'au moins 512 octets, même pour lire ou modifier un seul bit. Enfin, la dernière différence importante entre ces deux technologies est leur temps d'accès. Au niveau des mémoires RAM, les temps d'accès sont de l'ordre de quelques dizaines de nanosecondes. Pour un dispositif de stockage, les temps d'accès peuvent être de quelques dizaines de microsecondes pour un dispositif de type *Solid State Drive* ou *SSD* et jusqu'à quelques dizaines de millisecondes pour un disque dur. .. Les tableaux ci-dessous présentent les caractéristiques techniques de deux dispositifs de stockage [2] [3] à titre d'exemple.

La mémoire virtuelle utilise elle une unité intermédiaire qui est la *page*. Une *page* est une zone de mémoire contigüe. La taille des pages dépend de l'architecture du processeur et/ou du système d'exploitation utilisé. Une taille

1. En pratique, les adresses sur le disque dur ne sont pas stockées dans le *MMU* mais dans une table maintenue par le système d'exploitation. C'est le noyau qui est responsable des transferts entre le dispositif de stockage et la mémoire RAM.
2. Source : http ://www.intel.com/content/www/us/en/solid-state-drives/ssd-320-specification.html
3. Source : http ://www.seagate.com/staticfiles/support/disc/manuals/desktop/Barracuda%207200.11/100507013e.pdf

courante est de 4096 octets. L'appel getpagesize(2) permet d'obtenir la taille de page utilisée dans un système, comme montré dans l'exemple qui suit.

```
#include <unistd.h>
int sz = getpagesize();
```

Lorsqu'un programme est chargé en mémoire, par exemple lors de l'exécution de l'appel système execve(2), il occupe un nombre entier de pages (par exemple, si le segment text du programme contient 5192 octets, le segment text occupera deux pages, dont une partie de la deuxième page sera inutilisée). Grâce à la mémoire virtuelle, ces pages peuvent être stockée dans n'importe quelle zone de la mémoire RAM. La seule contrainte est que tous les octets qui font partie de la même page soient stockés à des adresses qui sont contigües. Cette contrainte permet de structurer les adresses virtuelles en deux parties comme représenté dans la figure ci-dessous. Une *adresse virtuelle* est donc un ensemble de bits. Les bits de poids fort servent à identifier la *page* dans laquelle une donnée est stockée. Les bits de poids faible (12 lorsque l'on utilise des pages de 4 KBytes) identifient la position de la donnée par rapport au début de la page.

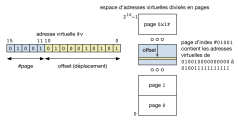


FIGURE 8.4 – Adresse virtuelle

Grâce à cette organisation des adresses virtuelles, il est possible de construire un mécanisme efficace qui permet de traduire une adresse virtuelle en une adresse réelle. La première solution qui a été proposée pour réaliser cette traduction est d'utiliser une *table des pages*. La *table des pages* est stockée en mémoire RAM et contient une ligne pour chaque page appartenant à la mémoire virtuelle. À titre d'exemple, un système utilisant des adresses virtuelles de 32 bits et des pages de 4 KBytes contient $2^{32-12} = 2^{20}$ pages. La table des pages de ce système contient donc 2^{20} lignes. Une ligne de la table des pages contient différentes informations que nous détaillerons par après. Les deux plus importantes sont :

– le *bit de validité* qui indique si la page est actuellement présente en mémoire physique (RAM) ou non,
– l'adresse en mémoire RAM à laquelle la page est actuellement stockée (si elle est présente en mémoire RAM, sinon une information permettant de trouver la page sur un dispositif de stockage)

La table des page est stockée en mémoire RAM comme un tableau en C. L'information correspondant à la page *0* est stockée à l'adresse de début de la table des pages. Cette adresse de début de la table des pages (`P`) est généralement stockée dans un registre du processeur pour être facilement accessible. Si une entrée [4] de la table des pages est encodée en *n* bytes, l'information correspondant à la page *1* sera stockée à l'adresse `P+n`, celle relative à la page *2* à l'adresse `P+2*n`, ... Cette organisation permet d'accéder facilement à l'entrée de la table des pages relative à la page *z*. Il suffit en effet d'y accéder à l'adresse `P+z*n`.

Grâce à cette table des pages, il est possible de traduire directement les adresses virtuelles en adresses physiques. Cette traduction est représentée dans la figure ci-dessous. Pour réaliser une traduction, il faut tout d'abord extraire de l'adresse virtuelle le numéro de la page. Celui-ci se trouve dans les bits de poids fort de l'adresse virtuelle. Le numéro de la page sert d'index pour récupérer l'entrée correspondant à cette page dans la table des pages. Cette entrée contient l'adresse en mémoire RAM à laquelle la page débute. Pour finaliser la traduction de l'adresse virtuelle, il suffit de concaténer les bits de poids faible de l'adresse virtuelle avec l'adresse de la page en mémoire RAM. Cette concaténation donne l'adresse réelle à laquelle la donnée est stockée en mémoire RAM. Cette adresse physique permet au processeur d'accéder directement à la donnée en mémoire.

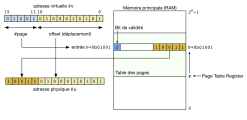


FIGURE 8.5 – Traduction d'adresses avec une table des pages

La table des pages permet de traduire les adresses virtuelles en adresses physiques. Ce faisant, elle introduit un mécanisme d'indirection entre les adresses (virtuelles) qui sont utilisées par les programmes et les adresses

4. Une entrée de la table de pages occupe généralement 32 ou 64 bits suivant les architectures de processeurs.

(réelles) qui sont utilisées par le hardware. Ce mécanisme d'indirection a de nombreuses applications comme nous le verrons par la suite.

Un point important à mentionner concernant l'utilisation d'un mécanisme de traduction des adresses est qu'il permet de découpler le choix de la taille des adresses (virtuelles) utilisées par les programmes des contraintes matérielles qui sont liées directement aux mémoires RAM utilisées. En pratique, il est tout à fait possible d'avoir des systèmes informatiques dans lesquels les adresses virtuelles sont plus longues, plus courtes ou ont la même longueur que les adresses physiques. Sur un ordinateur utilisant des adresses en 32 bits et équipé de 4 GBytes de mémoire, il est naturel d'utiliser des adresses virtuelles de 32 bits et des adresses physiques de 32 bits également pour pouvoir accéder à l'ensemble de la mémoire. Dans ce cas, la mémoire virtuelle permet d'accéder à toute la mémoire physique. Aujourd'hui, la majorité des ordinateurs de bureau et des serveurs utilisent des adresses sur 64 bits. Ceux-ci utilisent des adresses virtuelles de 64 bits, mais aucun ordinateur ne contient *2^64* bytes de mémoire. Par exemple, un serveur disposant de 128 GBytes de mémoire physique pourrait se contenter d'utiliser des adresses physiques de 37 bits. Dans ce cas, la mémoire virtuelle donne l'illusion qu'il est possible d'accéder à plus de mémoire que celle qui est réellement disponible. D'un autre côté, il est aussi possible de construire des serveurs qui utilisent des adresses virtuelles de 32 bits, mais disposent de plus de 4 GBytes de mémoire RAM. Dans ce cas, les adresses physiques pourront être plus longues que les adresses réelles. Quelles que soient les longueurs respectives des adresses virtuelles et physiques, la table des pages, sous le contrôle du système d'exploitation, permettra de réaliser efficacement les traductions entre les adresses virtuelles et les adresses physiques.

Pour bien comprendre la traduction des adresses virtuelles en utilisant la table des pages, considérons un système imaginaire qui utilise des adresses virtuelles encodées sur 7 bits et des adresses physiques qui sont elles encodées sur 6 bits. La table des pages correspondante est reprise dans le tableau ci-dessous. La ligne du bas du tableau est relative à la page *0*.

Index	Validité	Adresse
7	true	00
6	false	–
5	true	11
4	false	–
3	false	–
2	false	–
1	true	01
0	true	10

Cette mémoire virtuelle contient quatre pages. La première couvre les adresses physiques allant de `000000` à `001111`, la seconde de `010000` à `011111`, la troisième de `100000` à `101111` et la dernière de `110000` à `111111`. Les adresses virtuelles elles vont de `0000000` à `1111111`. La traduction s'effectue sur base de la table des pages. Ainsi, l'adresse `1010001` correspond à l'octet `0001` dans la page virtuelle `101`. Sur base de la table des pages, cette page se trouve en mémoire RAM (son bit de validité est vrai) et elle démarre à l'adresse `110000`. L'adresse virtuelle `1010001` est donc traduite en l'adresse réelle `110001`. L'adresse virtuelle `0110111` correspond elle à une page qui n'est pas actuellement en mémoire RAM puisque le bit de validité correspondant à la page `011` est faux.

Si on analyse la table des pages ci-dessus, on peut remarquer que la page contenant les adresses virtuelles les plus hautes se trouve dans la zone mémoire avec les adresses physiques les plus basses. Inversement, la page qui est en mémoire RAM à l'adresse la plus élevée correspond à des adresses virtuelles qui se trouvent au milieu de l'espace d'adressage. Ce découplage entre l'adresse virtuelle et la localisation physique de la page en mémoire est un des avantages importants de la mémoire virtuelle.

La mémoire virtuelle a aussi un rôle important à jouer lorsque plusieurs processus s'exécutent simultanément. Comme indiqué ci-dessus, l'adresse de la table des pages est stockée dans un des registre du processeur. L'utilisation de ce registre permet d'avoir une table des pages pour chaque processus. Pour cela, il suffit qu'une zone de mémoire RAM soit réservée pour chaque processus et que le système d'exploitation y stocke la table des pages du processus. Lors d'un changement de contexte, le système d'exploitation modifie le registre de table des pages de façon à ce qu'il pointe vers la table des pages du processus qui s'exécute. Ce mécanisme est particulièrement

utile et efficace.

A titre d'exemple, considérons un système imaginaire utilisant des adresses virtuelles sur 6 bits et des adresses physiques sur 8 bits. Deux processus s'exécutent sur ce système et ils utilisent chacun trois pages, deux pages dans le bas de l'espace d'adressage virtuel qui correspondent à leur segment de code et une page dans le haut de l'espace d'adressage virtuel qui correspond à leur pile. Le premier tableau ci-dessous présente la table des pages du processus `P1`.

Index	Validité	Adresse
3	true	0011
2	false	–
1	true	1001
0	true	1000

Le processus `P2` a lui aussi sa table des pages. Celle-ci pointe vers des adresses physiques qui sont différentes de celle utilisées par le processus `P1`. L'utilisation d'une table des pages par processus permet à deux processus distincts d'utiliser les mêmes adresses virtuelles.

Index	Validité	Adresse
3	true	0000
2	false	–
1	true	1111
0	true	1110

Lorsque le processus `P1` s'exécute, c'est sa table des pages qui est utilisée par le processeur pour la traduction des adresses virtuelles en adresses physiques. Ainsi, l'adresse `011101` est traduite en l'adresse `10011101`. Par contre, lorsque le processus `P2` s'exécute, cette adresse `011101` est traduite grâce à la table des pages de ce processus en l'adresse `11111101`.

Note : Performance de la mémoire virtuelle

Grâce à son mécanisme d'indirection entre les adresses virtuelles et les adresses physiques, la mémoire virtuelle permet de nombreuses applications comme nous le verrons dans les sections qui suivent. Cependant, la mémoire virtuelle peut avoir un impact important au niveau des performances des accès à une donnée en mémoire. Pour cela, il est intéressant d'analyser en détails ce qu'il se passe lors de chaque accès à la mémoire. Pour accéder à une donnée en mémoire, le *MMU* doit d'abord consulter la table des pages pour traduire l'adresse virtuelle en une adresse physique correspondante. Ce n'est qu'après avoir obtenu cette adresse physique que le processeur peut effectuer l'accès à la mémoire RAM. En pratique, l'utilisation d'une table des pages a comme conséquence de doubler le temps d'accès à une donnée en mémoire. Lorsque la mémoire virtuelle a été inventée, ce doublement du temps d'accès à la mémoire n'était pas une limitation car les mémoires RAM étaient nettement plus rapides que les processeurs. Aujourd'hui, la situation est complètement inversée puisque les processeurs sont déjà fortement ralentis par les temps d'accès à la mémoire RAM. Doubler ce temps d'accès aurait un impact négatif sur les performances des processeurs. Pour faire face à ce problème, les processeurs actuels disposent tous d'un *Translation Lookaside Buffer* (*TLB*). Ce *TLB* est en fait une sorte de *mémoire cache* qui permet de stocker dans une mémoire rapide se trouvant sur le processeur certaines lignes de la *table des pages*. Les détails de gestion du *TLB* sortent du cadre de ce cours [HennessyPatterson]. Avec l'utilisation du *TLB*, et grâce au principe de localité qui fait que les accès mémoire se font très souvent dans un ensemble limité de pages, la plupart des traductions des adresses virtuelles en adresses physique peuvent être obtenues sans devoir directement consulter la table des pages.

La table des pages d'un processus contrôle les adresses physiques auxquelles le processus a accès. Pour garantir la sécurité d'un système informatique, il faut bien entendu éviter qu'un processus ne puisse modifier lui-même et sans contrôle sa table des pages. Toutes les manipulations de la table des pages ou du registre de table des pages se font sous le contrôle du système d'exploitation. La modification du registre de table des pages est une opération privilégiée qui ne peut être exécutée que par le noyau du système d'exploitation.

Pour permettre la gestion des accès et des fonctionnalités de sécurité, une entrée de la table des pages contient également des bits de permission qui sont contrôlés par le système d'exploitation et spécifient quelles opérations peuvent être effectuées sur chaque page. Une entrée de la table des pages contient trois bits de permissions :

– *R* bit. Ce bit indique si le processus peut accéder en lecture à la page se trouvant en mémoire physique.
– *W* bit. Ce bit indique si le processus peut modifier le contenu de la page se trouvant en mémoire physique
– *X* bit. Ce bit indique si la page contient des instructions qui peuvent être exécutées par le processeur (si vrai), ou des données (si faux).

Ces bits de protection sont généralement fixés par le système d'exploitation. Par exemple, le segment code qui ne contient que des instructions à exécuter pourra être stocké dans des pages avec les bits *R* et *X* mais pas le bit *W* pour éviter que le processus ne modifie les instructions qu'il exécute. Le stack par contre sera placé dans des pages avec les bits *R* et *W* mais pas le bit *X*[5] . Le heap peut utiliser les mêmes bits de protection. Enfin, les pages qui n'ont pas été allouées au processus, notamment celles se trouvant entre le heap et le stack auront toutes leurs bits de protection mis à faux. L'appel système brk(2) ou sbrk(2) aura pour effet de réserver certaines pages afin de permettre d'y allouer de la mémoire dynamiquement. L'utilisation des bits de protection permet au processeur de détecter les accès à de la mémoire qui n'a pas été allouée au processus. Un tel accès provoquera la génération d'une *segmentation fault* et l'envoi du signal correspondant.

Même si ces bits de protection sont contrôlés par le système d'exploitation, il est parfois utile à un processus de modifier les bits de permissions qui sont associés à certaines de ses pages. Cela peut se faire via l'appel système mprotect(2).

```
#include <sys/mman.h>

int mprotect(const void *addr, size_t len, int prot);
```

Cet appel système prend trois arguments. Le premier est un pointeur vers le début de la zone mémoire dont il faut modifier les bits de protection. Le second est la longueur de la zone mémoire concernée et le dernier la protection souhaitée. Celle-ci est spécifiée en utilisant les constantes `PROT_NONE`, `PROT_READ`, `PROT_WRITE` et `PROT_EXEC` qui peuvent être combinées en utilisant une disjonction logique. La protection demandée ne peut pas être plus libérale que la protection qui est déjà fixée par le système d'exploitation. Dans ce cas, le système d'exploitation génère un signal `SIGSEGV`.

8.1.3 Utilisation des dispositifs de stockage

La mémoire virtuelle permet non seulement à des pages d'un processus d'être placées à différents endroits de la mémoire, mais aussi elle permet de combiner la mémoire RAM et les dispositifs de stockage de façon transparente pour les processus.

Une partie des pages qui composent la mémoire virtuelle peut être stockée sur un dispositif de stockage (disque dur, SSD, ...). En pratique, la mémoire RAM peut jouer le rôle d'une sorte de mémoire cache pour la mémoire virtuelle. Les pages qui sont le plus fréquemment utilisées sont placées en mémoire RAM par le système d'exploitation. Les pages les moins utilisées sont quant à elles placées sur un dispositif de stockage et ramenées en mémoire RAM lorsqu'elles sont utilisées par le processeur.

Pour bien comprendre cette utilisation de la mémoire virtuelle, il nous faut revenir à la table des pages. Celle-ci comprend autant d'entrées qu'il y a de pages dans l'espace d'adressage d'un processus. Nous avons vu qu'une entrée de cette table pouvait être structurée comme dans la figure ci-dessous.

FIGURE 8.6 – Entrée de la table des pages

Le bit de validité indique si la page est présente en mémoire RAM ou non. Lorsque la page est présente en mémoire RAM, les bits de poids faible de l'entrée de la table des pages contiennent l'adresse physique de la page en mémoire RAM. Lorsque le bit de validité a comme valeur *faux*, cela signifie que la page n'existe pas (elle n'a jamais été créée) ou qu'elle est actuellement stockée sur un dispositif de stockage. Si la page n'existe pas, aucun de ses bits de permission n'aura comme valeur *vrai* et tout accès à cette page provoquera une *segmentation fault*. Si par contre la page existe mais se trouve sur un dispositif de stockage, alors l'information de localisation pointera

5. Cette technique est utilisée dans les systèmes d'exploitation récents pour éviter qu'un problème de buffer overflow sur le stack ne conduise à l'exécution d'instructions qui ne font pas partie du processus.

vers une structure de données qui est maintenue par le système d'exploitation et contient la localisation physique de la donnée sur un dispositif de stockage.

Schématiquement, ces informations de localisation des pages peuvent être de deux types. Lorsqu'un dispositif de stockage, ou une partition d'un tel dispositif, est dédié au stockage de pages de la mémoire virtuelle, alors la localisation d'une page est composée de l'identifiant du dispositif et du numéro du secteur sur le dispositif. Ce sera le cas lorsque par exemple une *partition de swap* est utilisée. Sous Linux, le fichier `/proc/swaps` contient la liste des partitions de swap qui sont utilisées pour stocker les pages de la mémoire virtuelle avec leur type, leur taille et leur utilisation. Une telle partition de swap peut être créée avec l'utilitaire mkswap(8). Elle est activée en exécutant la commande swapon(8). Celle-ci est généralement lancée automatiquement lors du démarrage du système.

```
$ cat /proc/swaps
Filename                Type            Size    Used    Priority
/dev/sda3               partition       8193140 444948  -1
```

Outre les partitions de swap, il est également possible de stocker des pages de la mémoire virtuelle dans des fichiers. Dans ce cas, la localisation d'une page comprend le dispositif, l'identifiant du fichier (son *inode* comme nous le verrons lorsque nous couvrirons les systèmes de fichiers) et l'offset à partir duquel la page est accessible dans ce fichier. En pratique, les partitions de swap sont un peu plus rapides que les fichiers de swap car les secteurs qui composent une telle partition sont contigus, ce qui n'est pas toujours le cas avec un fichier de swap. D'un autre côté, il est plus facile d'ajouter ou de retirer des fichiers de swap que des partitions de swap sur un dispositif de stockage. En pratique, les deux techniques peuvent être utilisées.

A ce stade, il est utile d'analyser à nouveau le fonctionnement de la mémoire virtuelle. En toute généralité, celle-ci est découpée en pages et comprend une mémoire RAM et un ou plusieurs dispositifs de stockage. Pour simplifier la présentation, nous supposons qu'un seul disque dur est utilisé.

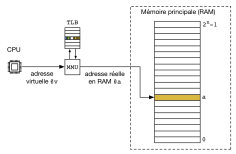


FIGURE 8.7 – La mémoire virtuelle

Les processus utilisent des adresses virtuelles pour représenter les positions des données et des instructions en mémoire virtuelles. Ces adresses virtuelles sont donc utilisées par le CPU chaque fois qu'il doit charger ou sauvegarder une donnée ou une instruction en mémoire. Comme nous l'avons vu, le *MMU* permet de traduire les adresses virtuelles en adresses réelles. Pour des raisons de performance, le *MMU* est intégré directement sur le processeur et il comprend un *TLB* qui sert de cache pour les entrées de la table des pages du processus qui est en train de s'exécuter.

Considérons une opération de lecture faite par le CPU. Pour réaliser cette opération, le CPU fournit l'adresse virtuelle au *MMU*. Celui-ci va consulter le *TLB* pour traduire l'adresse virtuelle demandée. Cette traduction peut nécessiter différentes opérations. Supposons que l'entrée de la table des pages demandées se trouve dans le *TLB*.

- Si le *bit de validité* de la page est *vrai*, la page demandée se situe en mémoire RAM. Dans ce cas, le *MMU* vérifie via les bits de permissions si l'accès demandé (dans ce cas une lecture, mais un raisonnement similaire est valable pour une écriture ou le chargement d'une instruction) est valide.
- Si l'accès est autorisé, le *MMU* retourne l'adresse réelle et le processeur accède aux données.
- Si l'accès n'est pas autorisé, le processeur génère une interruption. Le processus ayant tenté d'accéder à une zone de mémoire ne faisant pas partie de son espace d'adressage virtuel, c'est au système d'exploitation de réagir. Celui-ci enverra un signal segmentation fault, `SIGSEGV`, au processus qui a tenté cet accès.
- Si le *bit de validité* de la page est *faux*, la page demandée ne se trouve pas en mémoire RAM. Deux cas de figure sont alors possibles :
  - les bits de permission ne permettent aucun accès à la page. Dans ce cas, la page n'existe pas et le *MMU* va générer une interruption qui va provoquer l'exécution d'une routine de traitement d'interruption du système d'exploitation. Lors du traitement de cette opération, le noyau va envoyer un signal segmentation fault au processus qui a tenté cet accès.

– les bits de permission permettent l'accès à la page. On parle dans ce cas de *page fault*, c'est-à-dire qu'une page nécessaire à l'exécution du processus n'est pas disponible en mémoire RAM. Vu les temps d'accès et la complexité d'accéder à une page sur un disque dur (via une partition, un fichier de swap ou un fichier normal), le *MMU* ne peut pas accéder directement à la donnée sur le disque dur. Le *MMU* va donc générer une interruption qui va forcer l'exécution d'une routine de traitement d'interruption par le noyau. Cette routine va identifier la page manquante et préparer son transfert du disque dur vers la mémoire. Ce transfert peut durer plusieurs dizaines de millisecondes, ce qui est un temps très long par rapport à l'exécution d'instructions par le processeur. Tant que cette page n'est pas disponible en mémoire RAM, le processus ne peut pas continuer son exécution. Il passe dans l'état Blocked et le thread est alors associé à la structure d'attente correspondant à cette opération. Le scheduler va alors pouvoir sélectionner un autre thread en état Ready et lui allouer le processeur via un changement de contexte. Lorsque la page manquante aura été rapatriée depuis le disque dur en mémoire RAM, le thread pourra repasser en mode Ready et le scheduler redonnera accès à ce thread à un processeur, afin de retenter l'accès mémoire qui vient d'échouer.

Durant son exécution, un système doit pouvoir gérer des pages qui se trouvent en mémoire RAM et des pages qui sont stockées sur le disque dur. Lorsque la mémoire RAM est entièrement remplies de pages, il peut être nécessaire d'y libérer de l'espace mémoire en déplaçant des pages vers un des dispositifs de stockage. C'est le rôle des algorithmes de remplacement de pages.

8.1.4 Stratégies de remplacement de pages

C'est le système d'exploitation qui prend en charge les transferts de pages entre les dispositifs de stockage et la mémoire. Tant que la mémoire RAM n'est pas remplie, ces transferts sont simples, il suffit de ramener une ou plusieurs pages du dispositif de stockage vers la mémoire RAM. En général, le système d'exploitation cherchera à exploiter le principe de localité lors de ces transferts. Lorsqu'une page manque en mémoire RAM, le noyau programmera le chargement de cette page, mais aussi d'autres pages du même processus ayant des adresses proches.

Lorsque la mémoire RAM est remplie et qu'il faut ramener une page depuis un dispositif de stockage, le problème est plus délicat. Pour pouvoir charger cette nouvelle page en mémoire RAM, le système d'exploitation doit libérer de la mémoire. Pour cela, il doit implémenter une *stratégie de remplacement de pages* en mémoire. Cette stratégie définit quelle page doit être préférentiellement retirée de la mémoire RAM et placée sur le dispositif de stockage pour faire de la place pour la nouvelle page. Différentes stratégies sont possibles. Elles résultent en général d'un compromis entre la quantité d'information de contrôle qui doit être stockée dans la table des pages et les performances de la stratégie de remplacement des pages.

Une première stratégie de remplacement de pages pourrait être de sauvegarder les identifiants des pages dans une *file FIFO*. Chaque fois qu'une page est créée par le noyau, son identifiant est placé à la fin de cette *file FIFO*. Lorsque la mémoire est pleine et qu'une page doit être retirée de la mémoire RAM, le noyau pourrait choisir la page dont l'identifiant se trouve en tête de la *file FIFO*. Cette stratégie a l'avantage d'être simple à implémenter, mais remettre sur disque la page la plus anciennement créée n'est pas toujours la solution la plus efficace du point de vue des performances. En effet, cette page peut très bien être une des pages les plus utilisées par le processeur. Si elle est remise sur le disque, elle risque de devoir être récupérée peu de temps après.

Au niveau des performances, la meilleure stratégie de remplacement de pages serait de sauvegarder sur le disque dur les pages qui seront utilisées par le processeur d'ici le plus de temps possible. Malheureusement, cette stratégie nécessite de prévoir le futur, ce qui n'est évidemment pas possible (comme nous l'avons vu au chapitre précédent avec l'algorithme de scheduling hypothétique shortest-job-first) ... Une solution alternative serait de comptabiliser les accès aux différentes pages et de sauvegarder sur disque les pages qui ont été les moins utilisées. Cette solution est séduisante du point de vue théorique car en disposant de statistiques sur l'utilisation des pages, le système d'exploitation devrait pouvoir être capable de mieux prédire les pages qui seront nécessaires dans le futur et les conserver en mémoire RAM. Du point de vue de l'implémentation par contre, cette solution est loin d'être réaliste. En effet, pour maintenir un compteur du nombre d'accès à une page, il faut consommer de la mémoire supplémentaire dans chaque entrée de la table des pages. Mais il faut aussi que le *TLB* puisse incrémenter ce compteur lors de chaque accès à une de ces entrées. Cela augmente inutilement la complexité du *TLB*.

Stocker dans le *TLB* l'instant du dernier accès à une page de façon à pouvoir déterminer quelles sont les pages auxquelles le système a accédé depuis le plus longtemps est une autre solution séduisante d'un point de vue théorique. Du point de vue de l'implémentation, c'est loin d'être facilement réalisable. Tout d'abord, pour que

cet instant soit utile, il faut probablement disposer d'une résolution d'une milliseconde voire mieux. Une telle résolution consommera au moins quelques dizaines de bits dans chaque entrée de la table des pages. En outre, le *TLB* devra pouvoir mettre à jour cette information lors de chaque accès.

Face à ces difficultés d'implémentation, la plupart des stratégies de remplacement de pages s'appuient sur deux bits qui se trouvent dans chaque entrée de la table des pages [HennessyPatterson]. Il est relativement facile de supporter ces deux bits dans une mise en œuvre matérielle du *TLB* et leur présence n'augmente pas de façon significative la mémoire occupée par une entrée de la table des pages.

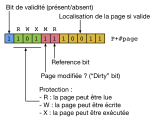


FIGURE 8.8 – Une entrée complète de la table des pages

Outre les bits de validité et de permission, une entrée de la table des pages contient les bits de contrôle suivants :
– le *bit de référence* est mis à vrai par le *MMU* dans le *TLB* à chaque accès à une donnée se trouvant dans la page correspondante, que cet accès soit en lecture ou en écriture
– le *bit de modification* ou *dirty bit* est mis à vrai par le *MMU* chaque fois qu'une opération d'*écriture* est réalisée dans cette page.

Ces deux bits sont mis à jour par le *MMU* à l'intérieur du *TLB*. Lorsqu'une entrée de la table des pages est retirée du *TLB* pour être remise en mémoire, que ce soit à l'occasion d'un changement de contexte ou parce que le *TLB* est plein, les valeurs de ces deux bits sont recopiées dans l'entrée correspondante de la table des pages. En somme, le *TLB* fonctionne comme une cache en *write-back* pour ces deux bits de contrôle.

Les bits de référence et de modification sont utilisés par la plupart des algorithmes de remplacement de pages. Le bit de référence est généralement utilisé par le système d'exploitation pour déterminer quelles sont les pages auxquelles un processus accède actuellement. Pour cela, le noyau va régulièrement remettre à *faux* les bits de validité des entrées des tables de pages. Lorsque une entrée de la table des pages est chargée dans le *TLB* suite à un *page fault*, son *bit de référence* est mis à *vrai*. Il en va de même chaque fois que le processeur accède à une donnée dans cette page.

La stratégie de remplacement utilisant une *file FIFO* que nous avons mentionné précédemment peut être améliorée en utilisant le *bit de référence*. Plutôt que de remettre sur disque la page dont l'identifiant est en tête de la file il suffit de regarder son *bit de référence*. Si celui-ci a la valeur *faux*, la page n'a pas été utilisée récemment et peut donc être retirée de la mémoire RAM. Sinon, le *bit de référence* est remis à *faux* et l'identifiant de la page est replacé en fin de file. L'algorithme de remplacement de page passe ensuite à la page suivante dans la file et continue jusqu'à trouver suffisamment de pages ayant leur bit de référence mis à *faux*.

Une autre stratégie est de combiner le *bit de référence* et le *dirty bit*. Dans ce cas, le noyau du système d'exploitation va régulièrement remettre à la valeur *faux* tous les bits de référence (par exemple toutes les secondes). Lorsque la mémoire RAM est pleine et qu'il faut libérer de l'espace mémoire pour charger de nouvelles pages, l'algorithme de remplacement de pages va grouper les pages en mémoire en quatre classes.

1. La première classe comprend les pages dont le *bit de référence* et le *bit de modification* ont comme valeur *faux*. Ces pages n'ont pas été utilisées récemment et sont identiques à la version qui est déjà stockée sur disque. Elles peuvent donc être retirée de la mémoire RAM sans nécessiter de transfert vers le disque.
2. La deuxième classe comprend les pages dont le *bit de référence* a comme valeur *faux* mais le *bit de modification* a comme valeur *vrai*. Ces pages n'ont pas été utilisées récemment, mais doivent être transférées vers le disque avant d'être retirées de la mémoire RAM.
3. La troisième classe comprend les pages dont le *bit de référence* a comme valeur *vrai* mais le *bit de modification* a comme valeur *faux*. Ces pages ont été utilisées récemment mais peuvent être retirées de la mémoire RAM sans nécessiter de transfert vers le disque.
4. La dernière classe comprend les pages dont les bits de référence et de modification ont comme valeur *vrai*. Ces pages ont étés utilisées récemment et il faut les transférer vers le disque avant de les retirer de la mémoire RAM.

Si l'algorithme de remplacement de pages doit retirer des pages de la mémoire RAM, il commencera par retirer des pages de la première classe, et ensuite de la deuxième, ...

Des algorithmes plus performants ont été proposés et sont utilisés en pratique. Une description détaillée de ces algorithmes sort du cadre de ce cours d'introduction mais peut être trouvée dans un livre consacré aux systèmes d'exploitation comme [Tanenbaum+2009].

Note : Swapping et pagination

Grâce à l'utilisation de la mémoire virtuelle qui est découpée en pages, il est possible de stocker certaines parties de processus sur un dispositif de stockage plutôt qu'en mémoire RAM. Cette technique de *pagination* permet au système d'exploitation de gérer efficacement la mémoire et de réserver la mémoire RAM aux parties de processus qui sont nécessaires à leur exécution. Grâce à la découpe en pages, il est possible de transférer de petites parties d'un processus temporairement sur un dispositif de stockage. Aujourd'hui, la *pagination* est très largement utilisée, mais ce n'est pas la seule technique qui permette de placer temporairement l'espace mémoire utilisé par un processus sur disque. Le *swapping* est une technique plus ancienne mais qui est encore utilisée en pratique. Le *swapping* est plus radical que la *pagination* puisque cette technique permet au noyau de sauvegarder sur disque la quasi totalité de la mémoire utilisée par un processus. Le noyau fera appel au swapping lorsque la mémoire RAM est surchargée et pour des processus qui sont depuis longtemps bloqués par exemple en attente d'un opération d'entrée/sortie. Lorsque le noyau manque de mémoire, il est plus efficace de sauvegarder un processus complet plutôt que de transférer des pages de différents processus. Un tel processus swappé sera réactivé et ramené en mémoire par le noyau lorsqu'il repassera dans l'état *Running*, par exemple suite à la réussite d'une opération d'entrée/sortie.

8.1.5 Utilisations de la mémoire virtuelle

Comme nous l'avons vu précédemment, la mémoire virtuelle est découpée en pages et elle permet de découpler les adresses utilisées par les processus des adresses physiques. Grâce à la table des pages, il est possible de placer les pages d'un processus à n'importe quel endroit de la mémoire RAM. Mais la mémoire virtuelle permet également d'interagir avec les dispositifs de stockage comme si ils faisaient partie de la mémoire accessible au processus.

8.1.6 Fichiers mappés en mémoire

Lorsqu'un processus Unix veut lire ou écrire des données dans un fichier, il utilise en général les appels systèmes open(2), read(2), write(2) et close(2) directement ou à travers une librairie de plus haut niveau comme la libraire d'entrées/sorties standard. Ce n'est pas la seule façon d'accéder à des données sur un dispositif de stockage. Grâce à la mémoire virtuelle, il est possible de placer le contenu d'un fichier ou d'une partie de fichier dans une zone de la mémoire du processus. Cette opération peut être effectuée en utilisant l'appel système mmap(2). Cet appel système permet de rendre des pages d'un fichier accessibles à travers la table des pages du processus comme illustré dans la figure ci-dessous.

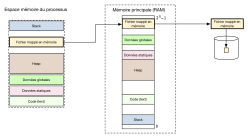


FIGURE 8.9 – Fichiers mappés en mémoire

```
#include <sys/mman.h>

void *mmap(void *addr, size_t length, int prot, int flags,
           int fd, off_t offset);
```

L'appel système mmap(2) prend six arguments, c'est un des appels systèmes qui utilise le plus d'arguments. Il permet de rendre accessible une portion d'un fichier via la mémoire d'un processus. Le cinquième argument est le descripteur du fichier qui doit être mappé. Celui-ci doit avoir été préalablement ouvert avec l'appel système open(2). Le sixième argument spécifie l'offset à partir duquel le fichier doit être mappé, `0` correspondant au début du fichier. Le premier argument est l'adresse à laquelle la première page du fichier doit être mappée. Généralement, cet argument est mis à `NULL` de façon à laisser le noyau choisir l'adresse la plus appropriée. Si cette adresse

est spécifiée, elle doit être un multiple de la taille des pages. Le deuxième argument est la longueur de la zone du fichier qui doit être mappée en mémoire. Le troisième argument contient des drapeaux qui spécifient les permissions d'accès aux données mappées. Cet argument peut soit être `PROT_NONE`, ce qui indique que la page est inaccessible soit une permission classique :
– `PROT_EXEC`, les pages mappées contiennent des instructions qui peuvent être exécutées
– `PROT_READ`, les pages mappées contiennent des données qui peuvent être lues
– `PROT_WRITE`, les pages mappées contiennent des données qui peuvent être modifiées
Ces drapeaux peuvent être combinés avec une disjonction logique. Le quatrième argument est un drapeau qui indique comment les pages doivent être mappées en mémoire. Ce drapeau spécifie comment un fichier qui est mappé par deux ou plusieurs processus doit être traité. Deux drapeaux sont possibles :
– `MAP_PRIVATE`. Dans ce cas, les pages du fichier sont mappées dans chaque processus, mais si un processus modifie une page, cette modification n'est pas répercutée aux autres processus qui ont mappé la même page de ce fichier.
– `MAP_SHARED`. Dans ce cas, plusieurs processus peuvent accéder et modifier la page qui est mappée en mémoire. Lorsqu'un processus modifie le contenu d'une page, la modification est visible aux autres processus. Par contre, le fichier qui est mappé en mémoire n'est modifié que lorsque le noyau du système d'exploitation décide d'écrire les données modifiées sur le dispositif de stockage. Ces écritures dépendent de nombreux facteurs, dont la charge du système. Si un processus veut être sûr des écritures sur disque des modifications qu'il a fait à un fichier mappé un mémoire, il doit exécuter l'appel système msync(2) ou supprimer le mapping via munmap(2).
Ces deux drapeaux peuvent dans certains cas particuliers être combinés avec d'autres drapeaux définis dans la page de manuel de mmap(2).

Lorsque mmap(2) réussit, il retourne l'adresse du début de la zone mappée en mémoire. En cas d'erreur, la constante `MAP_FAILED` est retournée et `errno` est mis à jour en conséquence.

L'appel système msync(2) permet de forcer l'écriture sur disque d'une zone mappée en mémoire. Le premier argument est l'adresse du début de la zone qui doit être écrite sur disque. Le deuxième argument est la longueur de la zone qui doit être écrite sur le disque. Enfin, le dernier contient un drapeau qui spécifie comment les pages correspondantes doivent être écrites sur le disque. La drapeau `MS_SYNC` indique que l'appel msync(2) doit bloquer tant que les données n'ont pas été écrites. Le drapeau `MS_ASYNC` indique au noyau que l'écriture doit être démarrée, mais l'appel système peut se terminer avant que toutes les pages modifiées aient été écrites sur disque.

```
#include <sys/mman.h>
int msync(void *addr, size_t length, int flags);
```

Lorsqu'un processus a fini d'utiliser un fichier mappé en mémoire, il doit d'abord supprimer le mapping en utilisant l'appel système munmap(2). Cet appel système prend deux arguments. Le premier doit être un multiple de la taille d'une page [6]. Le second est la taille de la zone pour laquelle le mapping doit être retiré.

```
#include <sys/mman.h>

int munmap(void *addr, size_t length);
```

A titre d'exemple d'utilisation de mmap(2) et munmap(2), le programme ci-dessous implémente l'équivalent de la commande cp(1). Il prend comme arguments deux noms de fichiers et copie le contenu du premier dans le second. La copie se fait en mappant le premier fichier entièrement en mémoire et en utilisant la fonction memcpy(3) pour réaliser la copie. Cette solution fonctionne avec de petits fichiers. Avec de gros fichiers, elle n'est pas très efficace car tout le fichier doit être mappé en mémoire.

```
#include <sys/types.h>
#include <sys/stat.h>
#include <sys/mman.h>
#include <fcntl.h>
#include <stdlib.h>
#include <stdio.h>
#include <unistd.h>
#include <string.h>

int main (int argc, char *argv[]) {
 int file1, file2;
```

6. Il est possible d'obtenir la taille des pages utilisée sur un système via les appels sysconf(3) ou getpagesize(2)

```
void *src, *dst;
struct stat file_stat;
char dummy=0;

if (argc != 3) {
  fprintf(stderr,"Usage : cp2 source dest\n");
  exit(EXIT_FAILURE);
}
// ouverture fichier source
if ((file1 = open (argv[1], O_RDONLY)) < 0) {
  perror("open(source)");
  exit(EXIT_FAILURE);
}

if (fstat (file1,&file_stat) < 0) {
  perror("fstat");
  exit(EXIT_FAILURE);
}
// ouverture fichier destination
if ((file2 = open (argv[2], O_RDWR | O_CREAT | O_TRUNC, file_stat.st_mode)) < 0) {
  perror("open(dest)");
  exit(EXIT_FAILURE);
}

// le fichier destination doit avoir la même taille que le source
if (lseek (file2, file_stat.st_size - 1, SEEK_SET) == -1) {
  perror("lseek");
  exit(EXIT_FAILURE);
}

// écriture en fin de fichier
if (write (file2, &dummy, sizeof(char)) != 1) {
  perror("write");
  exit(EXIT_FAILURE);
}

// mmap fichier source
if ((src = mmap (NULL, file_stat.st_size, PROT_READ, MAP_SHARED, file1, 0))  == (void *)(-1)) {
  perror("mmap(src)");
  exit(EXIT_FAILURE);
}

// mmap fichier destination
if ((dst = mmap (NULL, file_stat.st_size, PROT_READ | PROT_WRITE, MAP_SHARED, file2, 0)) == (voi
  perror("mmap(src)");
  exit(EXIT_FAILURE);
}

// copie complète
memcpy (dst, src, file_stat.st_size);

// libération mémoire
if(munmap(src,file_stat.st_size)<0) {
  perror("munmap(src)");
  exit(EXIT_FAILURE);
}

if(munmap(dst,file_stat.st_size)<0) {
  perror("munmap(dst)");
  exit(EXIT_FAILURE);
}
// fermeture fichiers
if(close(file1)<0) {
```

```
    perror("close(file1)");
    exit(EXIT_FAILURE);
  }

  if(close(file2)<0) {
    perror("close(file2)");
    exit(EXIT_FAILURE);
  }
  return(EXIT_SUCCESS);
}
```

8.1.7 Mémoire partagée

Dans les exemples précédents, nous avons supposé qu'il existait une correspondance biunivoque entre chaque page de la mémoire virtuelle et une page en mémoire RAM. C'est souvent le cas, mais ce n'est pas nécessaire. Il est tout à fait possible d'avoir plusieurs pages de la mémoire virtuelle qui appartiennent à des processus différents mais pointent vers la même page en mémoire physique. Ce partage d'une même page physique entre plusieurs pages de la mémoire virtuelle a plusieurs utilisations en pratique.

Revenons aux threads POSIX. Lorsqu'un processus crée un nouveau thread d'exécution, celui-ci a un accès complet au segment code, aux variables globales et au heap du processus. Par contre, le thread et le processus ont chacun un stack qui leur est propre. Comme nous l'avons indiqué lors de la présentation des threads, ceux-ci peuvent être implémentés en utilisant une librairie ou avec l'aide du système d'exploitation. Du point de vue de la mémoire, lorsqu'une librairie telle que *gnuth* est utilisée pour créer un thread, la librairie réserve une zone de mémoire sur le heap pour ce thread. Cette zone mémoire contient le stack qui est spécifique au thread. Celui-ci a été alloué en utilisant malloc(3) et a généralement une taille fixe. Avec la mémoire virtuelle, il est possible d'implémenter les threads plus efficacement avec l'aide du système d'exploitation. Lors de la création d'un thread, celui-ci va tout d'abord créer une nouvelle table des pages pour le thread. Celle-ci sera initialisée en copiant toutes les entrées de la table des pages du processus, sauf celles qui correspondent au stack. De cette façon, le processus *père* et le thread auront accès aux mêmes segments de code, aux même variables globales et au même heap. Toute modification faite par le processus père à une variable globale ou à une information stockée sur le heap sera immédiatement accessible au thread et inversement. L'entrée de la table des pages du thread correspondant à son stack pointera vers une page qui sera spécifique au thread. Cette page aura été initialisée par le système d'exploitation avec l'argument passé par le processus à la fonction pthread_create(3). La figure ci-dessous illustre ce partage de table des pages après la création d'un thread.

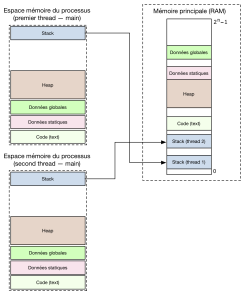


FIGURE 8.10 – Tables des pages après création d'un thread. Les segments de même nom et de même couleur sont partagés par les deux threads, sauf pour le segment de Stack qui est spécifique.

En exploitant intelligemment la table des pages, il est également possible de permettre à deux processus distincts d'avoir accès à la même zone de mémoire physique. Si deux processus peuvent accéder simultanément à la même zone de mémoire, ils peuvent l'utiliser pour communiquer plus efficacement qu'en utilisant des pipes par exemple. Cette technique porte le nom de *mémoire partagée*. Elle nécessite une modification de la table des pages des processus qui veulent partager une même zone mémoire. Pour comprendre le fonctionnement de cette *mémoire partagée*, considérons le cas de deux processus : *P1* et *P2* qui veulent pouvoir utiliser une page commune en mémoire. Pour cela, plusieurs interactions entre les processus et le système d'exploitation sont nécessaires comme nous allons le voir.

Avant de permettre à deux processus d'accéder à la même page en mémoire physique, il faut d'abord se poser la question de l'origine de cette page physique. Deux solutions sont possibles. La première est de prendre cette page parmi les pages qui appartiennent à l'un des processus, par exemple *P1*. Lorsque la page est partagée, le système d'exploitation peut modifier la table des pages du processus *P2* de façon à lui permettre d'y accéder. La seconde est que le noyau du système d'exploitation fournisse une nouvelle page qui pourra être partagée. Cette page "appartient" au noyau mais celui-ci la rend accessible aux processus *P1* et *P2* en modifiant leurs tables des pages. Linux utilise la seconde technique. Elle a l'avantage de permettre un meilleur contrôle par le système d'exploitation du partage de pages entre différents processus. De plus, lorsqu'une zone de mémoire partagée a été créée via le système d'exploitation, elle survit à la terminaison de ce processus. Une mémoire partagée créée par un processus peut donc être utilisée par d'autres processus.

Sous Linux, la mémoire partagée peut s'utiliser via les appels systèmes shmget(2), shmat(2) et shmdt(2). L'appel système shmget(2) permet de créer un segment de mémoire partagée. Le premier argument de shmget(2) est une clé qui identifie le segment de mémoire partagée. Cette clé est en pratique encodée sous la forme d'un entier qui identifie le segment de mémoire partagée. Elle sert d'identifiant du segment de mémoire partagée dans le noyau. Un processus doit connaître la clé qui identifie un segment de mémoire partagée pour pouvoir y accéder. Le deuxième argument de shmget(2) est la taille du segment. En pratique, celle-ci sera arrondie au multiple entier supérieur de la taille d'une page. Enfin, le troisième argument sont des drapeaux qui contrôlent la création du segment et les permissions qui y sont associées.

```
#include <sys/ipc.h>
#include <sys/shm.h>
int shmget(key_t key, size_t size, int shmflg);
```

L'appel système shmget(2) retourne un entier qui identifie le segment de mémoire partagée à l'intérieur du processus si il réussit et `-1` sinon. Il peut être utilisé de deux façons. Un processus peut appeler shmget(2) pour créer un nouveau segment de mémoire partagée. Pour cela, il choisit une clé unique qui identifie ce segment et utilise le drapeau `IPC_CREAT`. Celui-ci peut être combiné avec les drapeaux qui sont supportés par l'appel système open(2). Ainsi, le fragment de code ci-dessous permet de créer une page de mémoire partagée qui a `1252` comme identifiant et est accessible en lecture et en écriture par tous les processus qui appartiennent au même utilisateur ou au même groupe que le processus courant. Si cet appel à shmget(2) réussit, le segment de mémoire est initialisé à la valeur 0.

```
key_t key=1252;
int shm_id = shmget(key, 4096, IPC_CREAT | S_IRUSR | S_IWUSR | S_IRGRP | S_IWGRP );
if (shm_id == -1) {
  perror("shmget");
  exit(EXIT_FAILURE);
}
```

La fonction shmget(2) peut aussi être utilisée par un processus pour obtenir l'autorisation d'accéder à un segment de mémoire partagée qui a été créé par un autre processus. Dans ce cas, le drapeau `IPC_CREAT` n'est pas passé en argument.

Il est important de noter que si l'appel à shmget(2) réussit, cela indique que le processus dispose des permissions pour accéder au segment de mémoire partagée, mais à ce stade il n'est pas accessible depuis la table des pages du processus. Cette modification à la table des pages du processus se fait en utilisant shmat(2). Cet appel système permet d'attacher un segment de mémoire partagée à un processus. Il prend comme premier argument l'identifiant du segment de mémoire retourné par shmget(2). Le deuxième argument est un pointeur vers la zone mémoire via laquelle le segment doit être accessible dans l'espace d'adressage virtuel du processus. Généralement, c'est la valeur `NULL` qui est spécifiée comme second argument et le noyau choisit l'adresse à laquelle le segment de mémoire est attaché dans le processus. Il est aussi possible de spécifier une adresse dans l'espace d'adressage du processus. Le troisième argument permet, en utilisant le drapeau `SHM_RDONLY`, d'attacher le segment en lecture seule. shmat(2) retourne l'adresse à laquelle le segment a été attaché en cas de succès et `(void *) -1` en cas d'erreur.

```
#include <sys/types.h>
#include <sys/shm.h>

void *shmat(int shmid, const void *shmaddr, int shmflg);
```

```
int shmdt(const void *shmaddr);
```

L'appel système shmdt(2) permet de détacher un segment de mémoire qui avait été attaché en utilisant shmat(2). L'argument passé à shmdt(2) doit être l'adresse d'un segment de mémoire attaché préalablement par shmat(2). Lorsqu'un processus se termine, tous les segments auxquels il était attaché sont détachés lors de l'appel à exit(2). Cela n'empêche pas un programme de détacher correctement tous les segments de mémoire qu'il utilise avant de se terminer.

Le fragment de code ci-dessous présente comment un segment de mémoire peut être attaché et détaché après avoir été créé avec shmget(2).

```
void * addr = shmat(shm_id, NULL, 0);
if (addr == (void *) -1) {
  perror("shmat");
  exit(EXIT_FAILURE);
}
// ...
if(shmdt(addr)==-1) {
  perror("shmdt");
  exit(EXIT_FAILURE);
}
```

Note : Attention aux pointeurs en mémoire partagée

Lorsque deux processus partagent le même segment de mémoire partagée, ils ont tous les deux accès directement à la mémoire. Il est ainsi possible de stocker dans cette mémoire un tableau de nombres ou de caractères. Chacun des processus pourra facilement accéder aux données stockées dans le tableau. Il faut cependant être vigilant lorsque l'on veut stocker une structure de données utilisant des pointeurs dans un segment de mémoire partagée. Considérons une liste simplement chaînée. Cette liste peut être implémentée en utilisant une structure contenant la donnée stockée dans l'élément de la liste (par exemple un entier) et un pointeur vers l'élément suivant dans la liste (et `NULL` en fin de liste). Imaginons que les deux processus ont attaché le segment de mémoire destiné à contenir la liste avec l'appel shmat(2) présenté ci-dessus et que l'adresse retournée par shmat(2) est celle qui correspond au premier élément de la liste. Comme le système d'exploitation choisit l'adresse à laquelle le segment de mémoire partagée est stocké dans chaque processus, l'appel à shmat(2) retourne potentiellement une adresse différente dans les deux processus. Si ils peuvent tous les deux accéder au premier élément de la liste, il n'en sera pas de même pour le second élément. En effet, si cet élément a été créé par le premier processus, le pointeur contiendra l'adresse du second élément dans l'espace d'adressage virtuel du premier processus. Cette adresse ne correspondra en général pas à celle du second élément dans l'espace d'adressage du second processus. Pour cette raison, il est préférable de ne pas utiliser de pointeurs dans un segment de mémoire partagé.

Comme les segments de mémoire partagée sont gérés par le noyau du système d'exploitation, ils persistent après la terminaison du processus qui les a créé. C'est intéressant lorsque l'on veut utiliser des segments de mémoire partagée pour la communication entre plusieurs processus dont certains peuvent se terminer par une erreur. Malheureusement, le nombre de segments de mémoire partagée qui peuvent être utilisés sur un système Unix est borné. Lorsque la limite fixée par la configuration du noyau est atteinte, il n'est plus possible de créer de nouveau segment de mémoire partagée. Sous Linux ces limites sont visibles dans les fichiers `/proc/sys/kernel/shmmni` (nombre maximum d'identifiants de segments de mémoire partagée) et `/proc/sys/kernel/shmall` (taille totale maximale de la mémoire partagée) ou via shmctl(2). Cet appel système permet de réaliser de nombreuses fonctions de contrôle de la mémoire partagée et notamment la destruction de segments de mémoire partagée qui ont été créés par shmget(2). shmctl(2) s'appuie sur les structures de données qui sont maintenues par le noyau pour les segments de mémoire partagée. Lorsqu'un segment de mémoire partagée est crée, le noyau lui associe une structure de type `shmid_ds`.

```
struct shmid_ds {
  struct ipc_perm shm_perm;    /* Propriétaire et permissions */
  size_t          shm_segsz;   /* Taille du segment (bytes) */
  time_t          shm_atime;   /* Instant de dernier attach */
  time_t          shm_dtime;   /* Instant de dernier detach */
  time_t          shm_ctime;   /* Instant de dernière modification */
```

```
  pid_t           shm_cpid;     /* PID du créateur */
  pid_t           shm_lpid;     /* PID du dernier 'shmat(2)'_ / 'shmdt(2)'_ */
  shmatt_t        shm_nattch;   /* Nombre de processus attachés */
};
```

Ce descripteur de segment de mémoire partagée, décrit dans shmctl(2) contient plusieurs informations utiles. Son premier élément est une structure qui reprend les informations sur le propriétaire et les permissions qui ont été définies ainsi que la taille du segment. Le descripteur de segment comprend ensuite les instants auxquels les dernières opérations shmat(2), shmdt(2) et la dernière modification au segment ont été faites. Le dernier élément contient le nombre de processus qui sont actuellement attachés au segment. L'appel système shmctl(2) prend trois arguments. Le premier est un identifiant de segment de mémoire partagée retourné par shmget(2). Le deuxième est une constante qui spécifie une commande. Nous utiliserons uniquement la commande `IPC_RMID` qui permet de retirer le segment de mémoire partagée dont l'identifiant est passé comme premier argument. Si il n'y a plus de processus attaché au segment de mémoire partagée, celui-ci est directement supprimé. Sinon, il est marqué de façon à ce que le noyau retire le segment dès que le dernier processus s'en détache. shmctl(2) retourne `0` en cas de succès et `-1` en cas d'échec.

```
#include <sys/ipc.h>
#include <sys/shm.h>

int shmctl(int shmid, int cmd, struct shmid_ds *buf);
```

Le segment de mémoire partagée qui a été créé dans les exemples précédents peut être supprimé avec le fragment de code ci-dessous.

```
if (shmctl(shm_id, IPC_RMID, 0) != 0) {
   perror("shmctl");
   exit(EXIT_FAILURE);
}
```

En pratique, comme le noyau ne détruit un segment de mémoire partagée que lorsqu'il n'y a plus de processus qui y est attaché, il peut être utile de détruire le segment de mémoire partagée juste après avoir effectué l'appel shmat(2). C'est ce que l'on fera par exemple si un processus père utilise un segment de mémoire partagée pour communiquer avec son processus fils.

La mémoire partagée est utilisée non seulement pour permettre la communication entre processus, mais également avec les librairies partagées. Celles-ci sont chargées automatiquement lors de l'exécution d'un processus qui les utilise. Les instructions qui font partie de ces librairies partagées sont chargées dans la même zone mémoire que celle qui est utilisée pour la mémoire partagée. Sous Linux, cette zone mémoire est située entre le heap et le stack comme illustré dans la figure ci-dessous.

FIGURE 8.11 – Organisation en mémoire d'un processus

Lorsqu'il exécute un processus, le noyau maintient dans les structures de données qui sont relatives à ce processus la liste des segments de mémoire partagée et des librairies partagées qu'il utilise. Sous Linux, cette information est visible via le pseudo-système de fichiers `/proc`. Le fichier `/proc/PID/maps` représente de façon textuelle la table des segments de mémoire qui sont partagés dans le processus `PID`.

Un exemple d'un tel fichier *maps* est présenté ci-dessous. Il contient une carte de l'ensemble des pages qui appartiennent à un processus. Le fichier comprend six colonnes. La première est la zone de mémoire virtuelle. La seconde sont les bits de permission avec *r* pour la permission de lecture, *w* pour l'écriture et *x* pour l'exécution. Le dernier bit de permission est à la valeur *p* lorsque la page est en *copy-on-write* et *s* lorsqu'il s'agit d'un segment de mémoire partagé. Les trois dernières colonnes sont relatives au stockage des pages sur le disque.

```
00400000-00402000 r-xp 00000000 00:1a 49485798      /tmp/a.out
00602000-00603000 rw-p 00002000 00:1a 49485798      /tmp/a.out
3d3f200000-3d3f220000 r-xp 00000000 08:01 268543    /lib64/ld-2.12.so
3d3f41f000-3d3f420000 r--p 0001f000 08:01 268543    /lib64/ld-2.12.so
3d3f420000-3d3f421000 rw-p 00020000 08:01 268543    /lib64/ld-2.12.so
3d3f421000-3d3f422000 rw-p 00000000 00:00 0
3d3f600000-3d3f786000 r-xp 00000000 08:01 269510    /lib64/libc-2.12.so
3d3f786000-3d3f986000 ---p 00186000 08:01 269510    /lib64/libc-2.12.so
3d3f986000-3d3f98a000 r--p 00186000 08:01 269510    /lib64/libc-2.12.so
3d3f98a000-3d3f98b000 rw-p 0018a000 08:01 269510    /lib64/libc-2.12.so
3d3f98b000-3d3f990000 rw-p 00000000 00:00 0
3d3fa00000-3d3fa83000 r-xp 00000000 08:01 269516    /lib64/libm-2.12.so
3d3fa83000-3d3fc82000 ---p 00083000 08:01 269516    /lib64/libm-2.12.so
3d3fc82000-3d3fc83000 r--p 00082000 08:01 269516    /lib64/libm-2.12.so
3d3fc83000-3d3fc84000 rw-p 00083000 08:01 269516    /lib64/libm-2.12.so
7f7c57e42000-7f7c57e45000 rw-p 00000000 00:00 0
7f7c57e60000-7f7c57e61000 rw-s 00000000 00:04 66355276 /SYSV00000000
7f7c57e61000-7f7c57e63000 rw-p 00000000 00:00 0
7fffc479c000-7fffc47b1000 rw-p 00000000 00:00 0            [stack]
```

L'exemple ci-dessus présente la carte de mémoire d'un processus qui utilise trois librairies partagées. Le segment de mémoire partagée se trouve aux adresses virtuelles `7f7c57e60000-7f7c57e61000`. Il est accessible en lecture et en écriture.

8.1.8 Implémentation de fork(2)

La mémoire partagée joue un rôle clé dans l'exécution efficace des appels systèmes fork(2) et execve(2). Considérons d'abord fork(2). Cet appel est fondamental sur un système Unix. Au fil des années, les développeurs de Unix et de Linux ont cherché à optimiser ses performances. Une implémentation naïve de l'appel système fork(2) est de copier physiquement toutes les pages utilisées en mémoire RAM par le processus père. Ensuite, le noyau peut créer une table des pages pour le processus fils qui pointe vers les copies des pages du processus père. De cette façon, le processus père et le processus fils utilisent exactement les mêmes instructions. Ils peuvent donc poursuivre leur exécution à partir des mêmes données en mémoire. Mais chaque processus pourra faire les modifications qu'il souhaite aux données stockées en mémoire. Cette implémentation était utilisée par les premières versions de Unix, mais elle est peu efficace, notamment pour les processus qui consomment beaucoup de mémoire. Pour le shell, qui généralement exécute fork(2) suivi immédiatement par execve(2), copier l'entièreté de la mémoire du processus père est un pur gaspillage de ressources.

La mémoire virtuelle permet d'optimiser l'appel système fork(2) et de le rendre nettement plus rapide. Lors de la création d'un processus fils, le noyau du système d'exploitation commence par créer une table des pages pour le processus fils. En initialisant cette table avec les mêmes entrées que celles du processus père, le noyau permet aux deux processus d'accéder aux mêmes instructions et aux mêmes données. Pour les instructions se trouvant dans le segment code et dont les entrées de la table des pages sont généralement en *read-only*, cette solution fonctionne correctement. Le processus père et le processus fils peuvent exécuter exactement les mêmes instructions tout en n'utilisant qu'une seule copie de ces instructions en mémoire.

Malheureusement, cette solution ne fonctionne plus pour les pages contenant les données globales, le stack et le heap. En effet, ces pages doivent pouvoir être modifiées de façon indépendante par le processus père et le processus fils. C'est notamment le cas pour la zone mémoire qui contient la valeur de retour de l'appel système fork(2). Par définition, cette zone mémoire doit contenir une valeur différente dans le processus père et le processus fils. Pour éviter ce problème, le noyau pourrait copier physiquement les pages contenant les variables globales, le heap et le stack. Cela permettrait, notamment dans le cas de l'exécution de fork(2) par le shell d'améliorer les performances de fork(2) sans pour autant compromettre la sémantique de cet appel système. Il existe cependant une alternative à cette copie physique. Il s'agit de la technique du *copy-on-write*.

Sur un système qui utilise *copy-on-write*, l'appel système fork(2) est mis en œuvre de la façon suivante. Lors de l'exécution de fork(2), le noyau copie toutes les entrées de la table des pages du processus père vers la table des pages du processus fils. Ce faisant, le noyau modifie également les permissions de toutes les pages utilisées par le processus père. Les pages correspondant au segment de code sont toutes marquées en lecture seule. Les pages correspondant aux données globales, heap et stack sont marquées avec un statut spécial (*copy-on-write*).

Celui-ci autorise les accès en lecture à la page sans restriction. Si un processus tente un accès en écriture sur une de ces pages, le *MMU* interrompt l'exécution du processus et force l'exécution d'une routine d'interruption du noyau. Celle-ci analyse la tentative d'accès à la mémoire qui a échoué. Si la page était en lecture seule (par exemple une page du segment de code), un signal `SIGSEGV` est envoyé au processus concerné. Si par contre la page était marquée *copy-on-write*, alors le noyau alloue une nouvelle page physique et y recopie la page où la tentative d'accès a eu lieu. La table des pages du processus qui a fait la tentative d'accès est modifiée pour pointer vers la nouvelle page avec une permission en lecture et en écriture. La permission de l'entrée de la table des pages de l'autre processus est également modifiée de façon à autoriser les écritures et les lectures. Les deux processus disposent donc maintenant d'une copie différente de cette page et ils peuvent la modifier à leur guise. Cette technique de *copy-on-write* permet de ne copier que les pages qui sont modifiées par le processus père ou le processus fils. C'est un gain de temps appréciable par rapport à la copie complète de toutes les pages.

Dans le pseudo fichier `/proc/PID/maps` présenté précédemment, le bit *p* indique que les pages correspondantes sont en *copy-on-write*.

8.1.9 Interactions entre les processus et la mémoire

La mémoire virtuelle est gérée par le système d'exploitation. La plupart des processus supportent très bien d'avoir certaines de leurs pages qui sont sauvegardées sur disque lorsque la mémoire est saturée. Cependant, dans certains cas très particuliers, il peut être intéressant pour un processus de contrôler quelles sont ses pages qui restent en mémoire et quelles sont celles qui sont stockées sur le disque dur. Linux fournit plusieurs appels systèmes qui permettent à un processus de surveiller et éventuellement d'influencer la stratégie de remplacement des pages.

```
#include <unistd.h>
#include <sys/mman.h>
int mincore(void *addr, size_t length, unsigned char *vec);
```

L'appel système mincore(2) permet à un processus d'obtenir de l'information sur certaines de ses pages. Il prend comme premier argument une adresse, qui doit être un multiple de la taille des pages. Le deuxième est la longueur de la zone mémoire pour laquelle le processus demande de l'information. Le dernier argument est un pointeur vers une chaîne de caractères qui doit contenir autant de caractères que de pages dans la zone de mémoire analysée. Cette chaîne de caractères contiendra des drapeaux pour chaque page de la zone mémoire concernée. Les principaux sont :
– `MINCORE_INCORE` indique que la page est en mémoire RAM
– `MINCORE_REFERENCED` indique que la page a été référencée
– `MINCORE_MODIFIED` indique que la page a été modifiée
Un exemple d'utilisation de mincore(2) est repris dans le programme `/MemoireVirtuelle/src/mincore.c` ci-dessous.

```
#dfine _BSD_SOURCE

#define SIZE 10*4096

int main(int argc, char *argv[]) {

  char *mincore_vec;
  size_t page_size = getpagesize();
  size_t page_index;
  char *mem;

  mem=(char *)malloc(SIZE*sizeof(char));
  if(mem==NULL) {
    perror("malloc");
    exit(EXIT_FAILURE);
  }
  for(int i=0;i<SIZE/2;i++) {
    *(mem+i)='A';
  }

  mincore_vec = calloc(sizeof(char), SIZE/page_size);
```

```
  if(mincore_vec==NULL) {
    perror("calloc");
    exit(EXIT_FAILURE);
  }
  if(mincore(mem, SIZE, mincore_vec)!=0) {
    perror("mincore");
    exit(EXIT_FAILURE);
  }

  for (page_index = 0; page_index < SIZE/page_size; page_index++) {
    printf("%lu :", (unsigned long)page_index);
    if (mincore_vec[page_index]&MINCORE_INCORE) {
      printf(" incore");
    }
    if (mincore_vec[page_index]&MINCORE_REFERENCED) {
      printf(" referenced_by_us");
    }
    if (mincore_vec[page_index]&MINCORE_MODIFIED) {
      printf(" modified_by_us");
    }
    printf("\n");
  }
  free(mincore_vec);
  free(mem);
  return (EXIT_SUCCESS);
}
```

Certains processus doivent pouvoir contrôler la façon dont leurs pages sont stockées en mémoire RAM ou sur disque. C'est le cas notamment des processus qui manipulent des clés cryptographiques. Pour des raisons de sécurité, il est préférable que ces clés ne soient pas sauvegardées sur le disque dur. Pour des raisons de performances, certains processus préfèrent également éviter que leurs pages ne soient stockées sur disque dur. Linux comprend plusieurs appels systèmes qui permettent à un processus d'influencer la stratégie de remplacement des pages du noyau.

```
#include <sys/mman.h>

int mlock(const void *addr, size_t len);
int munlock(const void *addr, size_t len);

int mlockall(int flags);
int munlockall(void);
```

L'appel système mlock(2) permet de forcer un ensemble de pages à rester en mémoire RAM tandis que munlock(2) fait l'inverse. L'appel système mlockall(2) quant à lui permet à un processus de demander que tout l'espace d'adressage qu'il utilise reste en permanence en mémoire RAM. mlockall(2) prend comme argument des drapeaux. Actuellement deux drapeaux sont supportés. `MCL_CURRENT` indique que toutes les pages actuellement utilisées par le processus doivent rester en mémoire. `MCL_FUTURE` indique que toutes les pages qui seront utilisées par le processus devront être en mémoire RAM au fur et à mesure de leur allocation. L'appel système madvise(2) permet également à un processus de donner des indications sur la façon dont il utilisera ses pages dans le futur.

Ces appels systèmes doivent être utilisés avec précaution. En forçant certaines de ses pages à rester en mémoire, un processus perturbe le bon fonctionnement de la mémoire virtuelle, ce qui risque de perturber les performances globales du système. En pratique, en dehors d'applications cryptographiques et de processus avec des exigences d'exécution en temps réel dont la description dépasse le cadre de ce cours, il n'y a pas d'intérêt à utiliser ces appels système.

L'utilisation de la mémoire influence fortement les performances des processus. Plusieurs utilitaires sous Linux permettent de mesurer la charge d'un système. Certains offrent une interface graphique. D'autres, comme top(1) ou vmstat(8) s'utilisent directement depuis un terminal. La commande vmstat(8) permet de suivre l'évolution de la charge du système et de la mémoire virtuelle.

```
$ vmstat 5
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
3  0 662260 259360  49044 583808    0    0     1     1    1    2  5  0 95  0  0
3  0 662260 259352  49044 583808    0    0     0     0 6018 4182 61  0 39  0  0
7  1 662260 254856  49044 587328    0    0     0     0 6246 4309 61  0 38  2  0
4  0 662260 221668  49044 608652    0    0     0     0 9731 5520 62  1  9 28  0
3  0 662260 260368  49044 583836    0    0     0     2 8291 4868 70  1 12 17  0
4  0 662260 260548  49044 583836    0    0     0     0 6042 4174 61  0 39  0  0
3  0 662260 260720  49044 583836    0    0     0     0 6003 4242 61  0 39  0  0
```

Lorsqu'elle est utilisée, la commande vmstat(8) retourne de l'information sur l'état du système. Dans l'exemple ci-dessus, elle affiche cet état toutes les 5 secondes. La sortie de vmstat(8) comprend plusieurs informations utiles.
– la colonne `procs` reprend le nombre de processus qui sont dans l'état *Running* ou *Blocked*.
– la colonne `memory` reprend l'information relative à l'utilisation de la mémoire. La colonne *swpd* est la quantité de mémoire virtuelle utilisée. La colonne *free* est la quantité de mémoire RAM qui est actuellement libre. Les colonnes `buff` et `cache` sont relatives aux buffers et à la cache qui sont utilisés par le noyau.
– la colonne `swap` reprend la quantité de mémoire qui a été transférée sur le disque (*so*) ou lue du disque (*si*)
– la colonne `io` reprend le nombre de blocs lus du disque (*bi*) et écrits (*bo*)
– la colonne `system` reprend le nombre d'interruptions (*in*) et de changements de contexte (*cs*)
– la dernière colonne (`cpu`) fournit des statistiques sur l'utilisation du cpu
Enfin, notons que l'appel système getrusage(2) peut être utilisé par un processus pour obtenir de l'information sur son utilisation des ressources du système et notamment les opérations de transfert de pages entre le mémoire RAM et les dispositifs de stockage.

8.1.10 execve(2) et la mémoire virtuelle

Pour terminer ce survol de la mémoire virtuelle, il est intéressant de revenir sur le fonctionnement de l'appel système execve(2). Celui-ci permet de remplacer l'image mémoire du processus en cours d'exécution par un exécutable chargé depuis le disque. L'appel système execve(2) utilise fortement la mémoire virtuelle. Plutôt que de copier l'entièreté de l'exécutable en mémoire, il se contente de créer les entrées correspondants au segment de code et aux variables globales dans la table des pages du processus. Ces entrées pointent vers le fichier exécutable qui est sauvegardé sur le disque. En pratique, celui-ci est découpé en deux parties. La première contient les instructions. Celles-ci sont mappées sous forme de pages sur lesquelles le processus a les permissions d'exécution et de lecture. La seconde partie du fichier correspond à la zone data, contenant les variables globales. Celle-ci est mappée dans des pages qui sont accessibles en lecture et en écriture. En pratique, la technique du *copy-on-write* est utilisée pour ne créer une copie spécifique au processus que si celui-ci modifie certaines données en mémoire. Ensuite, les pages relatives au *heap* et au *stack* peuvent être créées.

La sortie ci-dessous présente le contenu de `/proc/PID/maps` lors de l'exécution du programme `/MemoireVirtuelle/src/cp2.c` présenté plus haut. Celui-ci utilise deux fichiers mappés en mémoire avec mmap(2). Ceux-ci apparaissent dans la table des pages du processus, tout comme l'exécutable. Au niveau des permissions, on remarquera que le fichier source est mappé avec des pages en lecture seule tandis que le fichier destination est mappé en lecture/écriture.

```
$ cat /proc/29039/maps
00400000-004ab000 r-xp 00000000 00:19 49479785                           /etinfo/users2/obo/sinf1
006ab000-006ac000 rw-p 000ab000 00:19 49479785                           /etinfo/users2/obo/sinf1
006ac000-006af000 rw-p 00000000 00:00 0
00bc7000-00bea000 rw-p 00000000 00:00 0                                  [heap]
7fc43d5e4000-7fc43d6a4000 rw-s 00000000 00:19 51380745                   /etinfo/users2/obo/sinf1
7fc43d6a4000-7fc43d764000 r--s 00000000 00:19 49479785                   /etinfo/users2/obo/sinf1
7fff5b912000-7fff5b927000 rw-p 00000000 00:00 0                          [stack]
7fff5b9ff000-7fff5ba00000 r-xp 00000000 00:00 0                          [vdso]
ffffffffff600000-ffffffffff601000 r-xp 00000000 00:00 0                  [vsyscall]
```

CHAPITRE 9

Fichiers

9.1 Gestion des utilisateurs

Unix est un système d'exploitation multi-utilisateurs. Un tel système impose des contraintes de sécurité qui n'existent pas sur un système mono-utilisateur. Il est intéressant de passer en revue quelques unes de ces contraintes :
– il doit être possible d'identifier et/ou d'authentifier les utilisateurs du système
– il doit être possible d'exécuter des processus appartenant à plusieurs utilisateurs simultanément et de déterminer quel utilisateur est responsable de chaque opération
– le système d'exploitation doit fournir des mécanismes simples qui permettent de contrôler l'accès aux différentes ressources (mémoire, stockage, ...).
– il doit être possible d'allouer certaines ressources à un utilisateur particulier à un moment donné
Aujourd'hui, la plupart des systèmes informatiques demandent une authentification de l'utilisateur sous la forme d'un mot de passe, d'une manipulation particulière voire d'une identification biométrique comme une empreinte digitale. Cette authentification permet de vérifier que l'utilisateur est autorisé à manipuler le système informatique. Cela n'a pas toujours été le cas et de nombreux systèmes informatiques plus anciens étaient conçus pour être utilisés par un seul utilisateur qui était simplement celui qui interagissait physiquement avec l'ordinateur.

Les systèmes Unix supportent différents mécanismes d'authentification. Le plus simple et le plus utilisé est l'authentification par mot de passe. Chaque utilisateur est identifié par un nom d'utilisateur et il doit prouver son identité en tapant son mot de passe au démarrage de toute session sur le système. En pratique, une session peut s'établir localement sur l'ordinateur via son interface graphique par exemple ou à distance en faisant tourner un serveur tel que sshd(8) sur le système Unix et en permettant aux utilisateurs de s'y connecter via Internet en utilisant un client ssh(1). Dans les deux cas, le système d'exploitation lance un processus login(1) qui permet de vérifier le nom d'utilisateur et le mot de passe fourni par l'utilisateur. Si le mot de passe correspond à celui qui est stocké sur le système, l'utilisateur est authentifié et son shell peut démarrer. Sinon, l'accès au système est refusé.

Lorsqu'un utilisateur se connecte sur un système Unix, il fournit son nom d'utilisateur ou *username*. Ce nom d'utilisateur est une chaîne de caractères qui est facile à mémoriser par l'utilisateur. D'un point de vue de mise en œuvre, un système d'exploitation préfère manipuler des valeurs numériques plutôt que des chaînes de caractères. Unix associe à chaque utilisateur un identifiant qui est stocké sous la forme d'un nombre entier positif. La table de correspondance entre l'identifiant d'utilisateur et le nom d'utilisateur est le fichier */etc/passwd*. Ce fichier texte, comme la grande majorité des fichiers de configuration d'un système Unix, comprend pour chaque utilisateur l'information suivante :
– nom d'utilisateur (*username*)
– mot de passe (sur les anciennes versions de Unix)
– identifiant de l'utilisateur (*userid*)
– identifiant du groupe principal auquel l'utilisateur appartient
– nom et prénom de l'utilisateur
– répertoire de démarrage de l'utilisateur
– shell de l'utilisateur

L'extrait ci-dessous présente un exemple de fichier `/etc/passwd`. Des détails complémentaires sont disponibles dans la page de manuel passwd(5). Un utilisateur peut modifier les informations le concernant dans ce fichier avec la commande passwd(1).

```
# Exemple de /etc/passwd
nobody:*:-2:-2:Unprivileged User:/var/empty:/usr/bin/false
root:*:0:0:System Administrator:/var/root:/bin/sh
daemon:*:1:1:System Services:/var/root:/usr/bin/false
slampion:*:1252:1252:Séraphin Lampion:/home/slampion:/bin/bash
```

Il y a en pratique trois types d'utilisateurs sur un système Unix. L'utilisateur *root* est l'administrateur du système. C'est l'utilisateur qui a le droit de réaliser toutes les opérations sur le système. Il peut créer de nouveaux utilisateurs, mais aussi formatter les disques, arrêter le système, interrompre des processus utilisateurs ou accéder à l'ensemble des fichiers sans restriction. Par convention, cet utilisateur a l'identifiant `0`. Ensuite, il y a tous les utilisateurs *normaux* du système Unix. Ceux-ci ont le droit d'accéder à leurs fichiers, d'interagir avec leurs processus mais en général ne peuvent pas manipuler les fichiers d'autres utilisateurs ou interrompre leurs processus. L'utilisateur *slampion* dans l'exemple ci-dessus est un utilisateur *normal*. Enfin, pour faciliter l'administration du système, certains systèmes Unix utilisent des utilisateurs qui correspondent à un service particulier comme l'utilisateur *daemon* dans l'exemple ci-dessus. Une discussion de ce type d'utilisateur sort du cadre de ces notes. Le lecteur intéressé pourra consulter une référence sur l'administration des système Unix telle que [Adelstein-Lubanovic2007] ou [Nemeth+2010].

Unix associe à chaque processus un identifiant d'utilisateur. Cet identifiant est stocké dans l'entrée du processus dans la table des processus. Un processus peut récupérer son identifiant d'utilisateur via l'appel système getuid(2). Outre cet appel système, il existe également l'appel système setuid(2) qui permet de modifier le *userid* du processus en cours d'exécution. Pour des raisons évidentes de sécurité, seul un processus appartenant à l'administrateur système (*root*) peut exécuter cet appel système. C'est le cas par exemple du processus login(1) qui appartient initialement à *root* puis exécute setuid(2) afin d'appartenir à l'utilisateur authentifié puis exécute execve(2) pour lancer le premier shell appartenant à l'utilisateur.

En pratique, il est parfois utile d'associer des droits d'accès à des groupes d'utilisateurs plutôt qu'à un utilisateur particulier. Par exemple, un département universitaire peut avoir un groupe correspondant à tous les étudiants et un autre aux membres du staff pour leur donner des permissions différentes. Un utilisateur peut appartenir à un groupe principal et plusieurs groupes secondaires. Le groupe principal est spécifié dans le fichier passwd(5) tandis que le fichier `/etc/group` décrit dans group(5) contient les groupes secondaires.

9.2 Systèmes de fichiers

Outre un processeur et une mémoire, la plupart des ordinateurs actuels sont équipés d'un ou plusieurs dispositifs de stockage. Les dispositifs les plus courants sont le disque dur ou le SSD, le lecteur de DVD-ROM, les clés USB et cartes mémoire, ... Ces dispositifs de stockage ont des caractéristiques techniques très différentes. Certains stockent l'information sous forme magnétique, d'autres sous forme électrique ou en creusant en utilisant un laser des trous dans un support physique. D'un point de vue logique, ils offrent pourtant tous une interface très similaire au système d'exploitation qui veut les utiliser.

Dans un système de fichiers Unix, l'ensemble des répertoires et fichiers est organisé sous la forme d'un arbre. La racine de cet arbre est le répertoire `/`. Il est localisé sur un des dispositifs de stockage du système. Le système de fichiers Unix permet d'intégrer facilement des systèmes de fichiers qui se trouvent sur différents dispositifs de stockage. Cette opération est en général réalisée par l'administrateur système en utilisant la commande mount(8). A titre d'exemple, voici quelques répertoires qui sont montés sur un système Linux.

```
$ df -k
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/sda1             15116836   9425020   4923912  66% /
tmpfs                  1559516      4724   1554792   1% /dev/shm
/dev/sda2             19840924   4374616  14442168  24% /xendomains
xenstore               1972388        40   1972348   1% /var/lib/xenstored
david:/mnt/student   258547072 188106176  59935296  76% /etinfo/users
david:/mnt/staff     254696800 177422400  64136160  74% /etinfo/users2
```

Dans l'exemple ci-dessus, la première colonne correspond au dispositif de stockage qui contient les fichiers et répertoires. Le premier dispositif de stockage `/dev/sda1/` est un disque local et contient le répertoire racine du système de fichiers. Le système de fichiers `david://mnt/student` est stocké sur le serveur `david` et est monté via mount(8) dans le répertoire `/etinfo/users`. Ainsi, tout accès à un fichier dans le répertoire `/etinfo/users` se fera via le serveur `david`.

Chaque répertoire du système de fichiers contient un ou plusieurs répertoires et un ou plusieurs fichiers. A titre d'exemple, il est intéressant de regarder le contenu de deux répertoires. Le premier est un extrait au contenu du répertoire racine obtenu avec la commande ls(1)

```
$ ls -la /
dr-xr-xr-x.  26 root root 233472 Feb 23 03:18 .
dr-xr-xr-x.  26 root root 233472 Feb 23 03:18 ..
-rw-r--r--    1 root root      0 Feb 13 16:45 .autofsck
-rw-r--r--    1 root root      0 Jul 27  2011 .autorelabel
dr-xr-xr-x.   4 root root   4096 Dec 15 05:50 boot
drwxr-xr-x   19 root root   4160 Mar 22 12:04 dev
drwxr-xr-x. 125 root root  12288 Mar 22 12:04 etc
drwxr-xr-x    4 root root      0 Mar 22 10:22 etinfo
drwxr-xr-x.   2 root root   4096 Jan  6  2011 home
dr-xr-xr-x.  14 root root   4096 Mar 22 03:26 lib
dr-xr-xr-x.  10 root root  12288 Mar 22 03:26 lib64
drwx------.   2 root root  16384 Jul 27  2011 lost+found
...
drwxrwxrwt. 104 root root   4096 Mar 22 12:05 tmp
drwxr-xr-x.  13 root root   4096 Jul 19  2011 usr
drwxr-xr-x.  23 root root   4096 Jul 27  2011 var
```

Le répertoire racine contient quelques fichiers et des répertoires. Tout répertoire contient deux répertoires spéciaux. Le premier répertoire, identifié par le caractère `.` (un seul point) est un alias vers le répertoire lui-même. Cette entrée de répertoire est présente dans chaque répertoire dès qu'il est créé avec une commande telle que mkdir(1). Le deuxième répertoire spécial est `..` (deux points consécutifs). Ce répertoire est un alias vers le répertoire parent du répertoire courant.

Les méta-données qui sont associées à chaque fichier ou répertoire contiennent, outre les informations de type, les bits de permission. Ceux-ci permettent d'encoder trois types de permissions et d'autorisation :
– `r` : autorisation de lecture
– `w` : autorisation d'écriture ou de modification
– `x` : autorisation d'exécution (fichiers) ou de parcours (répertoires)
Ces bits de permissions sont regroupés en trois blocs. Le premier bloc correspond aux bits de permission qui sont applicables pour les accès qui sont effectués par un processus qui appartient à l'utilisateur qui est propriétaire du fichier/répertoire. Le deuxième bloc correspond aux bits de permission qui sont applicables pour les opérations effectuées par un processus dont l'identifiant de groupe est identique à l'identifiant de groupe du fichier/répertoire mais n'appartient pas à l'utilisateur qui est propriétaire du fichier/répertoire. Le dernier bloc est applicable pour les opérations effectuées par des processus qui appartiennent à d'autres utilisateurs.

Les valeurs de ces bits sont représentés pas les symboles `rwx` dans l'output de la commande ls(1). Les bits de permission peuvent être modifiés en utilisant la commande chmod(1) qui utilise l'appel système chmod(2). Pour qu'un exécutable puisse être exécuté via l'appel système execve(2), il est nécessaire que le fichier correspondant possède les bits de permission `r` et `x`.

La signification du bit `x` est différente pour les répertoires et pour les fichiers. Ce bit indique qu'un fichier peut être utilisé par l'appel système execve(2) pour amorcer l'exécution d'un programme. Pour un répertoire, le bit `x` donne le droit de parcours, c'est à dire le droit d'accéder des fichiers ou sous-répertoires présents dans ce répertoire. L'opération de lister le contenu d'un répertoire (avec la commande ls(1) par exemple) nécessite le droit `r`, comme le montre l'exemple suivant.

```
$ mkdir -p repertoire
$ echo "LINFO1252" > repertoire/fichier
$ chmod 000 repertoire/
$ ls -al repertoire/
ls: cannot open directory repertoire/: Permission denied
```

```
$ cat repertoire/fichier
cat: repertoire/fichier: Permission denied
$ chmod +x repertoire/
$ ls -al repertoire/
ls: cannot open directory repertoire/: Permission denied
$ cat repertoire/fichier
LINFO1252
```

Pour un répertoire, il indique que le répertoire peut être listé (par exemple, en utilisant)

Note : Manipulation des bits de permission avec chmod(2)

L'appel système chmod(2) permet de modifier les bits de permission qui sont associés à un fichier. Ceux-ci sont encodés sous la forme d'un entier sur 16 bits.

- `S_IRUSR (00400)` : permission de lecture par le propriétaire
- `S_IWUSR (00200)` : permission d'écriture par le propriétaire
- `S_IXUSR (00100)` : permission d'exécution par le propriétaire
- `S_IRGRP (00040)` : permission de lecture par le groupe hormis le propriétaire
- `S_IWGRP (00020)` : permission d'écriture par le groupe hormis le propriétaire
- `S_IXGRP (00010)` : permission d'exécution par le groupe hormis le propriétaire
- `S_IROTH (00004)` : permission de lecture par tout utilisateur hormis le propriétaire et son groupe
- `S_IWOTH (00002)` : permission d'écriture par tout utilisateur hormis le propriétaire et son groupe
- `S_IXOTH (00001)` : permission d'exécution par tout utilisateur hormis le propriétaire et son groupe

```
#include <sys/stat.h>
int chmod(const char *path, mode_t mode);
```

Ces bits de permissions sont généralement spécifiés soit sous la forme d'une disjonction logique ou sous forme numérique. A titre d'exemple, un fichier qui peut être lu et écrit uniquement par son propriétaire aura comme permissions `00600` ou `S_IRUSR|S_IWUSR`.

Les bits de permission de bits de poids fort servent à encoder des permissions particulières relatives aux fichiers et répertoires. Par exemple, lorsque la permission `S_ISUID (04000)` est associée à un exécutable, elle indique que celui-ci doit s'exécuter avec les permissions du propriétaire de l'exécutable et pas les permissions de l'utilisateur. Cette permission spéciale est utilisée par des programmes comme passwd(1) qui doivent disposer des permissions de l'administrateur système pour s'exécuter correctement (passwd(1) doit modifier le fichier passwd(5) qui appartient à l'administrateur système).

Les exemples ci-dessous présentent le contenu partiel d'un répertoire.

```
$ ls -lai /etinfo/users/obo
total 1584396
drwx------ 78 obo  stafinfo       4096 Mar 17 00:34 .
drwxr-xr-x 93 root root           4096 Feb 22 11:37 ..
-rwxr-xr-x  1 obo  stafinfo      11490 Feb 28 00:43 a.out
-rw-------  1 obo  stafinfo       4055 Mar 22 15:13 .bash_history
-rw-r--r--  1 obo  stafinfo         55 Sep 18  1995 .bash_profile
-rw-r--r--  1 obo  stafinfo        101 Aug 28  2003 .bashrc
drwxr-xr-x  2 obo  stafinfo       4096 Nov 22  2004 bin
-rw-r--r--  1 obo  stafinfo        346 Feb 13 15:37 hello.c
drwxr-xr-x  3 obo  stafinfo       4096 Mar  2 09:30 sinf1252
drwxr-xr-x  2 obo  stafinfo       4096 May 17  2011 src
```

Dans un système Unix, que ce soit au niveau du shell ou dans n'importe quel processus écrit par exemple en langage C, les fichiers peuvent être spécifiés de deux façons. La première est d'indiquer le chemin complet depuis la racine qui permet d'accéder au fichier. Le chemin `/etinfo/users/obo` passé comme argument à la commande ls(1) ci-dessus en est un exemple. Le premier caractère `/` correspond à la racine du système de fichiers et ensuite ce caractère est utilisé comme séparateur entre les répertoires successifs. Ainsi, le fichier

`/etinfo/users/obo/hello.c` est un fichier qui a comme nom `hello.c` qui se trouve dans un répertoire nommé `obo` qui lui-même se trouve dans le répertoire `users` qui est dans le répertoire baptisé `etinfo` dans le répertoire racine. La seconde façon de spécifier un nom de fichier est de préciser son nom relatif. Pour éviter de forcer l'utilisateur à spécifier chaque fois le nom complet des fichiers et répertoires auxquels il veut accéder, le noyau maintient dans sa table des processus le *répertoire courant* de chaque processus. Par défaut, lorsqu'un processus est lancé, son répertoire courant est le répertoire à partir duquel le programme a été lancé. Ainsi, lorsque l'utilisateur tape une commande comme `gcc hello.c` depuis son shell, le processus gcc(1) peut directement accéder au fichier `hello.c` qui se situe dans le répertoire courant. Un processus peut modifier son répertoire courant en utilisant l'appel système chdir(2).

```
#include <unistd.h>

int chdir(const char *path);
```

Cet appel système prend comme argument une chaîne de caractères contenant le nom du nouveau répertoire courant. Ce nom peut être soit un nom complet (commençant par `/`), ou un nom relatif au répertoire courant actuel. Dans ce cas, il est parfois utile de pouvoir référer au répertoire parent du répertoire courant. Cela se fait en utilisant `..`. Dans chaque répertoire, cet alias correspond au répertoire parent. Ainsi, si le répertoire courant est `/etinfo/users`, alors le répertoire `../../bin` est le répertoire `bin` se trouvant dans le répertoire racine. Depuis le shell, il est possible de modifier le répertoire courant avec la commande cd(1posix). La commande pwd(1) affiche le répertoire courant actuel.

Il existe plusieurs appels systèmes et fonctions de la librairie standard qui permettent de parcourir le système de fichiers. Les principaux sont :

- l'appel système stat(2) permet de récupérer les méta-données qui sont associées à un fichier ou un répertoire. La commande stat(1) fournit des fonctionnalités similaires depuis le shell.
- les appels systèmes chmod(2) et chown(2) permettent de modifier respectivement le mode (i.e. les permissions), le propriétaire et le groupe associés à un fichier. Les commandes chmod(1), chown(1) et chgrp(1) permettent de faire de même depuis le shell.
- l'appel système utime(2) permet de modifier les dates de création/modification associées à un fichier/répertoire. Cet appel système est utilisé par la commande touch(1)
- l'appel système rename(2) permet de changer le nom d'un fichier ou d'un répertoire. Il est utilisé notamment par la commande rename(1)
- l'appel système mkdir(2) permet de créer un répertoire alors que l'appel système rmdir(2) permet d'en supprimer un
- les fonctions de la librairie opendir(3), closedir(3), et readdir(3) permettent de consulter le contenu de répertoires.

Les fonctions de manipulation des répertoires méritent que l'on s'y attarde un peu. Un répertoire est un fichier qui a une structure spéciale. Ces trois fonctions permettent d'en extraire de l'information en respectant le format d'un répertoire. Pour accéder à un répertoire, il faut d'abord l'ouvrir en utilisant opendir(3). La fonction readdir(3) permet d'accéder aux différentes entrées de ce répertoire et closedir(3) doit être utilisée lorsque l'accès n'est plus nécessaire. La fonction readdir(3) permet de manipuler la structure `dirent` qui est définie dans bits/dirent.h.

```
struct dirent {
    ino_t          d_ino;       /* inode number */
    off_t          d_off;       /* offset to the next dirent */
    unsigned short d_reclen;    /* length of this record */
    unsigned char  d_type;      /* type of file; not supported
                                   by all file system types */
    char           d_name[256]; /* filename */
};
```

Cette structure comprend le numéro de l'inode, c'est-à-dire la métadonnée qui contient les informations relatives au fichier/répertoire, la position de l'entrée `dirent` qui suit, la longueur de l'entrée, son type et le nom de l'entrée dans le répertoire. Chaque appel à readdir(3) retourne un pointeur vers une structure de ce type.

L'extrait de code ci-dessous permet de lister tous les fichiers présents dans le répertoire `name`.

```
#include <errno.h>
#include <stdio.h>
#include <stdlib.h>
```

```c
#include <dirent.h>

void exit_on_error(char *s) {
  perror(s);
  exit(EXIT_FAILURE);
}

int main (int argc, char *argv[])  {

  DIR *dirp;
  struct dirent *dp;
  char name[]=".";
  dirp = opendir(name);
  if(dirp==NULL) {
    exit_on_error("opendir");
  }
  while ((dp = readdir(dirp)) != NULL) {
    printf("%s\n",dp->d_name);
  }
  int err = closedir(dirp);
  if(err<0) {
    exit_on_error("closedir");
  }


}
```

La lecture d'un répertoire avec readdir(3) commence au début de ce répertoire. A chaque appel à readdir(3), le programme appelant récupère un pointeur vers une zone mémoire contenant une structure `dirent` avec l'entrée suivante du répertoire ou `NULL` lorsque la fin du répertoire est atteinte. Si une fonction doit relire à nouveau un répertoire, cela peut se faire en utilisant seekdir(3) ou rewinddir(3).

Note : readdir(3) et les threads

La fonction readdir(3) est un exemple de fonction non-réentrante qu'il faut éviter d'utiliser dans une application dont plusieurs threads doivent pouvoir parcourir le même répertoire. Ce problème est causé par l'utilisation d'une zone de mémoire `static` afin de stocker la structure dont le pointeur est retourné par readdir(3). Dans une application utilisant plusieurs threads, il faut utiliser la fonction readdir_r(3) :

```c
int  readdir_r(DIR *restrict dirp, struct dirent *restrict entry,
         struct dirent **restrict result);
```

Cette fonction prend comme arguments le pointeur `entry` vers un buffer propre à l'appelant qui permet de stocker le résultat de readdir_r(3).

Les appels systèmes link(2) et unlink(2) sont un peu particuliers et méritent une description plus détaillée. Sous Unix, un *inode* est associé à chaque fichier mais l'*inode* ne contient pas le nom de fichier parmi les méta-données qu'il stocke. Par contre, chaque *inode* contient un compteur (`nlinks`) du nombre de liens vers un fichier. Cela permet d'avoir une seule copie d'un fichier qui est accessible depuis plusieurs répertoires. Pour comprendre cette utilisation des liens sur un système de fichiers Unix, considérons le scénario suivant.

```
$ mkdir a
$ mkdir b
$ cd a
$ echo "test" > test.txt
$ cd ..
$ ln a/test.txt a/test2.txt
$ ls -li a
total 16
9624126 -rw-r--r--  2 obo  stafinfo  5 24 mar 21:14 test.txt
9624126 -rw-r--r--  2 obo  stafinfo  5 24 mar 21:14 test2.txt
```

```
$ ln a/test.txt b/test3.txt
$ stat --format "inode=%i nlinks=%h" b/test3.txt
inode=9624126 nlinks=3
$ ls -li b
total 8
9624126 -rw-r--r--  3 obo  stafinfo  5 24 mar 21:14 test3.txt
$ echo "complement" >> b/test3.txt
$ ls -li a
total 16
9624126 -rw-r--r--  3 obo  stafinfo  16 24 mar 21:15 test.txt
9624126 -rw-r--r--  3 obo  stafinfo  16 24 mar 21:15 test2.txt
$ ls -li b
total 8
9624126 -rw-r--r--  3 obo  stafinfo  16 24 mar 21:15 test3.txt
$ cat b/test3.txt
test
complement
$ cat a/test.txt
test
complement
$ rm a/test2.txt
$ ls -li a
total 8
9624126 -rw-r--r--  2 obo  stafinfo  16 24 mar 21:15 test.txt
$ rm a/test.txt
$ ls -li a
$ ls -li b
total 8
9624126 -rw-r--r--  1 obo  stafinfo  16 24 mar 21:15 test3.txt
```

Dans ce scénario, deux répertoires sont créés avec la commande mkdir(1). Ensuite, la commande echo(1) est utilisée pour créer le fichier `test.txt` contenant la chaîne de caractères `test` dans le répertoire `a`. Ce fichier est associé à l'*inode* `9624126`. La commande ln(1) permet de rendre ce fichier accessible sous un autre nom depuis le même répertoire. La sortie produite par la commande ls(1) indique que ces deux fichiers qui sont présents dans le répertoire `a` ont tous les deux le même `inode`. Ils correspondent donc aux mêmes données sur le disque. A ce moment, le compteur `nlinks` de l'*inode* `9624126` a la valeur `2`. La commande ln(1) peut être utilisée pour créer un lien vers un fichier qui se trouve dans un autre répertoire [1] comme le montre la création du fichier `test3.txt` dans le répertoire `b`. Ces trois fichiers correspondant au même *inode*, toute modification à l'un des fichiers affecte et est visible dans n'importe lequel des liens vers ce fichier. C'est ce que l'on voit lorsque la commande `echo "complement" >> b/test3.txt` est exécutée. Cette commande affecte immédiatement les trois fichiers. La commande `rm a/test2.txt` efface la référence du fichier sous le nom `a/test2.txt`, mais les deux autres liens restent accessibles. Le fichier ne sera réellement effacé qu'après que le dernier lien vers l'*inode* correspondant aie été supprimé. La commande rm(1) utilise en pratique l'appel système unlink(2) qui en toute généralité décrémente le compteur de liens de l'inode correspondant au fichier et l'efface lorsque ce compteur atteint la valeur `0`.

Une description détaillée du fonctionnement de ces appels systèmes et fonctions de la librairie standard peut se trouver dans les livres de référence sur la programmation en C sous Unix [Kerrisk2010], [Mitchell+2001], [StevensRago2008].

Note : Liens symboliques

Les liens créés dans l'exemple précédent avec la commande ln(1) sont des liens dits *durs*, permettant à un même fichier d'avoir plusieurs identités (noms dans la hiérarchie du système de fichier).

Il est quelquefois souhaitable de ne conserver qu'un seul lien dur entre un nom de fichier f et les données, et d'utiliser des liens symboliques vers ce nom de fichier f plutôt qu'un lien direct vers ces données. Contrairement à un lien dur, le lien symbolique n'a pas besoin d'être nécessairement au sein du même système de fichiers. Un lien

1. Dans un système de fichiers Unix, un lien ne peut être créé avec ln(1) ou link(2) que lorsque les deux répertoires concernés sont situés sur le même système de fichiers. Si ce n'est pas le cas, il faut utiliser un *lien symbolique*. Ceux-ci peuvent être créés en utilisant l'appel système symlink(2) ou via la commande ln(1) avec l'argument `-s`.

symbolique contient le nom du fichier lié, et permet donc une redirection pour accéder à ce dernier. Le désavantage d'un lien symbolique est que si le fichier lié est renommé ou effacé alors les données ne sont plus accessibles.

On peut créer un lien symbolique en utilisant l'option `-s` de la commande ln(1) comme le montre l'exemple suivant.

```
$ ls -l
-rw-rw-r-- 1 utilisateur groupe 28 Dec  2 16:00 test1.txt
$ cat test1.txt
Contenu du fichier de test.
$ ln test1.txt lien_dur.txt
$ ls -l
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 lien_dur.txt
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 test1.txt
$ ln -s test1.txt lien_symb.txt
$ ls -l
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 lien_dur.txt
lrwxrwxrwx 1 utilisateur groupe  9 Dec  2 16:01 lien_symb.txt -> test1.txt
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 test1.txt
$ mv test1.txt test_1.txt
$ ls -l
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 lien_dur.txt
lrwxrwxrwx 1 utilisateur groupe  9 Dec  2 16:01 lien_symb.txt -> test1.txt
-rw-rw-r-- 2 utilisateur groupe 28 Dec  2 16:00 test_1.txt
$ cat lien_symb.txt
cat: lien_symb.txt: No such file or directory
$ cat lien_dur.txt
Contenu du fichier de test.
```

9.2.1 Utilisation des fichiers

Si quelques processus manipulent le système de fichiers et parcourent les répertoires, les processus qui utilisent des données sauvegardées dans des fichiers sont encore plus nombreux. En plus du mapping des fichiers en mémoire que nous avons abordé précédemment dans le syllabus, un système Unix offre deux possibilités pour écrire et lire des données dans un fichier. La première utilise directement les appels systèmes open(2), read(2)/ write(2) et close(2). La seconde s'appuie sur les fonctions fopen(3), fread(3)/ fwrite(3) et fclose(3) de la librairie stdio(3). Seuls les appels systèmes seront traités dans ce cours. Des détails complémentaires sur les fonctions de la libraire peuvent être obtenus dans [Kerrisk2010], [Mitchell+2001] ou [StevensRago2008].

Du point de vue des appels systèmes de manipulation des fichiers, un fichier est une séquence d'octets. Avant qu'un processus ne puisse écrire ou lire dans un fichier, il doit d'abord demander au système d'exploitation l'autorisation d'accéder au fichier. Cela se fait en utilisant l'appel système open(2).

```
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>

int open(const char *pathname, int flags);
int open(const char* pathname, int flags, mode_t mode);
```

Il existe deux variantes de l'appel système open(2). La première permet d'ouvrir des fichiers existants. Elle prend deux arguments. La deuxième permet de créer un nouveau fichier et l'ouvre ensuite. Elle prend trois arguments. Le premier argument est le nom absolu ou relatif du fichier dont l'ouverture est demandée. Le deuxième argument est un entier qui contient un ensemble de drapeaux binaires qui précisent la façon dont le fichier doit être ouvert. Ces drapeaux sont divisés en deux groupes. Le premier groupe est relatif à l'accès en lecture et/ou en écriture du fichier. Lors de l'ouverture d'un fichier avec open(2), il est nécessaire de spécifier l'un des trois drapeaux d'accès suivants :

– `O_RDONLY` : indique que le fichier est ouvert uniquement en lecture. Aucune opération d'écriture ne sera effectuée sur le fichier.

– `O_WRONLY` : indique que le fichier est ouvert uniquement en écriture. Aucune opération de lecture ne sera effectuée sur le fichier.
– `O_RDWR` : indique que le fichier est ouvert pour des opérations de lecture et d'écriture.

En plus de l'un des trois drapeaux ci-dessus, il est également possible de spécifier un ou plusieurs drapeaux optionnels. Ces drapeaux sont décrits en détails dans la page de manuel open(2). Les plus utiles sont probablement :

– `O_CREAT` : indique que si le fichier n'existe pas, il doit être créé lors de l'exécution de l'appel système open(2). L'appel système creat(2) peut également être utilisé pour créer un nouveau fichier. Lorsque le drapeau `O_CREAT` est spécifié, l'appel système open(2) prend comme troisième argument les permissions du fichier qui doit être créé. Celles-ci sont spécifiées de la même façon que pour l'appel système chmod(2). Si elles ne sont pas spécifiées, le fichier est ouvert avec comme permissions les permissions par défaut du processus définies par l'appel système umask(2)
– `O_APPEND` : indique que le fichier est ouvert de façon à ce que les données écrites dans le fichier par l'appel système write(2) s'ajoutent à la fin du fichier.
– `O_TRUNC` : indique que si le fichier existe déjà et qu'il est ouvert en écriture, alors le contenu du fichier doit être supprimé avant que le processus ne commence à y accéder.
– `O_SYNC` : ce drapeau indique que toutes les opérations d'écriture sur le fichier doivent être effectuées immédiatement sur le dispositif de stockage sans être mises en attente dans les buffers du noyau du système d'exploitation
– `O_CLOEXEC` : ce drapeau qui est spécifique à Linux indique que le fichier doit être automatiquement fermé lors de l'exécution de l'appel système execve(2). Normalement, les fichiers qui ont été ouverts par open(2) restent ouverts lors de l'exécution de execve(2).

Ces différents drapeaux binaires doivent être combinés en utilisant une disjonction logique entre les différents drapeaux. Ainsi, `O_CREAT|O_RDWR` correspond à l'ouverture d'un fichier qui doit à la fois être créé si il n'existe pas et ouvert en lecture et écriture.

Lors de l'exécution de open(2), le noyau du système d'exploitation vérifie si le processus qui exécute l'appel système dispose des permissions suffisantes pour accéder au fichier. Si oui, le système d'exploitation ouvre le fichier et retourne au processus appelant le *descripteur de fichier* correspondant. Si non, le processus récupère une valeur de retour négative et `errno` indique le type d'erreur.

Sous Unix, un *descripteur de fichier* est représenté sous la forme d'un entier positif. L'appel système open(2) retourne toujours le plus petit *descripteur de fichier* disponible. Par convention,

– `0` est le *descripteur de fichier* correspondant à l'entrée standard.
– `1` est le *descripteur de fichier* correspondant à la sortie standard.
– `2` est le *descripteur de fichier* correspondant à la sortie d'erreur standard.

Si l'appel système open(2) échoue, il retourne `-1` comme *descripteur de fichier* et `errno` donne plus de précisions sur le type d'erreur. Il peut s'agir d'une erreur liée aux droits d'accès au fichier (`EACESS`), une erreur de drapeau (`EINVAL`) ou d'une erreur d'entrée sortie lors de l'accès au dispositif de stockage (`EIO`). Le noyau du système d'exploitation maintient une table de l'ensemble des fichiers qui sont ouverts par tous les processus actifs. Si cette table est remplie, il n'est plus possible d'ouvrir de nouveau fichier et open(2) retourne une erreur. Il en va de même si le processus tente d'ouvrir plus de fichiers que le nombre maximum de fichiers ouverts qui est autorisé.

Note : Seul open(2) vérifie les permissions d'accès aux fichiers

Sous Unix, seul l'appel système open(2) vérifie qu'un processus dispose des permissions suffisantes pour accéder à un fichier qui est ouvert. Si les permissions ou le propriétaire d'un fichier change alors que ce fichier est ouvert par un processus, ce processus continue à pouvoir y accéder sans être affecté par la modification de droits. Il en va de même lorsqu'un fichier est effacé avec l'appel système unlink(2). Si un processus utilisait le fichier qui est effacé, il continue à pouvoir l'utiliser même si le fichier n'apparaît plus dans le répertoire.

Toutes les opérations qui sont faites sur un fichier se font en utilisant le *descripteur de fichier* comme référence au fichier. Un *descripteur de fichier* est une ressource limitée dans un système d'exploitation tel que Unix et il est important qu'un processus n'ouvre pas inutilement un grand nombre de fichiers [2] et ferme correctement les fichiers ouverts lorsqu'il ne doit plus y accéder. Cela se fait en utilisant l'appel système close(2). Celui-ci prend comme argument le *descripteur de fichier* qui doit être fermé.

2. Il y a une limite maximale au nombre de fichiers qui peuvent être ouverts par un processus. Cette limite peut être récupérée avec l'appel système getdtablesize(2).

```
#include <unistd.h>

int close(int fd);
```

Tout processus doit correctement fermer tous les fichiers qu'il a utilisé. Par défaut, le système d'exploitation ferme automatiquement les descripteurs de fichiers correspondant `0`, `1` et `2` lorsqu'un processus se termine. Les autres descripteurs de fichiers doivent être explicitement fermés par le processus. Si nécessaire, cela peut se faire en enregistrant une fonction permettant de fermer correctement les fichiers ouverts via atexit(3). Il faut noter que par défaut un appel à execve(2) ne ferme pas les descripteurs de fichiers ouverts par le processus. C'est nécessaire pour permettre au programme exécuté d'avoir les entrées et sorties standard voulues.

Lorsqu'un fichier a été ouvert, le noyau du système d'exploitation maintient un *offset pointer*. Cet *offset pointer* est la position actuelle de la tête de lecture/écriture du fichier. Lorsqu'un fichier est ouvert, son *offset pointer* est positionné au premier octet du fichier, sauf si le drapeau `O_APPEND` a été spécifié lors de l'ouverture du fichier, dans ce cas l'*offset pointer* est positionné juste après le dernier octet du fichier de façon à ce qu'une écriture s'ajoute à la suite du fichier.

Les deux appels systèmes permettant de lire et d'écrire dans un fichier sont respectivement read(2) et write(2).

```
#include <unistd.h>

ssize_t read(int fd, void *buf, size_t count);
ssize_t write(int fd, const void *buf, size_t count);
```

Ces deux appels systèmes prennent trois arguments. Le premier est le *descripteur du fichier* sur lequel l'opération doit être effectuée. Le second est un pointeur `void *` vers la zone mémoire à lire ou écrire et le dernier est la quantité de données à lire/écrire. Si l'appel système réussit, il retourne le nombre d'octets qui ont été écrits/lus et sinon une valeur négative et la variable `errno` donne plus de précisions sur le type d'erreur. read(2) retourne `0` lorsque la fin du fichier a été atteinte.

Il est important de noter que read(2) et write(2) permettent de lire et d'écrire des séquences contiguës d'octets. Lorsque l'on écrit ou lit des chaînes de caractères dans lesquels chaque caractère est représenté sous la forme d'un byte, il est possible d'utiliser read(2) et write(2) pour lire et écrire d'autres types de données que des octets comme le montre l'exemple ci-dessous.

```
#include <stdio.h>
#include <stdlib.h>
#include <errno.h>
#include <sys/types.h>
#include <sys/stat.h>
#include <fcntl.h>
#include <unistd.h>
#include <string.h>

void exit_on_error(char *s) {
  perror(s);
  exit(EXIT_FAILURE);
}

int main (int argc, char *argv[])  {
  int n=1252;
  int n2;
  short ns=1252;
  short ns2;
  long nl=125212521252;
  long nl2;
  float f=12.52;
  float f2;
  char *s="sinf1252";
  char *s2=(char *) malloc(strlen(s)*sizeof(char )+1);
  int err;
  int fd;
```

```
  fd=open("test.dat",O_WRONLY|O_CREAT|O_TRUNC, S_IRUSR|S_IWUSR);
  if(fd==-1) {
    perror("open");
    exit(EXIT_FAILURE);
  }
  if( write(fd, (void *) s, strlen(s)) == -1 )
    exit_on_error("write s");
  if (write(fd, (void *) &n, sizeof(int )) == -1)
    exit_on_error("write n");
  if (write(fd, (void *) &ns, sizeof(short int))==-1)
    exit_on_error("write ns");
  if (write(fd, (void *) &nl, sizeof(long int))==-1)
    exit_on_error("write nl");
  if (write(fd, (void *) &f, sizeof(float))==-1)
    exit_on_error("write f");
  if (close(fd)==-1)
    exit_on_error("close ");

  // lecture
  fd=open("test.dat",O_RDONLY);
  if(fd==-1) {
    perror("open");
    exit(EXIT_FAILURE);
  }
  printf("Fichier ouvert\n");

  if(read(fd, (void *) s2, strlen(s))==-1)
    exit_on_error("read s");
  printf("Donnée écrite : %s, lue: %s\n",s,s2);

  if(read(fd, (void *) &n2, sizeof(int))==-1)
    exit_on_error("read n");
  printf("Donnée écrite : %d, lue: %d\n",n,n2);

  if(read(fd, (void *) &ns2, sizeof(short))==-1)
    exit_on_error("read ns");
  printf("Donnée écrite : %d, lue: %d\n",ns,ns2);

  if(read(fd, (void *) &nl2, sizeof(long))==-1)
    exit_on_error("read nl");
  printf("Donnée écrite : %ld, lue: %ld\n",nl,nl2);

  if(read(fd, (void *) &f2, sizeof(float))==-1)
    exit_on_error("read f");
  printf("Donnée écrite : %f, lue: %f\n",f,f2);
  err=close(fd);
  if(err==-1){
   perror("close");
    exit(EXIT_FAILURE);
  }

  return(EXIT_SUCCESS);
}
```

Lors de son exécution, ce programme affiche la sortie ci-dessous.

```
Fichier ouvert
Donnée écrite : sinf1252, lue: sinf1252
Donnée écrite : 1252, lue: 1252
Donnée écrite : 1252, lue: 1252
Donnée écrite : 125212521252, lue: 125212521252
Donnée écrite : 12.520000, lue: 12.520000
```

Si il est bien possible de sauvegarder dans un fichier des entiers, des nombres en virgule flottante voire même des structures, il faut être bien conscient que l'appel système write(2) se contente de sauvegarder sur le disque le contenu de la zone mémoire pointée par le pointeur qu'il a reçu comme second argument. Si comme dans l'exemple précédent c'est le même processus qui lit les données qu'il a écrit, il pourra toujours récupérer les données correctement.

Par contre, lorsqu'un fichier est écrit sur un ordinateur, envoyé via Internet et lu sur un autre ordinateur, il peut se produire plusieurs problèmes dont il faut être conscient. Le premier problème est que deux ordinateurs différents n'utilisent pas nécessairement le même nombre d'octets pour représenter chaque type de données. Ainsi, sur un ordinateur équipé d'un ancien processeur [IA32], les entiers sont représentés sur 32 bits (i.e. 4 bytes) alors que sur les processeurs plus récents ils sont souvent représentés sur 64 bits (i.e. 8 bytes). Cela implique qu'un tableau de 100 entiers en 32 bits sera interprété comme un tableau de 50 entiers en 64 bits.

Le second problème est que les fabricants de processeurs ne se sont pas mis d'accord sur la façon dont il fallait représenter les entiers sur 16 et 32 bits en mémoire. Il y a deux techniques qui sont utilisées : *big endian* et *little endian*.

Pour comprendre ces deux techniques, regardons comment l'entier 16 bits `0b1111111100000000` est stocké en mémoire. En *big endian*, le byte `11111111` sera stocké à l'adresse *x* et le byte `00000000` à l'adresse *x+1*. En *little endian*, c'est le byte `00000000` qui est stocké à l'adresse *x* et le byte `11111111` qui est stocké à l'adresse *x+1*. Il en va de même pour les entiers encodés sur 32 bits comme illustré dans les deux figures ci-dessous [3].

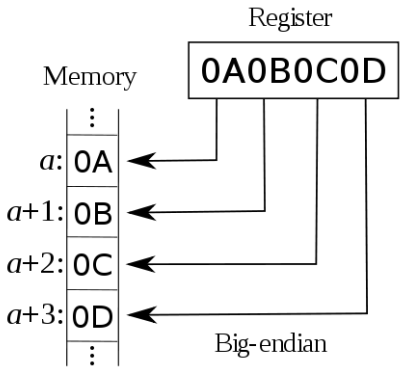


FIGURE 9.1 – Ecriture d'un entier 32 bits en mémoire en *big endian*

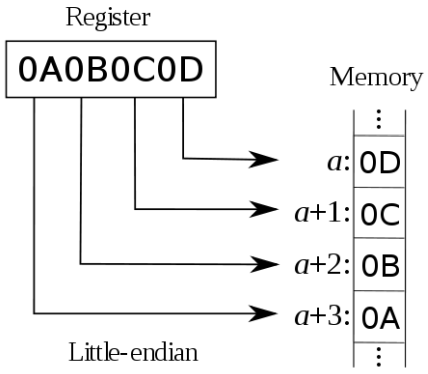


FIGURE 9.2 – Ecriture d'un entier 32 bits en mémoire en *little endian*

Pour les nombres en virgule flottante, ce problème ne se pose heureusement pas car tous les processeurs actuels utilisent la même norme pour représenter les nombres en virgule flottant en mémoire.

3. Source : http ://en.wikipedia.org/wiki/Endianness

Les processeurs [IA32] utilisent la représentation *little endian* tandis que les PowerPC utilisent *big endian*. Certains processeurs sont capables d'utiliser les deux représentations.

Il est également possible en utilisant l'appel système lseek(2) de déplacer l'*offset pointer* associé à un *descripteur de fichier*.

```
#include <sys/types.h>
#include <unistd.h>

off_t lseek(int fd, off_t offset, int whence);
```

Cet appel système prend trois arguments. Le premier est le *descripteur de fichier* dont l'*offset pointer* doit être modifié. Le second est un entier qui est utilisé pour le calcul de la nouvelle position *offset pointer* et le troisième indique comment l'*offset pointer* doit être calculé. Il y a trois modes de calcul possibles pour l'*offset pointer* :
– `whence==SEEK_SET` : dans ce cas, le deuxième argument de l'appel système indique la valeur exacte du nouvel *offset pointer*
– `whence==SEEK_CUR` : dans ce cas, le nouvel *offset pointer* sera sa position actuelle à laquelle le deuxième argument aura été ajouté
– `whence==SEEK_END` : dans ce cas, le nouvel *offset pointer* sera la fin du fichier à laquelle le deuxième argument aura été ajouté

Note : Fichiers temporaires

Il est parfois nécessaire dans un programme de créer des fichiers temporaires qui sont utilisés pour effectuer des opérations dans le processus sans pour autant être visible dans d'autres processus et sur le système de fichiers. Il est possible d'utiliser open(2) pour créer un tel fichier temporaire, mais il faut dans ce cas prévoir tous les cas d'erreur qui peuvent se produire lorsque par exemple plusieurs instances du même programme s'exécutent au même moment. Une meilleure solution est d'utiliser la fonction de la librairie mkstemp(3). Cette fonction prend comme argument un modèle de nom de fichier qui se termine par `XXXXXX` et génère un nom de fichier unique et retourne un descripteur de fichier associé à ce fichier. Elle s'utilise généralement comme suit :

```
char template[]="/tmp/sinf1252PROCXXXXXX";

int fd=mkstemp(template);
if(fd==-1)
  exit_on_error("mkstemp");
// template contient le nom exact du fichier généré
unlink(template);
// le fichier est effacé, mais reste accessible
// via son descripteur jusqu'à close(fd)

// Accès au fichier avec read et write

if(close(fd)==-1)
  exit_on_error("close");
// le fichier n'est plus accessible
```

L'utilisation de unlink(2) permet de supprimer le fichier du système de fichiers dès qu'il a été créé. Ce fichier reste cependant accessible au processus tant que celui-ci dispose d'un descripteur de fichier qui y est associé.

Note : Duplication de descripteurs de fichiers

Dans certains cas il est utile de pouvoir dupliquer un descripteur de fichier. C'est possible avec les appels systèmes dup(2) et dup2(2). L'appel système dup(2) prend comme argument un descripteur de fichier et retourne le plus petit descripteur de fichier libre. Lorsqu'un descripteur de fichier a été dupliqué avec dup(2) les deux descripteurs de fichiers partagent le même *offset pointer* et les mêmes modes d'accès au fichier.

9.2.2 Fichiers mappés en mémoire

L'utilisation de fichiers mappés en mémoire avec l'appel système mmap(2) est une alternative à l'utilisation des appels open(2), read(2), write(2) et close(2). Ce sujet est détaillé dans le chapitre précédent de ce syllabus, dédié à l'utilisation de la mémoire virtuelle.

Sur un système Unix, il est fréquent que des processus différents doivent communiquer entre eux et coordonner leurs activités. Dans ce chapitre, nous détaillons plusieurs mécanismes permettant à des processus ayant un ancêtre commun ou non de coordonner leurs activités.

Historiquement, le premier mécanisme de coordination entre processus a été l'utilisation des signaux. Nous avons déjà utilisé des signaux sans les avoir détaillés explicitement pour terminer des processus avec la commande kill(1). Nous décrivons plus en détails les principales utilisations des signaux dans ce chapitre.

Lorsque deux processus ont un ancêtre commun, il est possible de les relier en utilisant un pipe(2). Si ils n'ont pas d'ancêtre commun, il est possible d'utiliser une *fifo* pour obtenir un effet similaire. Une *fifo* peut être créée en utilisant la commande mkfifo(1) ou la fonction mkfifo(3). mkfifo(3) s'utilise comme l'appel système open(2). Pour créer une nouvelle *fifo*, il suffit de lui donner un nom sous la forme d'une chaîne de caractères et tout processus qui connait le nom de la *fifo* pourra l'utiliser.

Les sémaphores que nous avons utilisés pour coordonner plusieurs threads sont également utilisables entre processus moyennant quelques différences que nous détaillerons.

Enfin, des processus liés ou non doivent parfois accéder aux mêmes fichiers. Ces accès concurrents, si ils ne sont pas correctement coordonnés, peuvent conduire à des corruptions de fichiers. Nous présenterons les mécanismes de locking (verrouillage) qui sont utilisés dans Unix et Linux pour coordonner l'accès à des fichiers.

9.3 Signaux

L'envoi et la réception de signaux est le mécanisme de communication entre processus le plus primitif sous Unix. Un *signal* est une forme d'interruption logicielle [StevensRago2008]. Comme nous l'avons vu précédemment, un microprocesseur utilise les interruptions pour permettre au système d'exploitation de réagir aux événements extérieurs qui surviennent, et aux demandes d'action de la part du processeur ou d'un programme. Un *signal* Unix est un mécanisme qui permet à un processus de réagir de façon *asynchrone* à un événement qui s'est produit. Certains de ces événements sont directement liés au fonctionnement du matériel. D'autres sont provoqués par le processus lui-même ou un autre processus s'exécutant sur le système.

Pour être capable d'utiliser les signaux à bon escient, il est important de bien comprendre comment ceux-ci sont implémentés dans le système d'exploitation.

Il existe deux types de signaux.

– Un *signal synchrone* est un *signal* qui a été directement causé par l'exécution d'une instruction du processus. Un exemple typique de *signal synchrone* est le signal `SIGFPE` qui est généré par le système d'exploitation lorsqu'un processus provoque une exception lors du calcul d'expressions mathématiques. C'est le cas notamment lors d'une division par zéro. La sortie ci-dessous illustre ce qu'il se produit lors de l'exécution du programme `/Fichiers/src/sigfpe.c`.

```
$ ./sigfpe
Calcul de : 1252/0
Floating point exception
```

– Un *signal asynchrone* est un *signal* qui n'a pas été directement causé par l'exécution d'une instruction du processus. Il peut être produit par le système d'exploitation ou généré par un autre processus comme lorsque nous avons utilisé kill(1) dans un shell pour terminer un processus.

Le système Unix fournit plusieurs appels systèmes qui permettent de manipuler les signaux. Un processus peut recevoir des signaux synchrones ou asynchrones. Pour chaque signal, le système d'exploitation définit un traitement par défaut. Pour certains signaux, leur réception provoque l'arrêt du processus. Pour d'autres, le signal est ignoré et le processus peut continuer son exécution. Un processus peut redéfinir le traitement des signaux qu'il reçoit en utilisant l'appel système signal(2). Celui-ci permet d'associer un *handler* ou fonction de traitement de signal à chaque signal. Lorsqu'un *handler* a été associé à un signal, le système d'exploitation exécute ce *handler*

dès que ce signal survient. Unix permet également à un processus d'envoyer un signal à un autre processus en utilisant l'appel système kill(2).

Avant d'analyser en détails le fonctionnement des appels systèmes signal(2) et kill(2), il est utile de parcourir les principaux signaux. Chaque *signal* est identifié par un entier positif et signal.h définit des constantes pour chaque signal reconnu par le système. Sous Linux, les principaux signaux sont :
- `SIGALRM`. Ce signal survient lorsqu'une alarme définie par la fonction alarm(3posix) ou l'appel système setitimer(2) a expiré. Par défaut, la réception de ce signal provoque la terminaison du processus.
- `SIGBUS`. Ce signal correspond à une erreur au niveau matériel. Par défaut, la réception de ce signal provoque la terminaison du processus.
- `SIGSEGV`. Ce signal correspond à une erreur dans l'accès à la mémoire, typiquement une tentative d'accès en dehors de la zone mémoire allouée au processus. Par défaut, la réception de ce signal provoque la terminaison du processus.
- `SIGFPE`. Ce signal correspond à une erreur au niveau de l'utilisation des fonctions mathématiques, notamment en virgule flottante mais pas seulement. Par défaut, la réception de ce signal provoque la terminaison du processus.
- `SIGTERM`. Ce signal est le signal utilisé par défaut par la commande kill(1) pour demander la fin d'un processus. Par défaut, la réception de ce signal provoque la terminaison du processus.
- `SIGKILL`. Ce signal permet de forcer la fin d'un processus. Alors qu'un processus peut définir un handler pour le signal `SIGTERM`, il n'est pas possible d'en définir un pour `SIGKILL`. Ce signal est le seul qui ne peut être traité et ignoré par un processus.
- `SIGUSR1` et `SIGUSR2` sont deux signaux qui peuvent être utilisés par des processus sans conditions particulières. Par défaut, la réception d'un tel signal provoque la terminaison du processus.
- `SIGCHLD`. Ce signal indique qu'un processus fils s'est arrêté ou a fini son exécution. Par défaut ce signal est ignoré.
- `SIGHUP`. Aux débuts de Unix, ce signal servait à indiquer que la connexion avec le terminal avait été rompue. Aujourd'hui, il est parfois utilisé par des processus serveurs qui rechargent leur fichier de configuration lorsqu'ils reçoivent ce signal.
- `SIGINT`. Ce signal est envoyé par le shell lorsque l'utilisateur tape *Ctrl-C* pendant l'exécution d'un programme. Il provoque normalement la terminaison du processus.

Une description détaillée des différents signaux sous Unix et Linux peut se trouver dans signal(7), [BryantOHallaron2011], [StevensRago2008] ou encore [Kerrisk2010].

Note : Comment arrêter proprement un processus ?

Trois signaux permettent d'arrêter un processus : `SIGTERM`, `SIGINT` et `SIGKILL`. Lorsque l'on doit forcer l'arrêt d'un processus, il est préférable d'utiliser le signal `SIGTERM` car le processus peut prévoir une routine de traitement de ce signal. Cette routine peut par exemple fermer proprement toutes les ressources (fichiers, ...) utilisées par le processus. Lorsque l'on tape *Ctrl-C* dans le shell, c'est le signal `SIGINT` qui est envoyé au processus qui est en train de s'exécuter. Le signal `SIGKILL` ne peut pas être traité par le processus. Lorsque l'on envoie ce signal à un processus, on est donc certain que celui-ci s'arrêtera. Malheureusement, cet arrêt sera brutal et le processus n'aura pas eu le temps de libérer proprement toutes les ressources qu'il utilisait. Quand un processus ne répond plus, il est donc préférable de d'abord essayer `SIGINT` ou `SIGTERM` et de n'utiliser `SIGKILL` qu'en dernier recours.

9.3.1 Envoi de signaux

Un processus peut envoyer un signal à un autre processus en utilisant l'appel système kill(2).

```
#include <sys/types.h>
#include <signal.h>

int kill(pid_t pid, int sig);
```

Cet appel système prend deux arguments. Le second est toujours le numéro du signal à délivrer ou la constante définissant ce signal dans signal.h. Le premier argument indique à quel(s) processus le signal doit être délivré :
- `pid>0`. Dans ce cas, le signal est délivré au processus ayant comme identifiant `pid`.

– `pid==0`. Dans ce cas, le signal est délivré à tous les processus qui font partie du même groupe de processus [4] que le processus qui exécute l'appel système kill(2).
– `pid==-1`. Dans ce cas, le signal est délivré à tous les processus pour lesquels le processus qui exécute kill(2) a les permissions suffisantes pour leur envoyer un signal.
– `pid<-1`. Dans ce cas, le signal est délivré à tous les processus qui font partie du groupe `abs(pid)`.
Par défaut, un processus ne peut envoyer un signal qu'à un processus qui s'exécute avec les mêmes permissions que le processus qui exécute l'appel système kill(2).

9.3.2 Traitement de signaux

Pour des raisons historiques, il existe deux appels systèmes permettant à un processus de spécifier comment un signal particulier doit être traité. Le premier est l'appel système signal(2). Il prend deux arguments : le numéro du signal dont le traitement doit être modifié et la fonction à exécuter lorsque ce signal est reçu. Le second est sigaction(2). Cet appel système est nettement plus générique et plus complet que signal(2) mais il ne sera pas traité en détails dans ces notes. Des informations complémentaires sur l'utilisation de sigaction(2) peuvent être obtenues dans [Kerrisk2010] ou [StevensRago2008].

```
#include <signal.h>
typedef void (*sighandler_t)(int);

sighandler_t signal(int signum, sighandler_t handler);
```

Le premier argument de l'appel système signal(2) est généralement spécifié en utilisant les constantes définies dans signal.h. Le second argument est un pointeur vers une fonction de type `void` qui prend comme argument un entier. Cette fonction est la fonction qui sera exécutée par le système d'exploitation à la réception du signal. Ce second argument peut également être la constante `SIG_DFL` si le signal doit être traité avec le traitement par défaut documenté dans signal(7) et `SIG_IGN` si le signal doit être ignoré. La valeur de retour de l'appel système signal(2) est la façon dont le système d'exploitation gérait le signal passé en argument avant l'exécution de l'appel système (typiquement `SIG_DFL` ou `SIG_IGN`) ou `SIG_ERR` en cas d'erreur.

L'exemple ci-dessous est un programme simple qui compte le nombre de signaux `SIGUSR1` et `SIGUSR2` qu'il reçoit et se termine dès qu'il a reçu cinq signaux.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <signal.h>
#include <unistd.h>
#include <stdint.h>
#include <stdio.h>

volatile sig_atomic_t n_sigusr1=0;
volatile sig_atomic_t n_sigusr2=0;

static void sig_handler(int);

int main (int argc, char *argv[])  {

  if(signal(SIGUSR1,sig_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
  if(signal(SIGUSR2,sig_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
```

4. Chaque processus appartient à groupe de processus. Ce groupe de processus peut être récupéré via l'appel système getpgrp(2). Par défaut, lorsqu'un processus est créé, il appartient au même groupe de processus que son processus père, mais il est possible de changer de groupe de processus en utilisant l'appel système setpgid(2). En pratique, les groupes de processus sont surtout utilisés par le shell. Lorsqu'un utilisateur exécute une commande combinée telle que `cat /tmp/t | ./a.out`, il souhaite pouvoir l'arrêter en tapant sur *Ctrl-C* si nécessaire. Pour cela, il faut pouvoir délivrer le signal `SIGINT` aux processus `cat` et `a.out`.

```
  while( (n_sigusr1+n_sigusr2) <5) {
    // vide
  }

  printf("Fin du processus\n");
  printf("Reçu %d SIGUSR1 et %d SIGUSR2\n",n_sigusr1,n_sigusr2);
  return(EXIT_SUCCESS);
}

static void sig_handler(int signum) {

  if(signum==SIGUSR1) {
    n_sigusr1++;
  }
  else {
    if(signum==SIGUSR2) {
      n_sigusr2++;
    }
    else {
      char *msg="Reçu signal inattendu\n";
      write(STDERR_FILENO,msg,strlen(msg));
      _exit(EXIT_FAILURE);
    }
  }
}
```

Lors de son exécution, ce programme affiche :

```
$ ./sigusr &
[1] 45602
$ kill -s SIGUSR1 45602
$ kill -s SIGUSR2 45602
$ kill -s SIGUSR2 45602
$ kill -s SIGUSR1 45602
$ kill -s SIGUSR1 45602
$ Fin du processus
Reçu 3 SIGUSR1 et 2 SIGUSR2
[1]+  Done                    ./sigusr
```

Il est intéressant d'analyser le code source du programme ci-dessus. Commençons d'abord par une lecture rapide pour comprendre la logique du programme sans s'attarder sur les détails. La fonction `main` utilise l'appel système signal(2) pour enregistrer un handler pour les signaux `SIGUSR1` et `SIGUSR2`. La fonction `sig_handler` sera exécutée dès réception d'un de ces signaux. Cette fonction prend comme argument le numéro du signal reçu. Cela permet de traiter plusieurs signaux dans la même fonction. Ensuite, la boucle `while` est une boucle active qui ne se terminera que lorsque la somme des variables `n_sigusr1` et `n_sigusr2` sera égale à 5. Ces deux variables sont modifiées uniquement dans la fonction `sig_handler`. Elles permettent de compter le nombre de signaux de chaque type qui ont été reçus.

Une lecture plus attentive du code ci-dessus révèle plusieurs points importants auxquels il faut être attentif lorsque l'on utilise les signaux.

Tout d'abord, lorsqu'un processus comprend une (ou plusieurs) fonction(s) de traitement de signaux, il y a plusieurs séquences d'instructions qui peuvent être exécutées par le processus. La première est la suite d'instructions du processus lui-même qui démarre à la fonction `main`. Dès qu'un signal est reçu, cette séquence d'instructions est interrompue pour exécuter la séquence d'instructions de la fonction de traitement du signal. Ce n'est que lorsque cette fonction se termine que la séquence principale peut reprendre son exécution à l'endroit où elle a été interrompue.

L'existence de deux ou plusieurs séquences d'instructions peut avoir des conséquences importantes sur le bon fonctionnement du programme et peut poser de nombreuses difficultés d'implémentation. En effet, une fonction de traitement de signal doit pouvoir être exécutée à n'importe quel moment. Elle peut donc démarrer à n'importe quel endroit de la séquence d'instructions du processus. Si le processus et une fonction de traitement de signal

accèdent à la même variable, il y a un risque que celle-ci soit modifiée par la fonction de traitement du signal pendant qu'elle est utilisée dans le processus. Si l'on n'y prend garde, ces accès venant de différentes séquences d'instructions peuvent poser des problèmes similaires à ceux posés par l'utilisation de threads. Une routine de traitement de signal est cependant moins générale qu'un thread et les techniques utilisables dans les threads ne sont pas applicables aux fonctions de traitement des signaux. En effet, quand la fonction de traitement de signal démarre, il est impossible de bloquer son exécution sur un mutex pour revenir au processus principal. Celle-ci doit s'exécuter jusqu'à sa dernière instruction.

Lorsque l'on écrit une routine de traitement de signal, plusieurs précautions importantes doivent être prises. Tout d'abord, une fonction de traitement de signal doit manipuler les variables avec précautions. Comme elle est potentiellement exécutée depuis n'importe quel endroit du code, elle ne peut pas s'appuyer sur le stack. Elle ne peut utiliser que des variables globales pour influencer le processus principal. Comme ces variables peuvent être utilisées à la fois dans le processus et la routine de traitement de signal, il est important de les déclarer en utilisant le mot-clé `volatile`. Cela force le compilateur à recharger la valeur de la variable de la mémoire à chaque fois que celle-ci est utilisée. Mais cela ne suffit pas car il est possible que le processus exécute l'instruction de chargement de la valeur de la variable, puis qu'un signal lui soit délivré, ce qui provoquera l'exécution de la fonction de traitement du signal. Lorsque celle-ci se terminera le processus poursuivra son exécution sans recharger la valeur de la variable potentiellement modifiée par la fonction de traitement du signal. Face à ce problème, il est préférable d'utiliser uniquement des variables de types `sig_atomic_t` dans les fonctions de traitement de signaux. Ce type permet de stocker un entier. Lorsque ce type est utilisé, le compilateur garantit que tous les accès à la variable se feront de façon atomique sans accès concurrent possible entre le processus et la fonction de traitement des signaux.

L'utilisation de `sig_atomic_t` n'est pas la seule précaution à prendre lorsque l'on écrit une fonction de traitement des signaux. Il faut également faire attention aux fonctions de la librairie et aux appels systèmes que l'on utilise. Sachant qu'un signal peut être reçu à n'importe quel moment, il est possible qu'un processus reçoive un signal et exécute une fonction de traitement du signal pendant l'exécution de la fonction `fct` de la librairie standard. Si la fonction de traitement du signal utilise également la fonction `fct`, il y a un risque d'interférence entre l'exécution de ces deux fonctions. Ce sera le cas notamment si la fonction utilise un buffer statique ou modifie la variable `errno`. Dans ces cas, la fonction de traitement du signal risque de modifier une valeur ou une zone mémoire qui a déjà été modifiée par le processus principal et cela donnera un résultat incohérent. Pour éviter ces problèmes, il ne faut utiliser que des fonctions *réentrantes* à l'intérieur des fonctions de traitement des signaux. Des fonctions comme printf(3), scanf(3) ne sont pas réentrantes et ne doivent pas être utilisées dans une section de traitement des signaux. La (courte) liste des fonctions qui peuvent être utilisées sans risque est disponible dans la section 2.4 de l'Open Group Base Specification

Ces restrictions sur les instructions qui peuvent être utilisées dans une fonction de traitement des signaux ne sont pas les seules qui affectent l'utilisation des signaux. Ceux-ci souffrent d'autres limitations.

Pour bien les comprendre, il est utile d'analyser comment ceux-ci sont supportés par le noyau. Il y a deux stratégies possibles pour implémenter les signaux sous Unix. La première stratégie est de considérer qu'un signal est un message qui est envoyé depuis le noyau ou un processus à un autre processus. Pour traiter ces messages, le noyau contient une queue qui stocke tous les signaux destinés à un processus donné. Avec cette stratégie d'implémentation, l'appel système kill(2) génère un message et le place dans la queue associée au processus destination. Le noyau stocke pour chaque processus un tableau de pointeurs vers les fonctions de traitement de chacun des signaux. Ce tableau est modifié par l'appel système signal(2). Chaque fois que le noyau réactive un processus, il vérifie si la queue associée à ce processus contient un ou plusieurs messages concernant des signaux. Si un message est présent, le noyau appelle la fonction de traitement du signal correspondant. Lorsque la fonction se termine, l'exécution du processus reprend à l'instruction qui avait été interrompue.

La seconde stratégie est de représenter l'ensemble des signaux qu'un processus peut recevoir sous la forme de drapeaux binaires. En pratique, il y a un drapeau par signal. Avec cette stratégie d'implémentation, l'appel système kill(2) modifie le drapeau correspondant du processus destination du signal (sauf si ce signal est ignoré par le processus, dans ce cas le drapeau n'est pas modifié). L'appel système signal(2) modifie également le tableau contenant les fonctions de traitement des signaux associé au processus. Chaque fois que le noyau réactive un processus, que ce soit après un changement de contexte ou après l'exécution d'un appel système, il vérifie les drapeaux relatifs aux signaux du processus. Si un des drapeaux est vrai, le noyau appelle la fonction de traitement associée à ce signal.

La plupart des variantes de Unix, y compris Linux, utilisent la seconde stratégie d'implémentation pour les signaux. L'avantage principal de l'utilisation de drapeaux pour représenter les signaux reçus par un processus est

qu'il suffit d'un bit par signal qui peut être reçu par le processus. La première stratégie nécessite de maintenir une queue par processus et la taille de cette queue varie en fonction du nombre de signaux reçus. Par contre, l'utilisation de drapeaux a un inconvénient majeur : il n'y a pas de garantie sur la délivrance des signaux. Lorsqu'un processus reçoit un signal, cela signifie qu'il y a *au moins* un signal de ce type qui a été envoyé au processus, mais il est très possible que plus d'un signal ont été envoyés au processus.

Pour illustrer ce problème, considérons le programme ci-dessous qui compte simplement le nombre de signaux `SIGUSR1` reçus.

```
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <unistd.h>

volatile sig_atomic_t n_sigusr1=0;

static void sig_handler(int);

int main (int argc, char *argv[])  {

  if(signal(SIGUSR1,sig_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
  int duree=60;
  while(duree>0) {
    printf("Exécution de sleep(%d)\n",duree);
    duree=sleep(duree);
  }
  printf("Fin du processus\n");
  printf("Reçu %d SIGUSR1\n",n_sigusr1);
  return(EXIT_SUCCESS);
}

static void sig_handler(int signum) {

  if(signum==SIGUSR1) {
    n_sigusr1++;
  }
}
```

Depuis un shell, il est possible d'envoyer plusieurs fois le signal `SIGUSR1` rapidement avec le script `/Fichiers/src/nkill.sh`. Ce script prend deux arguments : le nombre de signaux à envoyer et le processus destination.

```
#!/bin/bash
if [ $# -ne 2 ]
then
    echo "Usage: `basename $0` n pid"
    exit 1
fi
n=$1
pid=$2
for (( c=1; c<=$n; c++ ))
do
    kill -s SIGUSR1 $pid
done
```

La sortie ci-dessous présente une exécution de ce script avec le processus `/Fichiers/src/sigusrcount.c` en tâche de fond.

```
$ ./sigusrcount &
[1] 47845
```

```
$ Exécution de sleep(60)
$./nkill.sh 10 47845
Exécution de sleep(52)
$ ./nkill.sh 10 47845
Exécution de sleep(46)
$ ./nkill.sh 10 47845
Exécution de sleep(31)
$ Fin du processus
Reçu 3 SIGUSR1
```

Il y a plusieurs points intéressants à noter concernant l'exécution de ce programme. Tout d'abord, même si 30 signaux `SIGUSR1` ont été générés, seuls 3 de ces signaux ont effectivement étés reçus. Les signaux ne sont manifestement pas fiables sous Unix et cela peut s'expliquer de deux façons. Premièrement, les signaux sont implémentés sous la forme de bitmaps. La réception d'un signal modifie simplement la valeur d'un bit dans le bitmap du processus. En outre, durant l'exécution de la fonction qui traite le signal `SIGUSR1`, ce signal est bloqué par le système d'exploitation pour éviter qu'un nouveau signal n'arrive pendant que le premier est traité.

Un deuxième point important à relever est l'utilisation de sleep(3). Par défaut, cette fonction de la librairie permet d'attendre un nombre de secondes passé en argument. Si sleep(3) a mis le processus en attente pendant au moins le temps demandé, elle retourne `0` comme valeur de retour. Mais sleep(3) est un des exemples de fonctions ou appels systèmes qui sont dits *lents*. Un *appel système lent* est un appel système dont l'exécution peut être interrompue par la réception d'un signal. C'est notamment le cas pour l'appel système read(2) [5]. Lorsqu'un signal survient pendant l'exécution d'un tel appel système, celui-ci est interrompu pour permettre l'exécution de la fonction de traitement du signal. Après l'exécution de cette fonction, l'appel système se termine [6] en retournant une erreur et met `errno` à `EINTR` de façon à indiquer que l'appel système a été interrompu. Le processus doit bien entendu traiter ces interruptions d'appels systèmes si il utilise des signaux.

Nous terminons cette section en analysant deux cas pratiques d'utilisation des signaux. Le premier est relatif aux signaux synchrones et nous développons une fonction de traitement du signal `SIGFPE` pour éviter qu'un programme ne s'arrête suite à une division par zéro. Ensuite, nous utilisons alarm(3posix) pour implémenter un temporisateur simple.

9.3.3 Traitement de signaux synchrones

Le programme ci-dessous prend en arguments en ligne de commande une séquence d'entiers et divise la valeur `1252` par chaque entier passé en argument. Il enregistre la fonction `sigfpe_handler` comme fonction de traitement du signal `SIGFPE`.

```
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <string.h>
#include <unistd.h>

static void sigfpe_handler(int);

int main (int argc, char *argv[])  {

  int n=1252;
  void (*handler)(int);

  handler=signal(SIGFPE,sigfpe_handler);
  if(handler==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
```

5. Les autres appels systèmes lents sont open(2), write(2), sendto(2), recvfrom(2), sendmsg(2), recvmsg(2), wait(2) ioctl(2).
6. L'appel système sigaction(2) permet notamment de spécifier pour chaque signal si un appel système interrompu par ce signal doit être automatiquement redémarré lorsque le signal survient ou non.

```
  for(int i=1;i<argc;i++) {
    char *endptr;
    printf("Traitement de argv[%d]=%s\n",i,argv[i]);
    fflush(stdout);
    long val=strtol(argv[i],&endptr,10);
    if(*endptr=='\0') {
      int resultat=n/(int) val;
      printf("%d/%d=%d\n",n,(int) val,resultat);
    }
    else {
      printf("Argument incorrect : %s\n",argv[i]);
    }
  }
  return(EXIT_SUCCESS);
}

static void sigfpe_handler(int signum) {
  char *msg="Signal SIGFPE reçu\n";
  write(STDOUT_FILENO,msg,strlen(msg));

}
```

Lors de son exécution, ce programme affiche la sortie ci-dessous :

```
$ ./sigfpe2 1 2 aa 0 9
Traitement de argv[1]=1
1252/1=1252
Traitement de argv[2]=2
1252/2=626
Traitement de argv[3]=aa
Argument incorrect : aa
Traitement de argv[4]=0
Signal SIGFPE reçu
Signal SIGFPE reçu
...
```

La fonction `sigfpe_handler` traite bien le signal `SIGPFE` reçu, mais après son exécution, elle tente de recommencer l'exécution de la ligne `int resultat=n/(int) val;`, ce qui provoque à nouveau un signal `SIGPFE`. Pour éviter ce problème, il faut permettre à la fonction de traitement du signal de modifier la séquence d'exécution de la fonction `main` après la réception du signal. Une solution pourrait être d'utiliser l'instruction `goto` comme suit :

```
  if(*endptr=='\0') {
    int resultat=n/(int) val;
    printf("%d/%d=%d\n",n,(int) val,resultat);
    goto fin:
erreur:
    printf("%d/%d=NaN\n",n,(int) val);
fin:
  }
  else {
    printf("Argument incorrect : %s\n",argv[i]);
   }
// ...

static void sigfpe_handler(int signum) {
 goto erreur:
}
```

En C, ce genre de construction n'est pas possible car l'étiquette d'un `goto` doit nécessairement se trouver dans la même fonction que l'invocation de `goto`. La seule possibilité pour faire ce genre de saut entre fonctions en C, est d'utiliser les fonctions setjmp(3) et longjmp(3). Dans un programme normal, ces fonctions sont à éviter encore plus que les `goto` car elles rendent le code très difficile à lire.

```
#include <setjmp.h>

int setjmp(jmp_buf env);

void longjmp(jmp_buf env, int val);
```

La fonction setjmp(3) est équivalente à la déclaration d'une étiquette. Elle prend comme argument un `jmp_buf`. Cette structure de données, définie dans setjmp.h permet de sauvegarder l'environnement d'exécution, c'est-à-dire les valeurs des registres y compris `%eip` et `%esp` au moment où elle est exécutée. Lorsque setjmp(3) est exécutée dans le flot normal des instructions du programme, elle retourne la valeur `0`. La fonction longjmp(3) prend deux arguments. Le premier est une structure de type `jmp_buf` et le second un entier. Le `jmp_buf` est l'environnement d'exécution qu'il faut restaurer lors de l'exécution de longjmp(3) et le second argument la valeur de retour que doit avoir la fonction setjmp(3) correspondante après l'exécution de longjmp(3).

Le programme ci-dessous illustre l'utilisation de setjmp(3) et longjmp(3).

```
#include <stdio.h>
#include <stdlib.h>
#include <setjmp.h>

jmp_buf label;

void f() {

  printf("Début fonction f\n");
  if(setjmp(label)==0) {
    printf("Exécution normale\n");
  }
  else {
    printf("Exécution après longjmp\n");
  }
}

void g() {
  printf("Début fonction g\n");
  longjmp(label,1);
  printf("Ne sera jamais affiché\n");
}

int main (int argc, char *argv[])  {
  f();
  g();
  return(EXIT_SUCCESS);
}
```

Le programme débute en exécutant la fonction `f`. Dans cette exécution, la fonction setjmp(3) retourne la valeur `0`. Ensuite, la fonction `main` appelle la fonction `g` qui elle exécute `longjmp(label,1)`. Cela provoque un retour à la fonction `f` à l'endroit de l'exécution de `setjmp(label)` qui cette fois-ci va retourner la valeur `1`. Lors de son exécution, le programme ci-dessus affiche :

```
Début fonction f
Exécution normale
Début fonction g
Exécution après longjmp
```

Avec les fonctions setjmp(3) et longjmp(3), il est presque possible d'implémenter le traitement attendu pour le signal `SIGFPE`. Il reste un problème à résoudre. Lorsque la routine de traitement du signal `SIGFPE` s'exécute, ce signal est bloqué par le système d'exploitation jusqu'à ce que cette fonction se termine. Si elle effectue un longjmp(3), elle ne se terminera jamais et le signal continuera à être bloqué. Pour éviter ce problème, il faut utiliser les fonctions sigsetjmp(3) et siglongjmp(3). Ces fonctions sauvegardent dans une structure de données `sigjmp_buf` non seulement l'environnement d'exécution mais aussi la liste des signaux qui sont actuellement bloqués. La fonction sigsetjmp(3) prend un argument supplémentaire : un entier qui doit être non nul si l'on veut

effectivement sauvegarder la liste des signaux bloqués. Lorsque siglongjmp(3) s'exécute, l'environnement et la liste des signaux bloqués sont restaurés.

Le programme ci-dessous présente l'utilisation de sigsetjmp(3) et siglongjmp(3).

```
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <setjmp.h>

sigjmp_buf buf;

static void sigfpe_handler(int);

int main (int argc, char *argv[])  {

  int n=1252;
  if(signal(SIGFPE,sigfpe_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }

  for(int i=1;i<argc;i++) {
    char *endptr;
    int r;
    printf("Traitement de argv[%d]=%s\n",i,argv[i]);
    long val=strtol(argv[i],&endptr,10);
    if(*endptr=='\0') {
      r=sigsetjmp(buf,1);
      if(r==0) {
        int resultat=n/(int) val;
        printf("%d/%d=%d\n",n,(int) val,resultat);
      }
      else {
        printf("%d/%d=NaN\n",n,(int) val);
      }
    }
    else {
      printf("Argument incorrect : %s\n",argv[i]);
    }
  }
  return(EXIT_SUCCESS);
}

static void sigfpe_handler(int signum) {
  // ignorer la donnée et passer à la suivante
  siglongjmp(buf,1);
}
```

Lors de son exécution, il affiche la sortie standard suivante.

```
./sigfpe3 1 2 3 0 a 0 3
Traitement de argv[1]=1
1252/1=1252
Traitement de argv[2]=2
1252/2=626
Traitement de argv[3]=3
1252/3=417
Traitement de argv[4]=0
1252/0=NaN
Traitement de argv[5]=a
Argument incorrect : a
Traitement de argv[6]=0
1252/0=NaN
```

```
Traitement de argv[7]=3
1252/3=417
```

9.3.4 Temporisateurs

Parfois il est nécessaire dans un programme de limiter le temps d'attente pour réaliser une opération. Un exemple simple est lorsqu'un programme attend l'entrée d'un paramètre via l'entrée standard mais peut remplacer ce paramètre par une valeur par défaut si celui-ci n'est pas entré endéans quelques secondes. Lorsqu'un programme attend une information via l'entrée standard, il exécute l'appel système read(2) directement ou via des fonctions de la librairie comme fgets(3) ou getchar(3). Par défaut, celui-ci est bloquant, cela signifie qu'il ne se terminera que lorsqu'une donnée aura été lue. Si read(2) est utilisé seul, il n'est pas possible de borner le temps d'attente du programme et d'interrompre l'appel à read(2) après quelques secondes. Pour obtenir ce résultat, une possibilité est d'utiliser un signal. En effet, read(2) est un appel système lent qui peut être interrompu par la réception d'un signal. Il y a plusieurs façons de demander qu'un signal soit généré après un certain temps. Le plus général est setitimer(2). Cet appel système permet de générer un signal `SIGALRM` après un certain temps ou périodiquement. L'appel système alarm(3posix) est plus ancien mais plus simple à utiliser que setitimer(2). Nous l'utilisons afin d'illustrer comment un signal peut permettre de limiter la durée d'attente d'un appel système.

```c
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <unistd.h>
#include <stdbool.h>

static void sig_handler(int);

int main (int argc, char *argv[])  {
  char c;
  printf("Tapez return en moins de 5 secondes !\n");
  fflush(stdout);
  if(signal(SIGALRM,sig_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
  // sigalrm interrompt les appels système
  if(siginterrupt(SIGALRM,true)<0) {
    perror("siginterrupt");
    exit(EXIT_FAILURE);
  }
  alarm(5);
  int r=read(STDIN_FILENO,&c,1);
  if((r==1)&&(c=='\n')) {
    alarm(0); // reset timer
    printf("Gagné\n");
    exit(EXIT_SUCCESS);
  }
  printf("Perdu !\n");
  exit(EXIT_FAILURE);
}

static void sig_handler(int signum) {
  // rien à faire, read sera interrompu
}
```

Ce programme utilise alarm(3posix) pour limiter la durée d'un appel système read(2). Pour ce faire, il enregistre d'abord une fonction pour traiter le signal `SIGALRM`. Cette fonction est vide dans l'exemple, son exécution permet juste d'interrompre l'appel système read(2). Par défaut, lorsqu'un signal survient durant l'exécution d'un appel système, celui-ci est automatiquement redémarré par le système d'exploitation pour éviter à l'application de devoir traiter tous les cas possibles d'interruption d'appels système. La fonction siginterrupt(3) permet de modifier ce comportement par défaut et nous l'utilisons pour forcer l'interruption d'appels systèmes lorsque le signal

`SIGALRM` est reçu. L'appel à `alarm(0)` permet de désactiver l'alarme qui était en cours.

Lors de son exécution, ce programme affiche la sortie suivante.

```
Tapez return en moins de 5 secondes !
Perdu !
```

En essayant le programme ci-dessus, on pourrait conclure qu'il fonctionne parfaitement. Il a cependant un petit défaut qui peut s'avérer genant si par exemple on utilise la même logique pour écrire une fonction `read_time` qui se comporte comme read(2) sauf que son dernier argument est un délai maximal. Sur un système fort chargé, il est possible qu'après l'exécution de `alarm(5)` le processus soit mis en attente par le système d'exploitation qui exécute d'autres processus. Lorsque l'alarme expire, la fonction de traitement de `SIGALRM` est exécutée puis seulement l'appel à read(2) s'effectue. Celui-ci étant bloquant, le processus restera bloqué jusqu'à ce que les données arrivent ce qui n'est pas le comportement attendu.

Pour éviter ce problème, il faut empêcher l'exécution de read(2) si le signal `SIGALRM` a déjà été reçu. Cela peut se réaliser en utilisant sigsetjmp(3) pour définir une étiquette avant l'exécution du bloc contenant l'appel à alarm(3posix) et l'appel à read(2). Si le signal n'est pas reçu, l'appel à read(2) s'effectue normalement. Si par contre le signal `SIGALRM` est reçu entre l'appel à alarm(3posix) et l'appel à read(2), alors l'exécution de siglongjmp(3) dans `sig_handler` empêchera l'exécution de l'appel système read(2) ce qui est bien le comportement attendu.

```c
#include <stdio.h>
#include <stdlib.h>
#include <signal.h>
#include <unistd.h>
#include <stdbool.h>
#include <setjmp.h>

sigjmp_buf env;

static void sig_handler(int);

int main (int argc, char *argv[])  {
  char c;
  printf("Tapez return en moins de 5 secondes !\n");
  fflush(stdout);
  if(signal(SIGALRM,sig_handler)==SIG_ERR) {
    perror("signal");
    exit(EXIT_FAILURE);
  }
  // sigalrm interrompt les appels système
  if(siginterrupt(SIGALRM,true)<0) {
    perror("siginterrupt");
    exit(EXIT_FAILURE);
  }
  int r=0;
  if(sigsetjmp(env,1)==0) {
    // sig_handler n'a pas encore été appelé
    alarm(5);
    r=read(STDIN_FILENO,&c,1);
  }
  else {
    // sig_handler a déjà été exécuté
    // le délai a déjà expiré, inutile de faire read
  }
  alarm(0); // arrêt du timer
  if((r==1)&&(c=='\n')) {
    printf("Gagné\n");
    exit(EXIT_SUCCESS);
  }
  else {
    printf("Perdu !\n");
```

```
    exit(EXIT_FAILURE);
  }
}

static void sig_handler(int signum) {
  siglongjmp(env,1);
}
```

L'appel système alarm(3posix) s'appuie sur setitimer(2), mais les deux types d'alarmes ne doivent pas être combinés. Il en va de même pour sleep(3) qui peut être implémenté en utilisant `SIGALRM`. Linux contient d'autres appels systèmes et fonctions pour gérer différents types de temporisateurs. Ceux-ci sont décrits de façon détaillée dans [Kerrisk2010].

Note : Signaux, threads, fork(2) et execve(2)

Le noyau du système d'exploitation maintient pour chaque processus une structure de données contenant la liste des signaux qui sont ignorés, ont été reçus et les pointeurs vers les fonctions de traitement pour chaque signal. Cette structure de données est associée à chaque processus. La création de threads ne modifie pas cette structure de données et lorsqu'un signal est délivré à un processus utilisant des threads, c'est généralement le thread principal qui recevra et devra traiter le signal. Lors de l'exécution de fork(2), la structure de données relative aux signaux du processus père est copiée dans le processus fils. Après fork(2), les deux processus peuvent évoluer séparément et le fils peut par exemple modifier la façon dont il traite un signal sans que cela n'affecte le processus père. Lors de l'exécution de execve(2), la structure de données relative aux signaux est réinitialisée avec les traitements par défaut pour chacun des signaux.

9.4 Sémaphores nommés

Nous avons présenté les sémaphores lors de l'étude du fonctionnement des threads et les mécanismes qui permettent de les coordonner. Chaque sémaphore utilise une structure de données qui est dans une mémoire accessible aux différents threads/processus qui doivent se coordonner. Lorsque des sémaphores sont utilisés pour coordonner des threads, cette structure de données est soit stockée dans la zone réservée aux variables globales, soit sur le heap (tas). Lorsque des processus doivent se coordonner, il ne partagent pas nécessairement [7] de la mémoire. Dans ce cas, la fonction sem_init(3) ne peut pas être utilisée pour créer de sémaphore. Par contre, il est possible d'utiliser des sémaphores "nommés" (named semaphores). Ces sémaphores utilisent une zone de mémoire qui est gérée par le noyau et qui peut être utilisée par plusieurs processus. Un *sémaphore nommé* est identifié par un *nom*. Sous Linux, ce nom correspond à un fichier sur un système de fichiers et tout processus qui connait le nom d'un sémaphore et possède les permissions l'autorisant à y accéder peut l'utiliser. Trois appels systèmes sont utilisés pour créer, utiliser et supprimer un sémaphore nommé. Une présentation générale des sémaphores est disponible dans la page de manuel sem_overview(7).

```
#include <fcntl.h>           /* For O_* constants */
#include <sys/stat.h>        /* For mode constants */
#include <semaphore.h>

sem_t *sem_open(const char *name, int oflag);
sem_t *sem_open(const char *name, int oflag,
                mode_t mode, unsigned int value);
int sem_close(sem_t *sem);
int sem_unlink(const char *name);
```

sem_open(3) permet d'ouvrir ou de créer un *sémaphore nommé*. Tout comme l'appel système open(2), il existe deux variantes de sem_open(3). La première prend 2 arguments et permet d'utiliser un *sémaphore nommé* déjà existant. La seconde prend quatre arguments et permet de créer un nouveau *sémaphore nommé* et l'initialise à la valeur du quatrième argument. La fonction sem_open(3) s'utilise de la même façon que l'appel système open(2) sauf qu'elle retourne un pointeur vers une structure de type `sem_t` et qu'en cas d'erreur elle retourne `SEM_FAILED` et utilise `errno` pour fournir une information sur le type d'erreur. La fonction sem_close(3)

7. Nous verrons dans le prochain chapitre comment plusieurs processus peuvent partager une même zone mémoire.

permet à un processus d'arrêter d'utiliser un *sémaphore nommé*. Enfin, la fonction sem_unlink(3) permet de supprimer un sémaphore qui avait été créé avec sem_open(3). Tout comme il est possible d'utiliser l'appel système unlink(2) sur un fichier ouvert par un processus, il est possible d'appeler sem_unlink(3) avec comme argument un nom de sémaphore utilisé actuellement par un processus. Dans ce cas, le sémaphore sera complètement libéré lorsque le dernier processus qui l'utilise aura effectué sem_close(3). Les fonctions sem_post(3) et sem_wait(3) s'utilisent de la même façon qu'avec les sémaphores non-nommés que nous avons utilisé précédemment.

A titre d'exemple, considérons un exemple simple d'utilisation de sémaphores nommés dans lequel un processus doit attendre la fin de l'exécution d'une fonction dans un autre processus pour pouvoir s'exécuter. Avec les threads, nous avions résolu ce problème en initialisant un sémaphore à `0` dans le premier thread alors que le second démarrait par un sem_wait(3). Le première exécute sem_post(3) dès qu'il a fini l'exécution de sa fonction critique.

Le programme ci-dessous illustre le processus qui s'exécute en premier.

```
sem_t *semaphore;

void before() {
  // do something
  for(int j=0;j<1000000;j++) {
  }
  printf("before done, pid=%d\n",(int) getpid());
  sem_post(semaphore);
}


int main (int argc, char *argv[])  {

  int err;

  semaphore=sem_open("lsinf1252",O_CREAT,S_IRUSR | S_IWUSR,0);
  if(semaphore==SEM_FAILED) {
    error(-1,"sem_open");
  }
  sleep(20);
  before();
  err=sem_close(semaphore);
  if(err!=0) {
      error(err,"sem_close");
  }
  return(EXIT_SUCCESS);
}
```

Ce processus commence par utiliser sem_open(3) pour créer un sémaphore qui porte le nom `lsinf1252` et est initialisé à zéro puis se met en veille pendant vingt secondes. Ensuite il exécute la fonction `before` qui se termine par l'exécution de `sem_post(semaphore)`. Cet appel a pour résultat de libérer le second processus dont le code est présenté ci-dessous :

```
sem_t *semaphore;

void after() {
  sem_wait(semaphore);
  // do something
  for(int j=0;j<1000000;j++) {
  }
  printf("after done, pid=%d\n",(int) getpid());
}

int main (int argc, char *argv[])  {
  int err;

  // semaphore a été créé par before
  semaphore=sem_open("lsinf1252",0);
  if(semaphore==SEM_FAILED) {
```

```
        error(-1,"sem_open");
    }
    after();

    err=sem_close(semaphore);
    if(err!=0) {
        error(err,"sem_close");
    }
    err=sem_unlink("lsinf1252");
    if(err!=0) {
        error(err,"sem_unlink");
    }
    return(EXIT_SUCCESS);
}
```

Ce processus utilise le sémaphore qui a été créé par un autre processus pour se coordonner. La fonction `after` démarre par l'exécution de sem_wait(3) qui permet d'attendre que l'autre processus ait terminé l'exécution de la fonction `before`. Les sémaphores nommés peuvent être créés et supprimés comme des fichiers. Il est donc normal que le sémaphore soit créé par le premier processus et supprimé par le second. Sous Linux, les sémaphores nommés sont créés comme des fichiers dans le système de fichiers virtuel `/dev/shm` :

```
$ ls -l /dev/shm/
-rw------- 1 obo          stafinfo       32 Apr 10 15:37 sem.lsinf1252
```

Les permissions du fichier virtuel représentent les permissions associées au sémaphore. La sortie ci-dessous présente une exécution des deux processus présentés plus haut.

```
$ ./process-sem-before &
[1] 5222
$ ./process-sem-after
before done, pid=5222
after done, pid=5223
[1]+  Done                    ./process-sem-before
```

Il est important de noter que les sémaphores nommés sont une ressource généralement limitée. Lorsqu'il a été créé, un sémaphore nommé utilise des ressources du système jusqu'à ce qu'il soit explicitement supprimé avec sem_unlink(3). Il est très important de toujours bien effacer les sémaphores nommés dès qu'ils ne sont plus nécessaires. Sans cela, l'espace réservé pour ces sémaphores risque d'être rempli et d'empêcher la création de nouveaux sémaphores par d'autres processus.

9.5 Partage de fichiers

Les fichiers sont l'un des principaux moyens de communication entre processus. L'avantage majeur des fichiers est leur persistence. Les données sauvegardées dans un fichier persistent sur le système de fichiers après la fin du processus qui les a écrites. L'inconvénient majeur de l'utilisation de fichiers par rapport à d'autres techniques de communication entre processus est la relative lenteur des dispositifs de stockage en comparaison avec les accès à la mémoire. Face à cette lenteur des dispositifs de stockage, la majorité des systèmes d'exploitation utilisent des buffers qui servent de tampons entre les processus et les dispositifs de stockage. Lorsqu'un processus écrit une donnée sur un dispositif de stockage, celle-ci est d'abord écrite dans un buffer géré par le système d'exploitation et le processus peut poursuivre son exécution sans devoir attendre l'exécution complète de l'écriture sur le dispositif de stockage. La taille de ces buffers varie généralement dynamiquement en fonction de la charge du système. Les données peuvent y rester entre quelques fractions de seconde et quelques dizaines de secondes. Un processus peut contrôler l'utilisation de ce buffer en utilisant l'appel système fsync(2). Celui-ci permet de forcer l'écriture sur le dispositif de stockage des données du fichier identifié par le descripteur de fichiers passé en argument. L'appel système sync(2) force quant à lui l'écriture de toutes les données actuellement stockées dans les buffers du noyau sur les dispositifs de stockage. Cet appel système est notamment utilisé par un processus système qui l'exécute toutes les trente secondes afin d'éviter que des données ne restent trop longtemps dans les buffers du noyau.

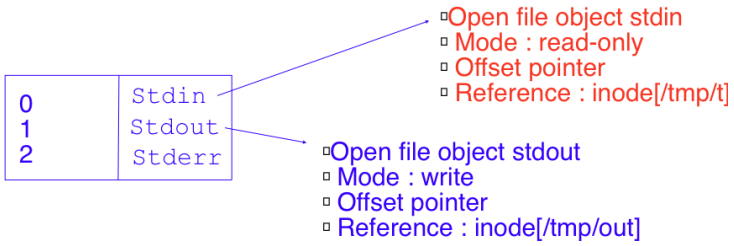
L'utilisation d'un même fichier par plusieurs processus est une des plus anciennes techniques de communication entre processus. Pour comprendre son fonctionnement, il est utile d'analyser les structures de données qui sont maintenues par le noyau du système d'exploitation pour chaque fichier et chaque processus. Comme nous l'avons présenté dans le chapitre précédent, le système de fichiers utilise des inodes pour stocker les méta-données et la liste des blocs de chaque fichier. Lorsqu'un processus ouvre un fichier, le noyau du système d'exploitation lui associe le premier descripteur de fichier libre dans la table des descripteurs de fichiers du processus. Ce descripteur de fichier pointe alors vers une structure maintenue par le noyau qui est souvent appelée un *open file object*. Un *open file object* contient toutes les informations qui sont nécessaires au noyau pour pouvoir effectuer les opérations de manipulation d'un fichier ouvert par un processus. Parmi celles-ci, on trouve notamment :

– le mode dans lequel le fichier a été ouvert (lecture seule, écriture, lecture/écriture). Ce mode est initialisé à l'ouverture du fichier. Le noyau vérifie le mode lors de l'exécution des appels systèmes read(2) et write(2) mais pas les permissions du fichier sur le système de fichiers.
– l'offset pointer qui est la tête de lecture/écriture dans le fichier
– une référence vers le fichier sur le système de fichiers. Dans un système de fichiers Unix, il s'agit généralement du numéro de l'*inode* du fichier ou d'un pointeur vers une structure contenant cet *inode* et des informations comme le dispositif de stockage sur lequel il est stocké.

A titre d'exemple, considérons l'exécution de la commande suivante depuis le shell :

```
$ cat < /tmp/t > /tmp/out
```

Lors de son exécution, deux open file objects sont créés dans le noyau. Le premier est relatif au fichier `/tmp/t` qui est associé au descripteur `stdin`. Le second est lié au fichier `/tmp/out` et est associé au descripteur `stdout`. Ces open-file objects sont représentés graphiquement dans la figure ci-dessous.



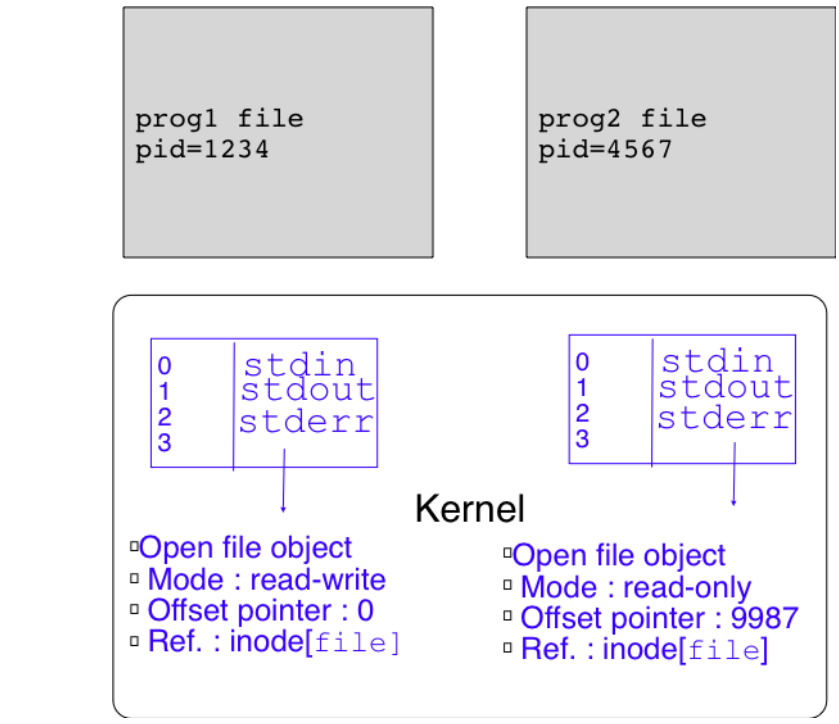
Les open file objects sont également utilisés lorsque plusieurs processus ouvrent le même fichier. Considérons l'exécution simultanée des deux commandes suivantes :

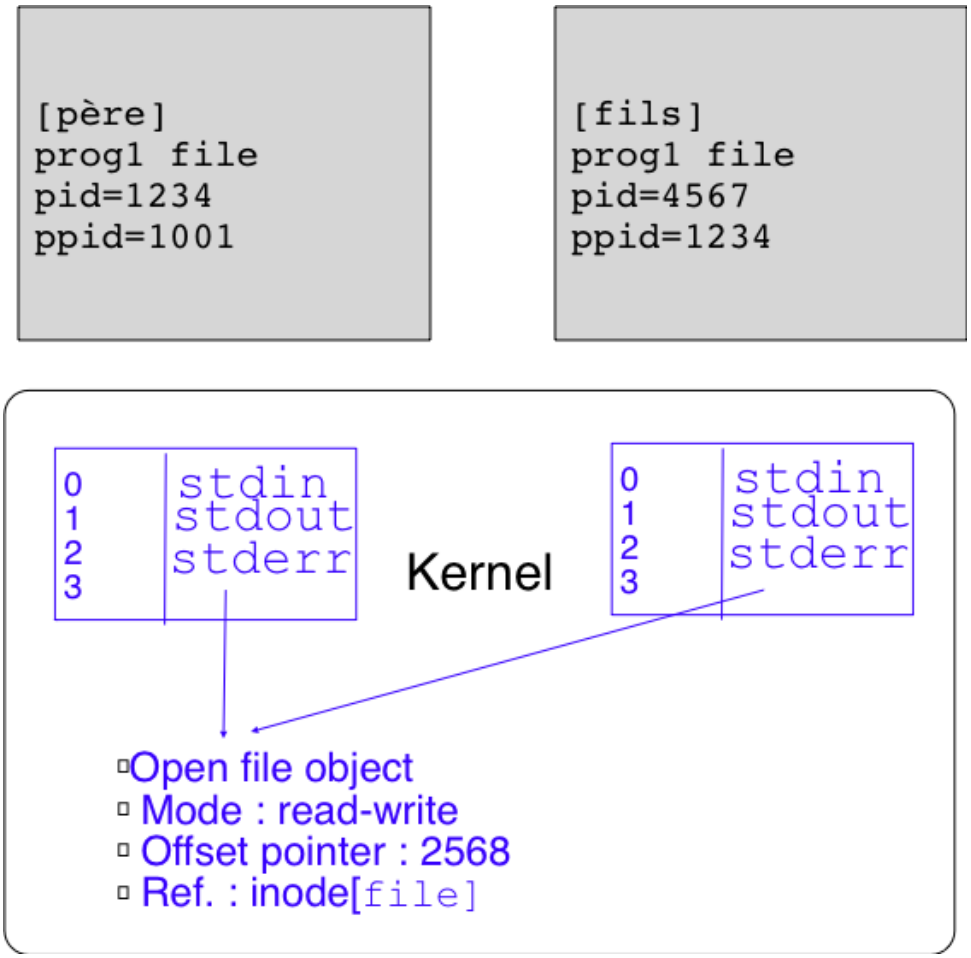
```
$ prog1 file &
$ prog2 file &
```

Lors de l'exécution de ces deux processus, le noyau va attribuer un descripteur de fichier à chacun d'eux. Si `file` est le premier fichier ouvert par chaque processus, il sera associé au descripteur `3`. Le noyau créera un open-file object pour le fichier `file` utilisé par le processus `prog1` et un autre open-file object pour le fichier `file` utilisé par le processus `prog2`. Ces deux open-file objects référencient le même inode et donc le même fichier, mais ils peuvent avoir des modes et des offset pointers qui sont totalement indépendants. Tous les accès faits par le processus `prog2` sont complètement indépendants des accès faits par le processus `prog1`. Cette utilisation d'un même fichier par deux processus distincts est représentée graphiquement dans la figure ci-dessous.

Sous Unix et Linux, il est important d'analyser également ce qu'il se passe lors de la création d'un processus en utilisant l'appel système fork(2). Imaginons que le processus `prog1` discuté ci-dessous effectue fork(2) après avoir ouvert le fichier `file`. Dans ce cas, le noyau du système d'exploitation va dupliquer la table des descripteurs de fichiers du processus père pour créer celle du processus fils. Le processus père et le processus fils ont donc chacun une table des descripteurs de fichiers qui leur est propre. Cela permet, comme nous l'avons vu lorsque nous avons présenté les pipes, que le fils ferme un de ses descripteurs de fichiers sans que cela n'ait d'impact sur l'utilisation de ce descripteur de fichier par le processus père. Par contre, l'exécution de l'appel système fork(2) ne copie pas les open-file objects. Après exécution de fork(2) le descripteur de fichiers `3` dans le processus père pointe vers l'open-file object associé au fichier `file` et le même descripteur dans le processus fils pointe vers le même open-file object. Cette situation est représentée schématiquement dans la figure ci-dessous.

Cette utilisation d'un même open-file object par le processus père et le processus fils est une particularité importante de Unix. Elle permet aux deux processus d'écrire des données séquentiellement dans un fichier qui avait


prog1 file
pid=1234
prog2 file
pid=4567
0
1
2
3
stdin
stdout
stderr
Kernel
Open file object
Mode : read-write
Offset pointer : 0
Ref. : inode[file]
Open file object
Mode : read-only
Offset pointer : 9987
Ref. : inode[file]


[père]
prog1 file
pid=1234
ppid=1001
[fils]
prog1 file
pid=4567
ppid=1234
0
1
2
3
stdin
stdout
stderr
Kernel
Open file object
Mode : read-write
Offset pointer : 2568
Ref. : inode[file]

été initialement ouvert par le processus père, mais pose régulièrement des problèmes lors de la manipulation de fichiers.

Lorsqu'un fichier est utilisé par plusieurs processus simultanément, il est nécessaire de coordonner les activités de ces processus pour éviter que le fichier ne devienne corrompu. Outre les appels systèmes classiques open(2), read(2), write(2) et close(2), Unix offre plusieurs appels systèmes qui sont utiles lorsque plusieurs processus accèdent au même fichier.

Considérons d'abord un processus père et un processus fils qui doivent lire des données à des endroits particuliers dans un fichier. Pour cela, il est naturel d'utiliser l'appel système lseek(2) pour déplacer l'offset pointer et d'ensuite réaliser la lecture avec read(2). Malheureusement lorsqu'un père et un ou plusieurs fils [8] utilisent ces appels systèmes, il est possible qu'un appel à lseek(2) fait par le fils soit immédiatement suivi d'un appel à lseek(2) fait par le père avant que le fils ne puisse exécuter l'appel système read(2). Dans ce cas, le processus fils ne lira pas les données qu'il souhaitait lire dans le fichier. Les appels systèmes pread(2) et pwrite(2) permettent d'éviter ce problème. Ils complètent les appels systèmes read(2) et write(2) en prenant comme argument l'offset auquel la lecture ou l'écriture demandée doit être effectuée. pread(2) et pwrite(2) garantissent que les opérations d'écriture et de lecture qui sont effectuées avec ces appels systèmes seront atomiques.

```
#include <unistd.h>
ssize_t pread(int fd, void *buf, size_t count, off_t offset);

ssize_t pwrite(int fd, const void *buf, size_t count, off_t offset);
```

Ce n'est pas le seul problème qui se pose lorsque plusieurs processus manipulent un même fichier. Considérons un logiciel de base de données qui comprend des processus qui lisent dans des fichiers qui constituent la base de données et d'autres qui modifient le contenu de ces fichiers. Ces opérations d'écritures et de lectures dans des fichiers partagés risquent de provoquer des problèmes d'accès concurrent similaires aux problèmes que nous avons dû traiter lorsque plusieurs threads se partagent une même mémoire. Pour réguler ces accès à des fichiers, Unix et Linux supportent des verrous (locks en anglais) que l'on peut associer à des fichiers. A première vue, un *lock* peut être comparé à un *mutex*. Un *lock* permet à un processus d'obtenir l'accès exclusif à un fichier ou une partie de fichier tout comme un *mutex* est utilisé pour réguler les accès à une variable. En théorie, il existe deux techniques d'utilisation de locks qui peuvent être utilisées sur un système Unix :

– *mandatory locking*. Dans ce cas, les processus placent des locks sur certains fichiers ou zones de fichiers et le système d'exploitation vérifie qu'aucun accès fait aux fichiers avec les appels systèmes standards ne viole ces locks.
– *advisory locking*. Dans ce cas, les processus doivent vérifier eux-mêmes que les accès qu'ils effectuent ne violent pas les locks qui ont été associés aux différents fichiers.

Certains systèmes Unix supportent les deux stratégies de locking, mais la plupart ne supportent que l'*advisory locking*. L'*advisory locking* est la stratégie la plus facile à implémenter dans le système d'exploitation. C'est aussi celle qui donne les meilleures performances. Nous limitons notre description à l'*advisory locking*. Le *mandatory locking* nécessite un support spécifique du système de fichiers qui sort du cadre de ce cours.

Deux appels systèmes sont utilisés pour manipuler les locks qui peuvent être associés aux fichiers : flock(2) et fcntl(2). flock(2) est la solution la plus simple. Cet appel système permet d'associer un verrou à un fichier complet.

```
#include <sys/file.h>
int flock(int fd, int operation);
```

Il prend comme argument un descripteur de fichier et une opération. Deux types de locks sont supportés. Un lock est dit partagé (shared lock, `operation==LOCK_SH`) lorsque plusieurs processus peuvent posséder un même lock vers un fichier. Un lock est dit exclusif (exclusive lock, `operation==LOCK_EX`) lorsqu'un seul processus peut posséder un lock vers un fichier à un moment donné. Il faut noter que les locks sont associés aux fichiers (et donc indirectement aux inodes) et non aux descripteurs de fichiers. Pour retirer un lock associé à un fichier, il faut utiliser `LOCK_UN` comme second argument à l'appel flock(2).

L'appel système fcntl(2) et la fonction lockf(3) sont nettement plus flexibles. Ils permettent de placer des locks sur une partie d'un fichier. lockf(3) prend trois arguments : un descripteur de fichiers, une commande et un entier qui indique la longueur de la section du fichier à associer au lock.

8. Même si il n'a pas été mentionné lors de l'utilisation de threads, ce problème se pose également lorsque plusieurs threads accèdent directement aux données dans un même fichier.

```
#include <unistd.h>
int lockf(int fd, int cmd, off_t len);
```

lockf(3) supporte plusieurs commandes qui sont chacune spécifiées par une constante définie dans unistd.h. La commande `F_LOCK` permet de placer un lock exclusif sur une section du fichier dont le descripteur est le premier argument. Le troisième argument est la longueur de la section sur laquelle le lock doit être appliqué. Par convention, la section s'étend de la position actuelle de l'offset pointer (`pos`) jusqu'à `pos+len-1` si `len` est positif et va de `pos-len` à `pos-1` si `len` est négatif. Si `len` est nul, alors la section concernée par le lock va de `pos` à l'infini (c'est-à-dire jusqu'à la fin du fichier, même si celle-ci change après l'application du lock).

L'appel à lockf(3) bloque si un autre processus a déjà un lock sur une partie du fichier qui comprend celle pour laquelle le lock est demandé. La commande `F_TLOCK` est équivalente à `F_LOCK` avec comme différence qu'elle ne bloque pas si un lock existe déjà sur le fichier mais retourne une erreur. `F_TEST` permet de tester la présence d'un lock sans tenter d'acquérir ce lock contrairement à `F_TLOCK`. Enfin, la commande `F_ULOCK` permet de retirer un lock placé précédemment.

En pratique, lockf(3) doit s'utiliser de la même façon qu'un mutex. C'est-à-dire qu'un processus qui souhaite écrire de façon exclusive dans un fichier doit d'abord obtenir via lockf(3) un lock sur la partie du fichier dans laquelle il souhaite écrire. Lorsque l'appel à lockf(3) réussit, il peut effectuer l'écriture via write(2) et ensuite libérer le lock en utilisant lockf(3). Tout comme avec les mutex, si un processus n'utilise pas lockf(3) avant d'écrire ou de lire, cela causera des problèmes.

L'appel système fcntl(2) est un appel "fourre-tout" qui regroupe de nombreuses opérations qu'un processus peut vouloir faire sur un fichier. L'application de locks est une de ces opérations, mais la page de manuel en détaille de nombreuses autres. Lorsqu'il est utilisé pour manipuler des locks, l'appel système fcntl(2) utilise trois arguments :

```
#include <unistd.h>
#include <fcntl.h>

int fcntl(int fd, int cmd, struct flock* );
```

Le premier argument est le descripteur de fichiers sur lequel le lock doit être appliqué. Le second est la commande. Tout comme lockf(3), fcntl(2) supporte différentes commandes qui sont spécifiées dans la page de manuel. Le troisième argument est un pointeur vers une structure `flock` qui doit contenir au minimum les champs suivants :

```
struct flock {
   ...
   short l_type;    /* Type of lock: F_RDLCK,
                       F_WRLCK, F_UNLCK */
   short l_whence;  /* How to interpret l_start:
                       SEEK_SET, SEEK_CUR, SEEK_END */
   off_t l_start;   /* Starting offset for lock */
   off_t l_len;     /* Number of bytes to lock */
   pid_t l_pid;     /* PID of process blocking our lock
                       (F_GETLK only) */
   ...
   };
```

Cette structure permet de spécifier plus finement qu'avec la fonction lockf(3) la section du fichier sur laquelle le lock doit être placé. L'utilisation de locks force le noyau à maintenir des structures de données supplémentaires pour stocker ces locks et les processus qui peuvent être en attente sur chacun de ces locks. Conceptuellement, cette structure de données est associée à chaque fichier comme représenté dans la figure ci-dessous.

Comme les locks sont associés à des fichiers, le noyau doit maintenir pour chaque fichier ouvert une liste de locks. Celle-ci peut être implémentée sous la forme d'une liste chaînée comme représenté ci-dessus ou sous la forme d'une autre structure de données. Le point important est que le noyau doit mémoriser pour chaque fichier utilisant des locks quelles sont les sections du fichier qui sont concernées par chaque lock, quel est le type de lock, quel est le processus qui possède le lock (pour autoriser uniquement ce processus à le retirer) et enfin une liste ou une queue des processus qui sont en attente sur ce lock.

Sous Linux, le système de fichiers virtuel `/proc` fournit une interface permettant de visualiser l'état des locks. Il suffit pour cela de lire le contenu du fichier `/proc/locks`.

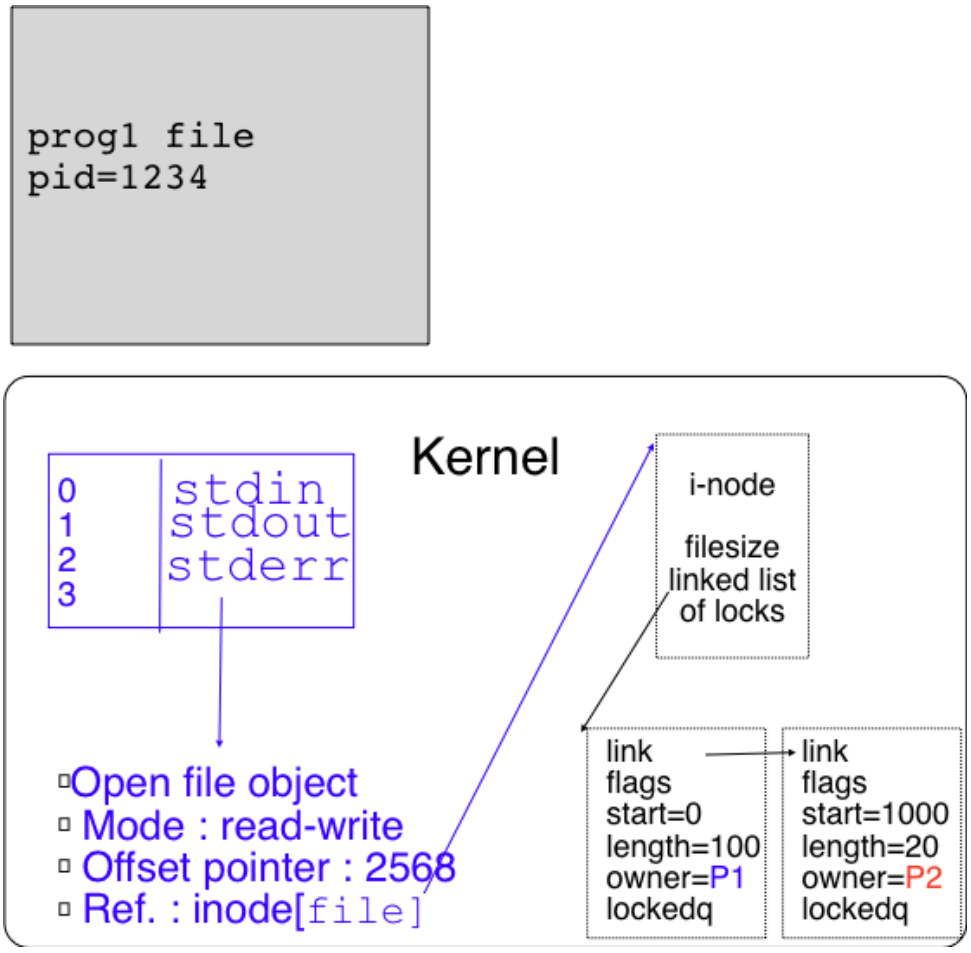


FIGURE 9.3 – Gestion des locks par le kernel

```
cat /proc/locks
1: POSIX  ADVISORY  WRITE 12367 00:17:20628657 0 0
2: POSIX  ADVISORY  WRITE 12367 00:17:7996086 0 0
3: POSIX  ADVISORY  WRITE 12367 00:17:24084665 0 0
4: POSIX  ADVISORY  WRITE 12367 00:17:12634137 0 0
5: POSIX  ADVISORY  WRITE 30677 00:17:14534587 1073741824 1073742335
6: POSIX  ADVISORY  READ  30677 00:17:25756362 128 128
7: POSIX  ADVISORY  READ  30677 00:17:25756359 1073741826 1073742335
8: POSIX  ADVISORY  READ  30677 00:17:25756319 128 128
9: POSIX  ADVISORY  READ  30677 00:17:25756372 1073741826 1073742335
10: POSIX  ADVISORY  WRITE 30677 00:17:25757269 1073741824 1073742335
11: POSIX  ADVISORY  WRITE 30677 00:17:25756354 0 EOF
12: FLOCK  ADVISORY  WRITE 22677 00:18:49578023 0 EOF
13: POSIX  ADVISORY  WRITE 3023 08:01:652873 0 EOF
14: POSIX  ADVISORY  WRITE 3014 08:01:652855 0 EOF
15: POSIX  ADVISORY  WRITE 2994 08:01:652854 0 EOF
16: FLOCK  ADVISORY  WRITE 2967 08:01:798437 0 EOF
17: FLOCK  ADVISORY  WRITE 2967 08:01:797391 0 EOF
18: FLOCK  ADVISORY  WRITE 1278 08:01:652815 0 EOF
```

Dans ce fichier, la première colonne indique le type de lock (`POSIX` pour un lock placé avec fcntl(2) ou lockf(3) et `FLOCK` pour un lock placé avec flock(2)). La deuxième indique si le lock est un *advisory lock* ou un *mandatory lock*. La troisième spécifie si le lock protège l'écriture et/ou la lecture. La quatrième colonne est l'identifiant du processus qui possède le lock. La cinquième précise le dispositif de stockage et le fichier concerné par ce lock (le dernier nombre est l'inode du fichier). Enfin, les deux dernières colonnes spécifient la section qui est couverte par le lock avec `EOF` qui indique la fin du fichier.

9.6 Mise en œuvre des systèmes de fichiers

Nous avons vu dans les sections précédentes comment un système d'exploitation comme UNIX permettait un accès et une interface unifiés aux périphériques de stockage via le système de fichier.

De nombreuses mises en œuvre de systèmes de fichiers existent. Elles répondent à des besoins différents en terme de performance ou de fiabilité, et visant des technologies de stockage variées. Par exemple, un système de fichier pour le stockage sur bande magnétique à très grande capacité, utilisées pour de l'archivage de données à long terme, sera l'objet de choix de conception très différents de ceux faits pour un système de fichiers généralistes destiné à héberger le système d'exploitation, les logiciels, et les données d'un utilisateur de machine personnelle sur un disque dur ou un SSD. Dans le cas d'un stockage sur bande, le système de fichier doit prendre en compte le fait que l'accès se fait de façon strictement linéaire (la bande magnétique est déroulée devant une tête de lecture), et que l'objectif est de maximiser la performance d'écriture des archives, si nécessaire au détriment de la rapidité d'accès à un fichier en particulier. Dans le cas d'un système de fichiers généraliste en revanche, il est nécessaire de supporter de manière efficace des accès aléatoires à des fichiers et répertoires, tout en conservant des performances correctes pour les accès linéaires à de grands fichiers.

Nous allons explorer dans cette section des éléments de mise en œuvre des systèmes de fichiers en nous concentrant sur le cas des systèmes de fichiers généralistes utilisés dans les systèmes UNIX et en particulier le système de fichier ext4. Ce système de fichier est le plus couramment utilisé par les distributions de Linux.

9.6.1 Organisation d'un disque et des partitions

Chaque périphérique de stockage (disque, clé USB, etc.) est géré par un gestionnaire de périphérique spécifique, associé à un driver de périphérique. Ces deux éléments forment l'interface entre le matériel et le logiciel, pilotée par le kernel du système d'exploitation. Les différents périphériques de stockage ont des organisations physiques très différentes. La figure suivante présente l'exemple d'un disque dur.

Un disque dur est formé de plusieurs plateaux (A), qui sont des disques rotatifs sur lesquels une tête de lecture/écriture (C) peut lire ou écrire des données en agissant sur la polarisation magnétique de la surface du

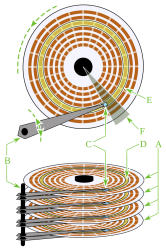


FIGURE 9.4 – Structure d'un disque dur. Crédits : Wikimedia / Domaine public

matériau. Le plus souvent, les deux faces d'un même plateau sont utilisées. L'écriture sur chaque se fait sur des pistes concentriques (E) et chaque piste est divisée en secteurs (F).

Différents disques durs auront des caractéristiques différentes comme leur nombre de plateau ou de têtes de lecture/écriture. Il en va de même pour un périphérique de stockage de type SSD, utilisant des puces de mémoire non volatile (mémoire flash) dont la disposition et la configuration peut différer de manière importante d'un modèle à l'autre.

L'utilisation d'un driver et d'un contrôleur de périphérique, mais aussi de l'électronique de gestion embarquée du disque dur ou du SSD permet toutefois au système d'exploitation de ne pas avoir à gérer directement ces notions spécifiques de plateau, secteur, pistes, ou de puces de mémoire flash, lors de la lecture ou de l'écriture de données. À la place, un périphérique de stockage est vu comme une suite de *secteurs* contigües, formant un espace d'adressage linéaire. Une taille typique de secteur pour un disque dur est de 512 octets, tandis qu'un SSD présente souvent les données stockées à une granularité de 4 Kilo-octets. Nous ne ferons pas de distinction entre disque dur et SSD dans la suite du chapitre, et utiliserons l'appellation générique de "disque" pour couvrir les deux types de périphériques de stockage.

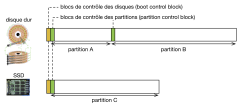


FIGURE 9.5 – Disques, partitions et blocs de contrôle.

Un disque comporte toujours une zone de démarrage permettant de stocker des métadonnées utile pour sa gestion. Ce bloc de contrôle du disque est souvent appelé en anglais *boot control block* car un de ses rôles principaux est de contenir, à une adresse fixe, le code de démarrage du système d'exploitation lorsque ce disque joue le rôle de disque de démarrage (les gestionnaires de démarrage comme `grub` permettent de démarrer des systèmes d'exploitation différents stockés sur des partitions différentes du même disque ou sur des disques différents). Outre ce code de démarrage, ce bloc de contrôle contient la liste des *partitions* du disque. Chaque partition peut stocker des données en utilisant un système de fichier différent. Sur notre exemple, le disque SSD ne contient qu'une partition, tandis que le disque dur en contient deux. La partition C sur le disque SSD pourrait, par exemple, utiliser le système de fichiers ext4, la partition A sur le disque dur être utilisée comme espace de *swap* pour le système de gestion de la mémoire virtuelle, et la partition B utiliser le système de fichiers ExFAT. Chaque partition démarre par son propre bloc de contrôle. Le rôle de ce bloc de contrôle est de contenir les métadonnées nécessaires pour réaliser l'opération de montage du système de fichier contenu dans la partition. Sa structure est spécifique au système de fichier utilisé, mais des informations classiques y figurant incluent l'organisation de l'index des fichiers ou de l'index de l'espace libre, ou encore une indication sur le fait que ls système de fichier a été *démonté* proprement ou non lors de sa dernière utilisation. Nous reviendrons sur l'ensemble de ces notions au cours de notre exploration de la mise en œuvre d'un système de fichier qui suit.

Note : Le cas de la partition de swap

Nous avons vu dans la section consacrée à la mémoire virtuelle qu'une *partition de swap* pouvait être crée avec mkswap(8) et activée avec swapon(8). La partition de swap ne contient pas à proprement parler de système de fichier. Elle est traitée comme un grand espace uniforme par le système d'exploitation pour y stocker les copies des pages mémoire évincées par l'algorithme de remplacement de page. La partition de swap est scindée en autant

de cadres de pages que possible (par exemple, de 4 Ko sur les systèmes de la famille x86 comme IA-32). Seul le premier cadre de page est utilisé pour stocker les métadonnées permettant l'utilisation de la partition de swap.

9.6.2 Structure du système de fichiers

Un système de fichier divise toujours l'espace de stockage disponible en un nombre de blocs. Souvent, la taille de ces blocs est la même que la taille des pages mémoires utilisées par le système, e.g., 4 Ko sur x86. Nous considérerons ici uniquement le cas de blocs de taille fixe. Un fichier occupe toujours un nombre entier de blocs, même si sa taille est inférieure à un multiple de la taille de ces blocs. Par exemple, un fichier de 100 octets occupera au moins un bloc (4 Ko = 4.096 octets) et un fichier de 16.500 octets occupera au moins 4 blocs (16Ko ou 16.536 octets). L'espace perdu est nommé la *fragmentation interne*, suivant la même définition que celle utilisée lors de la description des algorithmes de gestion dynamique de la mémoire dynamique.

On doit conserver, pour chaque fichier, la liste des blocs qu'il occupe. Des méta-données supplémentaires sont par ailleurs nécessaires pour permettre le contrôle d'accès (identifiant du propriétaire et du groupe du fichier, bits de permission, etc.) ou pour y collecter des métriques informatives à l'intention des utilisateurs et administrateurs du système (e.g., la date de la dernière écriture ou celle du dernier accès), comme nous l'avons vu au chapitre précédent.

Outre la liste des blocs de données associées à chaque fichier, il est nécessaire de maintenir la liste des blocs libres. Cette liste permet de réserver des blocs pour de nouveaux fichiers, ou pour étendre des fichiers existants.

Utilisation d'une table d'allocation des fichiers

Il existe deux grandes approches pour gérer le stockage des métadonnées et des listes de blocs. La première approche est représentée par les systèmes FAT (File Allocation Table) et ses successeurs (FAT32, ExFAT). Elle utilise une unique table d'allocation dédiée (d'où le nom FAT qui reprend cette notion de table) pour stocker les identifiants des blocs occupés par chaque fichier. Un fichier est identifié par un numéro unique, utilisé comme index dans cette table (on rappelle que l'association entre un nom de fichier en toute lettres, tel que manipulé par l'utilisateur ou un programme, est actée par l'existence d'une entrée dans un répertoire). L'entrée correspondant au numéro du fichier dans la table contient alors, outre les métadonnées, l'identifiant du *premier bloc* du fichier. On peut alors considérer deux approches pour stocker la liste des blocs associés à un fichier. Elles sont illustrées par la figure suivante.

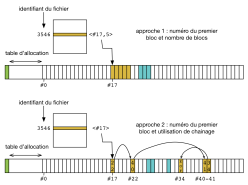


FIGURE 9.6 – Utilisation d'une table d'allocation et deux approches pour la conservation de la liste des blocs pour un fichier.

– Une **première approche** consiste à stocker dans l'entrée de la table correspondant au fichier l'identifiant du premier bloc et le nombre de blocs contigües occupés par le fichier. Outre sa simplicité, cette méthode a l'avantage de garantir que le fichier sera stocké sur des zones consécutives du disque. Particulièrement pour les disques dur (dispositifs mécaniques) cela permet d'assurer qu'un minimum de mouvements de la tête de lecture/écriture seront nécessaires pour lire le fichier de façon linéaire. Par contre, cette approche comporte deux désavantages :

– Tout d'abord, il n'est pas toujours possible d'augmenter la taille du fichier sans procéder à une coûteuse copie du fichier à un autre endroit du disque. Par exemple, on peut augmenter la taille du fichier jaune sur la partie haute de la figure de deux blocs seulement. Pour augmenter d'avantage sa taille, en revanche, il devient nécessaire de copier tous les blocs de ce fichier dans une zone libre différente sur le disque avant de pouvoir réaliser l'extension.

– Ensuite, cette approche tend à créer une importante fragmentation externe, c'est à dire de l'espace libre perdu entre des zones occupées, sans qu'il soit possible d'aménager la place nécessaire pour un grand fichier en une seule zone contigüe. Ce problème est similaire au problème de fragmentation externe rencontré par les gestionnaires de mémoire dynamique, et est illustré dans le chapitre correspondant.

– Une **seconde approche** est d'utiliser un *chaînage* entre les blocs formant le fichier. L'entrée de la table d'allocation pour un fichier ne contient alors que l'identifiant du premier bloc de ce fichier. Chaque bloc du fichier est utilisé pour stocker les données à l'exception d'une petite zone qui contient l'identifiant du prochain bloc (ou la valeur EOF, pour *end of file*, signifiant la fin du fichier). L'accès au fichier se fait en parcourant la liste chaînée de ses blocs. L'avantage de cette approche est que la taille du fichier n'est pas limitée par autre chose que l'espace disponible sur le disque. Par contre, outre sa plus grande complexité de mise en œuvre, elle présente elle aussi deux désavantages :

– Premièrement, il faut éviter de stocker les fichiers sur des blocs éparpillés sur le disque (non contigües), bien que le stockage chaîné de la liste des blocs le permette. En effet, l'éparpillement aura un impact important sur la performance, nécessitant des mouvements supplémentaires de la tête de lecture/écriture pour un disque dur et ne permettant pas de tirer partie de la grande granularité de lecture/écriture pour un disque dur ou un SSD. Il faut, en d'autres termes, privilégier la localité dans les allocations de blocs pour un même fichier.
– Deuxièmement, cette approche n'est pas adaptée pour des accès aléatoires au contenu des fichiers, c'est à dire à une adresse quelconque, car elle nécessite alors de lire et parcourir l'ensemble du fichier jusqu'au bloc désiré afin de suivre les informations de chaînage.

Les systèmes de fichiers de la famille FAT combinent en réalité ces deux approches. Les fichiers sont placés sur des zones (groupes de blocs) contigües et la table ne contient que l'identifiant du premier bloc. Toutefois, le dernier bloc de cette zone peut contenir un pointeur vers une nouvelle zone (un nouveau groupe de bloc) dans le cas où le fichier doit croître au delà de ce qui est possible en étendant le bloc existant. Cette approche hybride permet de combiner les avantages des deux approches discutées. Toutefois, elle ne règle pas le problème de la propension à la fragmentation externe et à l'éparpillement de ce type de système de fichiers : si un grand nombre de fichiers sont créés, supprimés, ou voient leur taille changer au cours du temps, et si le disque est fort rempli, alors les fichiers ont tendance à occuper de nombreuses petites zones éparpillées et la performance est sévèrement réduite.

Les systèmes FAT (et leurs successeurs comme NTFS) ont, outre leur simplicité, a leur avantage que l'espace utilisé pour stocker les métadonnées (la table d'allocation) est réduit au minimum nécessaire. Le reste du disque peut être utilisé pour stocker les données elles-mêmes.

Note : Utilisation d'un défragmenteur

Au contraire de la mémoire dynamique, il est possible d'agir pour réduire la fragmentation externe d'un système de fichier et pour augmenter la localité (i.e., le fait que les blocs pour un même fichier soient le plus possible contigües sur le disque).

Un appel à malloc(3) renvoie une adresse en mémoire qui est utilisée ensuite par l'application. Une fois que l'appel à malloc(3) a renvoyé cette adresse, il n'est plus possible de la modifier. L'algorithme de gestion de mémoire dynamique ne peut donc plus agir a posteriori pour diminuer la fragmentation externe, c'est à dire récupérer de l'espace perdu sous forme de "trous" entre des zones allouées mais non libérées.

L'association entre un numéro de fichier (dans la table d'allocation) et le placement sur le disque n'est jamais exposé directement aux applications, et n'a donc pas besoin d'être définitif. Il est donc tout à fait possible de modifier le placement des blocs du fichier dynamiquement, sans que ce changement ne soit visible par le reste du système d'exploitation et par les applications. Les systèmes d'exploitation de la famille Windows incluent ainsi un utilitaire système appelé le "défragmenteur" (*defrag* en anglais), s'appliquant aux systèmes de fichiers FAT et NTFS. Son objectif est d'appliquer un algorithme d'optimisation, regroupant les fichiers en des zones de blocs contigües uniques par des opérations de déplacement, augmentant ainsi leur localité et la performance de leurs accès. Cet utilitaire récupére par ailleurs l'espace perdu par la fragmentation externe en groupant ces zones en une zone unique, sans espace vide intermédiaire. Un nom plus exact pour cet outil serait donc le "rapprocheur/défragmenteur" ... On notera que l'utilisation d'un tel outil est rarement nécessaire pour les systèmes de fichiers utilisés sous UNIX comme ext4, qui prennent des mesures pour corriger le problème de façon dynamique lors des opérations d'écriture.

Stockage indexé

Une deuxième approche pour stocker la liste des blocs occupés par un fichier est d'utiliser directement un bloc complet comme *bloc d'index* vers des blocs de données. Ainsi, il n'y a plus de spécialisation de zone du disque pour stocker d'un côté les métadonnées et de l'autre les données. Cette approche est illustrée par la figure suivante.

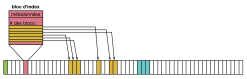


FIGURE 9.7 – Principe de l'indexation des blocs d'un fichier dans un *inode* occupant lui même un des blocs du disque.

Dans cet exemple, le bloc rose sert d'index pour le fichier. Il contient directement les métadonnées (propriétaire du fichier, groupe, etc.) et une liste de numéros des blocs jaunes formant le contenu du fichier.

Un avantage de cette approche est qu'il n'est pas nécessaire de limiter à l'avance le nombre maximal de fichiers en choisissant une taille pour la table d'indexation : il est possible de stocker des nouveaux fichiers tant qu'il existe au moins deux blocs libres, un pour les bloc d'index et un pour les données (voir aucun si le fichier est vide, e.g. il a été créé en utilisant touch(1)).

Un désavantage est que l'espace nécessaire pour tout fichier est toujours augmenté d'un bloc (par exemple de 4 Ko) ce qui est largement plus volumineux qu'une entrée dans une table d'indexation. Un fichier même très petit occupera donc au moins deux blocs : un fichier de 18 octets contenant uniquement la chaîne "Bonjour LINFO1252" occuperait ainsi deux blocs de 4 Ko sur le disque, soit un total de 8 Ko.

L'utilisation de blocs d'index indexant individuellement chaque bloc de données résout les problèmes de fragmentation externe, chaque bloc pouvant être utilisé même s'il est isolé, mais comme pour l'indexation avec une liste chaînée, peut amener à un éparpillement des blocs en particulier lorsque le disque est très rempli.

L'utilisation d'un bloc d'index unique a par ailleurs pour effet de limiter la taille maximale d'un fichier. Le nombre d'identifiants de blocs de donnée que l'on peut stocker dans ce bloc d'index est effectivement limité par la taille d'un bloc. Par exemple, si on sait stocker 800 index vers des blocs de données dans le bloc d'index, à la suite des métadonnées, alors la taille maximale d'un fichier sera de 800x 4 Ko soit un peu plus de 3 Mo.

Système de fichiers ext4

Le système de fichiers ext4 est le plus couramment utilisé sous Linux. Ce système de fichiers utilise une approche hybride entre les solutions discutées précédemment. Des blocs d'index appelés inodes contiennent les métadonnées associées aux fichiers ainsi que la liste des blocs de données. Une partition est scindée en groupes de blocs. Dans chacun de ces groupes, une zone est réservée pour stocker les inodes. Le reste est formé de blocs de données. L'avantage de la scission en groupes de blocs est de ne pas stocker trop loin sur le disque les métadonnées et les données correspondantes, pour ainsi éviter des va-et-vient trop important des têtes de lecture/écriture. Le stockage des inodes dans une zone spéciale permet de les limiter à une taille fixe plus réduite que celle d'un bloc complet. Le système ext4 supporte, par ailleurs, des tailles de blocs de données variables. Une description complète de ext4 sort du contexte de ce cours, mais la manière dont celui-ci résout le problème de la limitation du nombre d'entrées vers des blocs de données dans un inode est intéressante à étudier.

Dans ext4, et dans d'autres systèmes de fichiers pour UNIX avant lui, un inode contient un nombre limité de liens directs vers des blocs de données. Une configuration standard est de 12 liens "directs" de ce type. En utilisant seulement ces liens directs, et avec une taille de bloc de 4 Ko, cela indique qu'un fichier peut être d'une taille maximum de 12x4=48 Ko. La structure de l'inode permet d'utiliser des liens supplémentaires et donc des fichiers plus volumineux en utilisant plusieurs niveaux d'indirections. On compte trois niveaux d'indirection (simple, double, et triple), comme illustré sur la figure suivante.

Le premier pointeur d'indirection pointe vers un bloc d'index supplémentaire (choisi parmi les blocs de données) dont le contenu sera des numéros de blocs de données formant le contenu du fichier (premier niveau d'indirection). Si un bloc de données a une taille de 4 Ko et que chaque numéro de bloc occupe 4 octets, alors un total de 4 Mo (1.024 blocs de 4 Ko) pourra être indexé par ce bloc, en plus des 48 Ko indexés par les liens directs. Avec deux

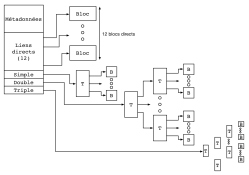


FIGURE 9.8 – Différents niveaux d'indirections pour supporter des grands fichiers à partir d'un inode de taille fixe.

niveaux d'indirection, le pointeur dans l'inode pointe vers un bloc contenant lui même les identifiants de blocs d'index. En utilisant les mêmes paramètres, le fichier peut contenir 4 Go de données supplémentaires. Le même principe est applicable avec un troisième niveau d'indirection, permettant d'atteindre une taille de fichier maximale de 4 To + 4 Go + 4 Mo + 48 Ko.

Stockage des répertoires

Un répertoire dans un système de fichier comme ext4 est stocké de la même manière qu'un fichier, à ceci près qu'un indicateur (le flag `d` dans les métadonnées) est mis à vrai, et que le bloc de données associés à l'inode comprendra alors une liste de structures de données `dirent`. Ces structures contiennent l'association entre des noms de fichiers et sous-répertoires et les inodes correspondants. Lorsque le nombre d'entrées `dirent` dans le répertoire dépasse la capacité d'un bloc, des blocs accessibles via les niveaux d'indirection sont utilisés, tout comme pour les fichiers. On notera que ext4 utilise une optimisation qui est de stocker directement dans l'inode du répertoire les entrées dirent lorsque leur nombre est très petit, évitant ainsi d'utiliser un bloc pour peu de données.

Gestion de l'espace libre

Il est nécessaire de conserver la liste des blocs disponibles afin de pouvoir rapidement réserver de l'espace pour la création d'un nouveau fichier ou l'accroissement de la taille d'un fichier existant. Bien entendu, cette information pourrait être retrouvée en passant en revue l'ensemble des inodes valides (i.e., accessibles depuis une entrée d'un répertoire lui même accessible depuis la racine du système de fichier) mais effectuer cette opération de recherche à chaque montage du système de fichier pour créer la structure de donnée correspondante en mémoire aurait un coût prohibitif. De plus, la taille de la structure de donnée résultante peut vite être très importante et occuper beaucoup d'espace. Avec un disque de 4 To et des blocs de 4 Ko, on peut estimer le nombre de bits nécessaires (si chaque bit représente si un bloc est ou non disponible) à 1.073.741.824, soit 128 Mo. Cette structure de donnée (un champ de bit) est donc stockée sur le disque lui-même. Dans le système ext4, chaque groupe de bloc inclue une zone réservée pour stocker ce "bitmap" des blocs libres.

Bien entendu, différents algorithmes existent pour choisir la zone la plus adéquate pour créer un nouveau fichier, en évitant si possible la fragmentation et en essayant de maximiser la localité. Leur description dépasse le contexte de ce cours, mais on retrouve des similarités entre ces algorithmes et ceux utilisés pour la gestion dynamique de la mémoire que nous avons abordé précédemment.

Note : Mon fichier est-il vraiment effacé ?

L'effacement d'un fichier avec la commande rm(1) consiste simplement en l'effacement de l'inode qui pointe vers ses blocs de contenu. Les blocs de contenu sont alors déclarés comme libres et pourront être réutilisés pour la création de nouveaux fichiers. Sauf précaution particulière, le contenu des blocs de données n'est pas modifié.

Des logiciels spécifiques permettent de passer en revue l'ensemble des blocs de données libres pour détecter des fichiers complets effacés, qui n'auraient pas encore été recouverts par le contenu de nouveaux fichiers. Ils se basent sur cela, entre autres, sur l'analyse des caractéristiques de fichiers classiques (fichiers d'images, vidéos, etc.). Ils sont utilisés par exemple par les forces de police pour recouvrer des preuves que des criminels n'ayant pas suivi LINFO1252 auraient tenté d'effacer avec un simple rm(1). Des administrateurs systèmes peuvent aussi utiliser des outils comme `extundelete` sur une version montée en lecture seule d'un système de fichier, pour recouvrer des fichiers effacés par mégarde (il n'y a pas de notion de corbeille en ligne de commande, contrairement à ce que l'on peut trouver dans un environnement graphique).

Si on souhaite effacer un fichier de façon permanente, c'est à dire en modifiant ses blocs de données plusieurs fois en y écrivant des données aléatoires (ou bien des 0 partout), il est possible d'utiliser l'utilitaire shred(1). Celui-ci permet de spécifier le nombre de passes d'écritures souhaitées sur les données. En effet, dans certains cas une seule passe n'est pas suffisante, en tout cas sur un disque dur. Avec du matériel spécialisé, il est possible de retrouver avec un probabilité qui décroit au fur et à mesure des écritures ultérieures, la polarisation passée d'un bit stocké sur le disque. L'exemple ci-dessous montre l'utilisation de shred(1) pour effacer définitivement un fichier sensible. L'option -v permet d'obtenir une sortie "verbeuse" détaillant les étapes des opérations. L'option -n permet de spécifier le nombre de passes d'effacement à effectuer. On voit ici que shred(1) alterne entre l'écriture de valeurs aléatoires, de 0, et de 1. L'option -z permet de demander l'écriture de 0 à la fin, pour rendre moins détectable l'opération d'effacement. On voit par ailleurs que l'utilitaire renomme de nombreuses fois le fichier, afin de faire disparaître l'entrée du contenu du répertoire.

```
$ cat ma_carte_de_credit
Linus Torvalds
Aktia Savings Bank
1234 5678 9123 4567
EXP 01/21
CRC 123
$ shred -vzu -n 5 ma_carte_de_credit
shred: ma_carte_de_credit: pass 1/6 (random)...
shred: ma_carte_de_credit: pass 2/6 (000000)...
shred: ma_carte_de_credit: pass 3/6 (random)...
shred: ma_carte_de_credit: pass 4/6 (ffffff)...
shred: ma_carte_de_credit: pass 5/6 (random)...
shred: ma_carte_de_credit: pass 6/6 (000000)...
shred: ma_carte_de_credit: removing
shred: ma_carte_de_credit: renamed to 000000000000000000
shred: 000000000000000000: renamed to 00000000000000000
shred: 00000000000000000: renamed to 0000000000000000
shred: 0000000000000000: renamed to 000000000000000
shred: 000000000000000: renamed to 00000000000000
shred: 00000000000000: renamed to 0000000000000
shred: 0000000000000: renamed to 000000000000
shred: 000000000000: renamed to 00000000000
shred: 00000000000: renamed to 0000000000
shred: 0000000000: renamed to 000000000
shred: 000000000: renamed to 00000000
shred: 00000000: renamed to 0000000
shred: 0000000: renamed to 000000
shred: 000000: renamed to 00000
shred: 00000: renamed to 0000
shred: 0000: renamed to 000
shred: 000: renamed to 00
shred: 00: renamed to 0
shred: ma_carte_de_credit: removed
```

9.6.3 Performance des systèmes de fichiers

Les accès aux périphériques de stockage sont particulièrement lents, même en utilisant des technologies SSD et des contrôleurs de périphériques de dernière génération. À titre d'exemple, les latences d'accès à la mémoire principales se comptent en dizaines ou centaines de nano-secondes, tandis que la latence d'accès à un SSD connecté avec un contrôleur de périphérique à la norme NVMe (la plus rapide disponible hors serveurs haute performance) est plutôt de l'ordre de quelques dizaines ou centaines de micro-secondes, soit un rapport de un à mille. La différence en bande passante, elle, est moins importante mais reste d'un facteur de 10 à 20 entre les mémoires un les SSD les plus performants. Un disque dur classique présente quand à lui des latences d'accès de quelques millisecondes, et une bande passante environ 5 à 10 fois moins mois élevée que celle d'un SSD.

L'utilisation du principe de cache permet d'augmenter sensiblement la performance des systèmes de fichier. On retrouve des caches à plusieurs niveaux :

– Tout d'abord, les périphériques de stockage eux-même (et/ou les contrôleurs de périphériques) disposent souvent d'un cache permettant de stocker un petit nombre d'opérations d'écriture en attente, et donc de diminuer la latence de ces opérations du point de vue du système d'exploitation.
– Ensuite, le système d'exploitation utilise une partie de la mémoire pour servir de cache pour les blocs lus et écrits par le système de fichiers. Sous Linux, l'ensemble des pages qui ne sont pas autrement utilisées par les applications sont inclues dans ce "Page Cache" (ou *disk cache*). Lors de la lecture d'un bloc ou d'un inode par le système de fichier, son contenu est ajouté à une page libre du page cache (si nécessaire, une page ancienne est évincée par l'algorithme de remplacement de page). Les accès en lecture suivants se font alors dans le cache. Les accès en écriture se font eux aussi dans le cache, et son répercutés lors de l'éviction de la page de la mémoire, ou lorsque le processus aura utilisé l'appel système fsync(2).

L'utilisation du page cache facilite la mise en œuvre du mapping des fichiers en mémoire partagé. Les pages correspondant au fichier mappé sont marquées dans la table des pages du processus ayant appelé l'appel système mmap(2), mais ne seront pas rappatriées directement en mémoire physique. Cela aurait peu d'intérêt, en effet, si le processus n'accède *in fine* qu'à un sous-ensemble du fichier. L'accès à ces pages provoquera des défauts de page qui seront servis en lisant le contenu du fichier au fur et à mesure de son utilisation.

Le page cache permet de mettre en œuvre des optimisations de performance, au delà de pouvoir servir les requêtes pour des données récemment lues sans devoir les relire depuis le disque :

– Le positionnement de la tête de lecture/écriture est généralement une opération beaucoup plus longue que la lecture elle même. Les accès aux fichiers se font par ailleurs le plus souvent de façon linéaire. Il est donc bénéfique de profiter du positionnement de la tête de lecture sur la même piste pour lire plusieurs secteurs en une seule opération. Ainsi, si le bloc de numéro 125 est lu, les blocs 126, 127, etc. seront lus en même temps et l'accès se fera alors directement depuis le page cache. Cette stratégie dite de pré-chargement (*prefetching* ou *read-ahead*) est très bénéfique en particulier lorsque la localité des fichiers sur le disque est élevée.
– Il est très commun que la lecture d'un fichier se fasse de façon linéaire, sans jamais revenir en arrière pour relire des données. Dans ce cas, il est possible et même souhaitable d'évincer les copies de pages dont la lecture est complète dès le chargement de la prochaine page (ou ensemble de pages avec le préchargement). Cette stratégie dite *free-behind* permet d'éviter que des pages qui ne seront plus jamais accédées mais l'ayant été récemment prennent la place de pages plus anciennes mais de plus grande importance.

Note : Le scheduling des accès à un disque dur

Il existe de nombreuses optimisations à différents niveaux de la mise en œuvre des systèmes de fichiers. L'une d'entre elles est l'utilisation d'un scheduling optimisé au niveau du traitement des commandes par le disque dur lui même. Lorsque plusieurs demandes d'accès au disque pour lire ou écrire des blocs sont effectuées de manière concurrente (par différents threads et/ou différents processus) il est rarement optimal de servir ces commandes dans leur ordre d'arrivée : souvent, les blocs seront à des positions éloignées du disque, sur des plateaux différents, ou sur des faces différentes du même plateau. L'opération de déplacement de la tête de lecture écriture, ou le changement de côté du plateau accédé, sont des opérations coûteuses en latence. On parle en anglais de "seek latency". Il est préférable de ré-ordonner les accès de façon à ce que les mouvements mécaniques du disque soient minimisés. On peut faire un parallèle avec les algorithmes utilisés pour les ascenseurs de très grands immeubles : ceux-ci ne répondent pas aux sollicitations dans l'ordre de l'appui sur les boutons mais bien en cherchant à maximimiser le nombre de personnes transportées, potentiellement en ne respectant pas de mesure d'équité – certains attendent alors plus que d'autres.

Robustesse et vérificateurs de systèmes de fichiers

Contrairement à la mémoire principale dont le contenu est effacé lors de l'arrêt de la machine, ce n'est pas le cas d'un périphérique de stockage dont le contenu doit pouvoir être monté de nouveau au prochain démarrage. Il faut évidemment qu'il n'y ait pas d'incohérence entre les différentes informations conservées sur le disque, comme le contenu des inodes, l'association entre les inodes et les blocs de contenu, et surtout la bitmap indiquant les blocs libres. Un scénario particulièrement dommageable est qu'un bloc soit indiqué comme libre alors qu'il est effectivement utilisé par un fichier. Ce bloc pourrait alors être alloué à deux fichiers différents, cassant les propriétés d'isolation (un processus n'ayant pas les droits sur le deuxième fichiers mais sur le premier pouvant voir du contenu appartenant du premier fichier) et des corruptions de données (le contenu d'un fichier écrasant le contenu de l'autre). Moins grave, mais pénalisant sur le long terme, un bloc indiqué comme occupé mais qui ne l'est pas ne sera jamais libéré et la capacité disponible du disque s'en trouvera ainsi réduite.

Il arrive qu'un arrêt brutal de la machine ne permette pas le *démontage* du système de fichier proprement. Or, l'utilisation du cache et le fait que les écritures sur le disques puissent ne pas être complètement répercutées peut entraîner des incohérences de l'état stocké sur ce disque. Des utilitaires système spéciaux, les vérificateurs de systèmes de fichiers, permettent de vérifier le contenu d'une partition avant son montage pour repérer et souvent, corriger, les erreurs rencontrées. Ils effectuent de nombreuses vérifications, dont un exemple est de reconstruire en mémoire le champs de bit correspondant aux blocs libres et de comparer celui-ci aux blocs effectivement liés par des fichiers. Un bloc marqué comme libre alors qu'il ne l'est en réalité par sera marqué comme tel, et un bloc effectivement libre sera ajouté au compte de l'espace disponible sur la partition. Sous Linux, l'utilitaire fsck(8) est le vérificateur pour les systèmes de fichiers de la famille ext, comme ext4.

Systèmes de fichiers journalisés

Le système de fichier ext4, à l'image de la plupart des systèmes de fichiers modernes, utilise le principe de la journalisation. Les opérations d'écriture sur le disque ne sont pas réalisées directement là où un bloc est stocké. À la place, une partie du disque est utilisée pour y écrire, dans l'ordre de leur arrivée, les modifications. Chaque modification est donc vue comme une transaction, et l'ensemble des transactions forme un journal. Les transactions peuvent être écrite de façon rapide sur le disque, et comme elles sont successives maximiser leur localité (i.e. elles vont être écrites les unes à côté des autres sur le disque). La propagation des changements vers les blocs eux-même peut ensuite être réalisée de façon paresseuse, avec une priorité moindre sur les accès directs. L'utilisation d'un système de fichiers journalisé améliore la performance mais aussi la robustesse. Si le disque est démonté brutalement ou perd son alimentation, il est possible de passer en revue le journal des transactions et de les appliquer de nouveau pour restaurer un état cohérent du système.

9.7 Annexes

9.7.1 Bibliographie

9.7.2 Glossaire

/proc Sous Linux, représentation de l'information stockée dans la *table des processus* sous la forme d'une arborescence directement accessible via les commandes du *shell*. Voir proc(5)

adresse Position d'un donnée en mémoire.

AND, conjonction logique Opération binaire logique.

Andrew Tanenbaum Andrew Tanenbaum est professeur à la VU d'Amsterdam.

appel système Fonction primitive fournie par le noyau du système d'exploitation et pouvant être appelée directement par les programmes applicatifs.

appel système bloquant Appel système qui ne retourne par de résultat immédiat. Dans ce cas, le noyau du système d'exploitation sélectionne un autre processus via le *scheduler* en attendant que le résultat de l'appel système soit disponible.

appel système lent Un appel système lent est un appel système qui peut attendre un temps indéfini pour se terminer. Par exemple, l'appel read(2) sur l'entrée standard ne retournera de résultat que lorsque l'utilisateur aura pressé une touche sur le clavier.

Application Programming Interface, API Un API est généralement un ensemble de fonctions et de structures de données qui constitue l'interface entre deux composants logiciels qui doivent collaborer. Par exemple, l'API du noyau d'un système Unix est composée de ses appels systèmes. Ceux-ci sont décrits dans la section 2 des pages de manuel (voir intro(2)).

Aqua Aqua est une interface graphique spécifique à *MacOS*. Voir https ://en.wikipedia.org/wiki/Aqua_(user_interface)

assembleur Programme permettant de convertir un programme écrit en langage d'assemblage dans le langage machine correspondant à un processeur donné.

benchmark Ensemble de programmes permettant d'évaluer les performances d'un système informatique.

big endian Ordre dans lequel les octets correspondants à des mots de plusieurs octets sont stockés en mémoire. Voir https ://fr.wikipedia.org/wiki/Boutisme#Little_endian

bit Plus petite unité d'information. Par convention, un bit peut prendre les valeurs `0` et `1`.

bit de poids faible Par convention, bit le plus à droite d'une séquence de n bits.

bit de poids fort Par convention, le bit le plus à gauche d'une séquence de n bits.

BSD Unix Variante de Unix développée à l'Université de Californie à Berkeley.

buffer overflow Erreur dans laquelle un programme informatique cherche à stocker plus de données en mémoire que la capacité de la zone réservée en mémoire. Donne généralement lieu à des problèmes, parfois graves, de sécurité. https ://en.wikipedia.org/wiki/Buffer_overflow

byte, octet Un bloc de huit bits consécutifs.

C Langage de programmation permettant d'interagir facilement avec le matériel.

C99 Standard international définissant le langage C [C99]

changement de contexte Passage de l'exécution du programme A au programme B.

CISC Complex Instruction Set Computer

contexte Structure de données maintenue pas le noyau du système d'exploitation qui contient toutes les informations nécessaires pour poursuivre l'exécution d'un programme.

cpp, préprocesseur Le préprocesseur C est un programme de manipulation de texte sur base de macros qui est utilisé avec le compilateur. Le préprocesseur de *gcc* est http ://gcc.gnu.org/onlinedocs/cpp/

CPU Central Processing Unit

CPU Central Processing Unit. Voir *microprocesseur*

deadlock Voir https ://en.wikipedia.org/wiki/Deadlock

debugger Logiciel

descripteur de fichier Identifiant (entier) retourné par le noyau du système d'exploitation lors de l'ouverture d'un fichier par l'appel système open(2).

DMA, Direct Memory Access Mécanisme permettant à un contrôleur de périphérique d'entrée/sortie d'écrire ou de lire directement depuis la mémoire principale en ne générant une *interruption* que lorsque le transfert est terminé.

double précision Représentation de nombre réels en virgule flottante (type `double` en C). La norme IEEE754 définit le format de ces nombres sur 64 bits.

DRAM, dynamic RAM Un des deux principaux types de mémoire. Dans une DRAM, l'information est mémorisée comme la présence ou l'absence de charge dans un minuscule condensateur. Les mémoires DRAM sont plus lentes que les *SRAM* mais ont une plus grande capacité.

eip, pc, compteur de programme, instruction pointer Registre spécial du processeur qui contient en permanence l'adresse de l'instruction en cours d'exécution. Le contenu de ce registre est incrémenté après chaque instruction et modifié par les instructions de saut.

errno Variable globale mise à jour par certains appels systèmes et fonctions de la librairie standard en cas d'erreur. Voir errno(3)

exclusion mutuelle Zone d'un programme multithreadé qui ne peut pas être exécutée par plus d'un thread à la fois.

fichier Une séquence composée d'un nombre entier d'octets stockée sur un dispositif de stockage. Un fichier est identifié par son nom et sa position dans l'arborescence du système de fichiers.

fichier header Fichier contenant des signatures de fonctions, des déclarations de types de données, des variables globales, permettant d'utiliser une librairie ou un API.

fichier objet Fichier résultat de la compilation d'une partie de programme. Ce fichier contient les instructions en langage machine à exécuter ainsi que les informations relatives aux différents symboles (variables, fonctions, ...) qui y sont définis.

FreeBSD Variante de BSD Unix disponible depuis http ://www.freebsd.org

FSF Free Software Foundation, http ://www.fsf.org

garbage collector Algorithme permettant de libérer la mémoire qui n'est plus utilisée notamment dans des langages tels que Java

gcc Compilateur pour la langage C développé par un groupe de volontaires qui est diffusé depuis http ://gcc.gnu.org gcc est utilisé dans plusieurs systèmes d'exploitation de type Unix, comme MacOS, Linux ou FreeBSD. Il existe d'autres compilateurs C. Une liste non-exhaustive est maintenue sur http ://en.wikipedia.org/wiki/List_of_compilers#C_compilers

GHz Mesure de fréquence en milliards de répétitions par seconde.

Gnome Environnement graphique utilisé par de nombreuses distributions Linux. Voir https ://en.wikipedia.org/wiki/GNOME

GNU is not Unix, GNU GNU est un projet open-source de la Free Software Foundation qui a permis le développement d'un grand nombre d'utilitaires utilisés par les systèmes d'exploitation de la famille Unix actuellement.

GNU/Linux Nom générique donné à un système d'exploitation utilisant les utilitaires *GNU* notamment et le noyau *Linux* .

heap, tas Partie de la mémoire d'un programme gérée par malloc(3) et free(3).

hiérarchie de mémoire Ensemble des mémoires utilisées sur un ordinateur. Depuis les registres jusqu'à la mémoire virtuelle en passant par la mémoire centrale et les mémoires caches.

inode structure de données utilisée par le système de fichiers Unix pour représenter un fichier/répertoire

interruption Signal extérieur (horloge, opération d'entrée/sortie, ...) qui force le processeur à arrêter l'exécution du programme en cours pour exécuter une routine du système d'exploitation et traiter l'interruption.

kHz Mesure de fréquence en milliers de répétitions par seconde.

libc Librairie C standard. Contient de nombreuses fonctions utilisables par les programmes écrits en langage C et décrites dans la troisième section des pages de manuel. Linux utilise la librairie GNU glibc qui contient de nombreuses extensions par rapport à la libraire standard.

lien symbolique Unix supporte deux types de liens. Les liens durs créés par ln(1) et les liens symboliques créés par ln(1) avec l'argument `-s`.

linker Editeur de liens. Partie du compilateur c permettant de combiner plusieurs fichiers objet en un exécutable.

Linus Torvalds Linus Torvalds est le créateur et le mainteneur principal du noyau *Linux*.

Linux Noyau de système d'exploitation compatible Unix développé initialement par Linus Torvalds.

little endian Ordre dans lequel les octets correspondants à des mots de plusieurs octets sont stockés en mémoire. Voir https ://fr.wikipedia.org/wiki/Boutisme#Little_endian

livelock Voir https ://en.wikipedia.org/wiki/Deadlock#Livelock

liveness, vivacité Propriété d'un programme informatique. Dans le problème de l'exclusion mutuelle, une propriété de vivacité est qu'un thread qui souhaite entrer en section critique finira par y accéder.

llvm Ensemble de compilateurs pour différents langages de programmation et différents processeurs développé par un groupe de volontaire. `llvm` est distribué depuis http ://llvm.org/

loi de Amdahl Voir https ://fr.wikipedia.org/wiki/Loi_d%27Amdahl

loi de Moore Voir https ://fr.wikipedia.org/wiki/Loi_de_Moore

MacOS Système d'exploitation développé par Apple Inc. comprenant de nombreux composantes provenant de *FreeBSD*.

makefile Fichier décrivant la façon dont make(1) doit compiler un programme.

memory leak Fuite de mémoire. Erreur concernant un programme qui a alloué de la mémoire avec malloc(3) et ne l'utilise plus sans avoir fait appel à free(3)

mémoire Dispositif électronique permettant de stocker de l'information

mémoire cache Mémoire rapide de faible capacité. La mémoire cache peut stocker des données provenant de mémoires de plus grande capacité mais qui sont plus lentes, et exploite le *principe de localité* en stockant de manière transparente les instructions et les données les plus récemment utilisées. Elle fait office d'interface entre le processeur et la mémoire principale et toutes les demandes d'accès à la mémoire principale passent par la mémoire cache, ce qui permet d'améliorer les performances de nombreux systèmes informatiques.

MHz Mesure de fréquence en millions de répétitions par seconde.

microprocesseur, processeur Unité centrale de l'ordinateur qui exécute les instructions en langage machine et interagit avec la mémoire.

Minix Famille de noyaux de systèmes d'exploitation inspiré de *Unix* développée notamment par *Andrew Tanenbaum*. Voir http ://www.minix3.org pour la dernière version de Minix.

MIPS Million d'instructions par seconde

MMU Unité de gestion de la mémoire, ou Memory Management Unit. Ce composant associé au processeur permet de gérer l'accès à la mémoire, particulièrement la traduction entre adresses virtuelles et adresses physiques et le contrôle d'accès.

multi-threadé, multi-coeurs Processeur contenant plusieurs unités permettant d'exécuter simultanément des instructions de programmes différents.

multitâche, multitasking Capacité d'exécuter plusieurs programmes simultanément.

multithreadé Programme utilisant plusieurs threads.

mutex Primitive de synchronisation permettant d'empêcher que deux threads accèdent simultanément à une même section critique.

nibble Un bloc de quatre bits consécutifs.

NOT, négation Opération binaire logique.

offset pointer Position de la tête de lecture associée à un fichier ouvert.

OpenBSD Variante de BSD Unix disponible depuis http ://www.openbsd.org

opération atomique Opération ne pouvant être interrompue.

OR, disjonction logique Opération binaire logique.

pid, process identifier identifiant de processus. Sous Unix, chaque processus est identifié par un entier unique. Cet identifiant sert de clé d'accès à la *table des processus*. Voir getpid(2) pour récupérer l'identifiant du processus courant.

pipe Mécanisme de redirection des entrées-sorties permettant de relier la sortie standard d'un programme à l'entrée standard d'un autre pour créer des pipelines de traitement.

pointeur Adresse d'une variable ou fonction en mémoire.

politique d'ordonnancement scheduling en anglais. Politique décidant dans quel ordre et sur quel processeur les processus disponibles à l'exécution doivent se voir allouer le ou les processeur(s) disponibles

portée Zone d'un programme dans laquelle une variable est déclarée.

portée globale Une variable ayant une portée globale est accessible dans tout le programme.

portée locale Une variable ayant une portée locale est accessible uniquement dans le bloc dans laquelle elle est définie.

principe de localité principe de fonctionnement de la mémoire indiquant que lorsqu'un programme accède à une adresse à un temps t, il accédera encore à des adresses proches dans les prochains instants

processus Ensemble cohérent d'instructions utilisant une partie de la mémoire, initié par le système d'exploitation et exécuté sur un des processeurs du système. Le système d'exploitation libère les ressources qui lui sont allouées à la fin de son exécution.

RAM, Random Access Memory Mémoire à accès aléatoire. Mémoire permettant au processeur d'accéder à n'importe quelle donnée en connaissant son adresse. Voir *DRAM* et *SRAM*.

raspberry pi Systèmes informatiques développés par la Raspberry Pi Foundation, voir https ://www.raspberrypi.org

registre Unité de mémoire intégrée au processeur. Les registres sont utilisés comme source ou destination pour la plupart des opérations effectuées par un processeur.

répertoire Branche de l'arborescence du système de fichiers. Un répertoire contient un ou plusieurs fichiers.

répertoire courant Répertoire dans lequel l'appel système open(2) cherchera à ouvrir les fichiers do

RISC Reduced Instruction Set Computer

root Racine de l'arborescence des fichiers mais aussi utilisateur ayant les privilèges les plus élevés sur un ordinateur utilisant Unix.

round-robin Voir https ://fr.wikipedia.org/wiki/Round-robin_(informatique)

scheduler Ordonnanceur. Algorithme utilisé par le noyau du système d'exploitation pour sélectionner le prochain programme à exécuter après une interruption d'horloge ou un appel système bloquant.

section critique Partie de programme ne pouvant pas être exécutée simultanément par deux threads différents.

segment de données Partie de la mémoire comprenant les segments des données initialisées et non-initialisées

segment des données initialisées Partie de la mémoire d'un programme contenant les données initialisées dans le code source du programme ainsi que les chaînes de caractères.

segment des données non-initialisées Partie de la mémoire d'un programme contenant les données (tableaux notamment) qui sont déclarés mais pas explicitement initialisés dans le code source du programme.

segmentation fault Erreur à l'exécution causée par un accès à une adresse mémoire non-autorisée pour le programme.

sémaphore Primitive de synchronisation permettant notamment l'exclusion mutuelle. Voir notamment [Downey2008]

shared library, librairie dynamique, librairie partagée Lorsqu'un librairie est dynamiquement liée à un programme exécutable, le code de celui-ci ne contient pas les instructions de la librairie, mais celle-ci est automatiquement chargée lors de chaque exécution du programme. Cela permet d'avoir une seule copie de chaque librairie. C'est la solution utilisée par défaut sous Linux.

shell Interpréteur de commandes sur un système Unix. bash(1) est l'interpréteur de commandes le plus utilisé de nos jours.

signal mécanisme permettant la communication entre processus. Utilisé notamment pour arrêter un processus via la commande kill(1)

simple précision Représentation de nombre réels en virgule flottante (type `float` en C). La norme IEEE754 définit le format de ces nombres sur 32 bits.

Solaris Système d'exploitation compatible Unix développé par Sun Microsystems et repris par Oracle. La version open-source, OpenSolaris, est disponible depuis http ://www.opensolaris.org

SRAM, static RAM Un des deux principaux types de mémoire. Dans une SRAM, l'information est mémorisée comme la présence ou l'absence d'un courant électrique. Les mémoires SRAM sont généralement assez rapides mais de faible capacité. Elles sont souvent utilisées pour construire des mémoires caches.

SSD, Solid State Drive Système de stockage de données s'appuyant uniquement sur de la mémoire flash.

stack, pile Partie de la mémoire d'un programme contenant les variables locales et adresses de retour des fonctions durant leur exécution.

static library, librairie statique Une librairie est statiquement liée à un programme exécutable lorsque tout son code est intégré dans l'exécutable. Voir les arguments `static` dans gcc(1)

stderr Sortie d'erreur standard sur un système Unix (par défaut l'écran)

stdin Entrée standard sur un système Unix (par défaut le clavier)

stdout Sortie standard sur un système Unix (par défaut l'écran)

sûreté, safety Propriété d'un programme informatique. Dans le problème de l'exclusion mutuelle, une propriété de sûreté est que deux threads ne seront jamais dans la même section critique.

table des pages Table contenant l'association entre les pages d'un espace de mémoire virtuelle et les frames (ou pages physiques) de la mémoire physique auxquelles elles sont allouées. Le contenu de la table des pages est mis à jour par le système d'exploitation et utilisé par la MMU pour traduire automatiquement les adresses virtuelles en adresses physiques.

table des processus Table contenant les identifiants (*pid*) de tous les processus qui s'exécutent à ce moment sur un système Unix. Outre les identifiants, cette table contient de nombreuses informations relatives à chaque *processus*. Voir également */proc*

text, segment text Partie de la mémoire d'un programme contenant les instructions en langage machine à exécuter.

thread-safe Une fonction est dite thread-safe si elle peut être simultanément exécutée sans contrainte par différents threads d'un même programme.

trap interruption logicielle permettant à un processus en mode utilisateur de donner la main au noyau en mode protégé, soit dans le cas d'un appel système soit lors de l'utilisation d'une instruction interdite

Unicode Norme d'encodage de caractères supportant l'ensemble des langues écrites, voir notamment https ://en.wikipedia.org/wiki/Unicode

Unix Système d'exploitation développé initialement par AT&T Bell Labs.

userid Identifiant d'utilisateur. Sous Unix, un entier unique est associé à chaque utilisateur.

von Neumann Un des inventaires des premiers ordinateurs. A défini l'architecture de base des premiers ordinateurs qui est maintenant connue comme le modèle de von Neumann [Krakowiak2011]

warning Message d'avertissement émis par un compilateur C. Un *warning* n'empêche pas la compilation et la génération du code objet. Cependant, la plupart des warnings indiquent un problème dans le programme compilé et il est nettement préférable de les supprimer du code.

X11 Interface graphique développée au MIT pour Unix. Voir https ://en.wikipedia.org/wiki/X_Window_System

x86 Famille de microprocesseurs développée par intel. Le 8086 est le premier processeur de cette famille. Ses successeurs (286, 386, Pentium, Centrino, Xeon, ...) sont restés compatibles avec lui tout en introduisant chacun de nouvelles instructions et de nouvelles fonctionnalités. Aujourd'hui, plusieurs fabricants développent des processeurs qui supportent le même langage machine que les processeurs de cette famille.

XOR, ou exclusif Opération binaire logique.

Bibliographie

[ABS] Cooper, M., *Advanced Bash-Scripting Guide*, 2011, http ://tldp.org/LDP/abs/html/

[AdelsteinLubanovic2007] Adelstein, T., Lubanovic, B., *Linux System Administration*, OReilly, 2007, http ://books.google.be/books ?id=-jYe2k1p5tIC

[Bashar1997] Bashar, N., *Ariane 5 : Who Dunnit ?*, IEEE Software 14(3) : 15–16. May 1997. https ://doi.ieeecomputersociety.org/10.1109/MS.1997.589224

[C99] http ://www.open-std.org/jtc1/sc22/wg14/www/docs/n1256.pdf

[Cooper2011] Cooper, M., *Advanced Bash-Scripting Guide*, http ://tldp.org/LDP/abs/html/, 2011

[CPP] C preprocessor manual, http ://gcc.gnu.org/onlinedocs/cpp/

[Dijkstra1965b] Dijkstra, E., *Cooperating sequential processes*, 1965, http ://www.cs.utexas.edu/users/EWD/transcriptions/EWD01

[Dijkstra1965] Dijkstra, E., *Solution of a problem in concurrent programming control*. Commun. ACM 8, 9 (September 1965), 569 http ://doi.acm.org/10.1145/365559.365617

[Downey2008] Downey, A., *The Little Book of Semaphores*, Second Edition, Green Tea Press, 2008, https ://greenteapress.com/wp/semaphores/

[DrepperMolnar2005] Drepper, U., Molnar, I., *The Native POSIX Thread Library for Linux*, http ://www.akkadia.org/drepper/nptl-design.pdf

[Goldberg1991] Goldberg, D., *What every computer scientist should know about floating-point arithmetic*. ACM Comput. Surv. 23, 1 (March 1991), 5-48. http ://doi.acm.org/10.1145/103162.103163 ou http ://www.validlab.com/goldberg/paper.pdf

[Gove2011] Gove, D., *Multicore Application Programming for Windows, Linux and Oracle Solaris*, Addison-Wesley, 2011, http ://books.google.be/books ?id=NF-C2ZQZXekC

[GNUPTH] Engelschall, R., *GNU Portable Threads*, http ://www.gnu.org/software/pth/

[IA32] intel, *Intel® 64 and IA-32 Architectures : Software Developer's Manual*, Combined Volumes : 1, 2A, 2B, 2C, 3A, 3B and 3C, December 2011, http ://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-manual-325462.pdf

[Kamp2011] Kamp, P., *The Most Expensive One-byte Mistake*, ACM Queue, July 2011, http ://queue.acm.org/detail.cfm ?id=2010365

[Kerrisk2010] Kerrisk, M., *The Linux Programming Interface*, No Starch Press, 2010, http ://my.safaribooksonline.com/book/programming/linux/9781593272203

[Kernighan] Kernighan, B., *Programming in C - A Tutorial*, http ://cm.bell-labs.com/cm/cs/who/dmr/ctut.pdf

[KernighanRitchie1998] Kernighan, B., and Ritchie, D., *The C programming language, second edition*, Addison Wesley, 1998, http ://cm.bell-labs.com/cm/cs/cbook/

[King2008] King, K., *C programming : a modern approach*, W. W. Norton & company, 2008

[Krakowiak2011] Krakowiak, S., *Le modele d'architecture de von Neumann*, http ://interstices.info/le-modele-darchitecture-de-von-neumann

[Leroy] Leroy, X., *The LinuxThreads library*, http ://pauillac.inria.fr/~xleroy/linuxthreads/

[McKenney2005] McKenney, P., *Memory Ordering in Modern Microprocessors, Part I*, Linux Journal, August 2005, http ://www.linuxjournal.com/article/8211

[Mitchell+2001] Mitchell, M., Oldham, J. and Samuel, A., *Advanced Linux Programming*, New Riders Publishing, ISBN 0-7357-1043-0, June 2001, http ://www.advancedlinuxprogramming.com/

[Nemeth+2010] Nemeth, E., Hein, T., Snyder, G., Whaley, B., *Unix and Linux System Administration Handbook*, Prentice Hall, 2010, http ://books.google.be/books ?id=rgFIAnLjb1wC

[Stallings2011] Stallings, W., *Operating Systems : Internals and Design Principles*, Prentice Hall, 2011, http ://williamstallings.com/OperatingSystems/index.html

[StevensRago2008] Stevens, R., and Rago, S., *Advanced Programming in the UNIX Environment*, Addison-Wesley, 2008, http ://books.google.be/books ?id=wHI8PgAACAAJ

[Stokes2008] Stokes, J., *Classic.Ars : Understanding Moore's Law*, http ://arstechnica.com/hardware/news/2008/09/moore.ars

[Tanenbaum+2009] Tanenbaum, A., Woodhull, A., *Operating systems : design and implementation*, Prentice Hall, 2009, https ://www.pearson.com/us/higher-education/program/Tanenbaum-Operating-Systems-Design-and-Implementation-3rd-Edition/PGM228096.html

[Walls2006] Walls, D., *How to Use the restrict Qualifier in C*. Sun Microsystems, 2006, http ://developers.sun.com/solaris/articles/cc_restrict.html