Chapitre 1 INTRODUCTION

Les logiciels peuvent être classés en deux catégories :
- les programmes d'application des utilisateurs
- les programmes système qui permettent le fonctionnement de l'ordinateur. Parmi ceux-ci, le système d'exploitation (SE dans la suite).

Le SE soustrait le matériel au regard du programmeur et offre une présentation agréable des fichiers. Un SE a ainsi deux objectifs principaux :

- présentation :Il propose à l'utilisateur une abstraction plus simple et plus agréable que le matériel : une machine virtuelle

- gestion : il ordonne et contrôle l'allocation des processeurs, des mémoires, des icônes et fenêtres, des périphériques, des réseaux entre les programmes qui les utilisent. Il assiste les programmes utilisateurs. Il protège les utilisateurs dans le cas d'usage partagé.

1. HISTORIQUE

Les premiers ordinateurs étaient mis à la disposition d'un programmeur selon un calendrier de réservation : un usager avec un travail unique utilisait seul la machine à un moment donné. Puis vint l'époque du traitement par lots (batch) : enchaînement, sous le contrôle d'un moniteur, d'une suite de travaux avec leurs données, confiés à l'équipe d'exploitation de la machine (inconvénient : temps d'attente des résultats pour chaque utilisateur).

Cette pratique a nécessité trois innovations :
- le contrôle des E/S et leur protection pour éviter le blocage d'un lot
- un mécanisme de comptage de temps et de déroutement autoritaire des programmes pour éviter le blocage d'un lot à cause d'une séquence trop longue. Ce furent les premières interruptions
- les premiers langages de commande (JCL) sous forme de cartes à contenu particulier introduites dans le paquet ($JOB, $LOAD, $RUN, etc...)

- la multiprogrammation : partitionnement de la mémoire permettant au processeur d'accueillir une tâche dans chaque partie et donc d'être utilisé plus efficacement par rapport aux temps d'attente introduits par les périphériques (le processeur est ré-alloué)

- les E/S tamponnées : adjonction à l'UC d'un processeur autonome capable de gérer en parallèle les E/S ou canal ou unité d'échange. Cela nécessite une politique de partage du bus ou d'autres mécanismes (vol de cycle, DMA).

- les SE en réseaux : ils permettent à partir d'une machine de se connecter sur une machine distante, de transférer des données. Mais chaque machine dispose de son propre SE

- les SE distribués ou répartis : l'utilisateur ne sait pas où sont physiquement ses données, ni où s'exécute son programme. Le SE gère l'ensemble des machines connectées. Le système informatique apparaît comme un mono-processeur.

2. ELEMENTS DE BASE D'UN SYSTEME D'EXPLOITATION

Les principales fonctions assurées par un SE sont les suivantes :

- gestion de la mémoire principale et des mémoires secondaires,
- exécution des E/S à faible débit (terminaux, imprimantes) ou haut débit (disques, bandes),
- multiprogrammation, temps partagé, parallélisme : interruption, ordonnancement, répartition en mémoire, partage des données
- lancement des outils du système (compilateurs, environnement utilisateur,...) et des outils pour l'administrateur du système (création de points d'entrée, modification de privilèges,...),
- lancement des travaux,
- protection, sécurité ; facturation des services,
- réseaux

2.1 Les processus

Un processus est un programme qui s'exécute, ainsi que ses données, sa pile, son compteur ordinal, son pointeur de pile et les autres contenus de registres nécessaires à son exécution.
Les appels système relatifs aux processus permettent généralement d'effectuer au moins les actions suivantes :
- création d'un processus (fils) par un processus actif (d'où la structure d'arbre de processus gérée par un SE)
- destruction d'un processus
- mise en attente, réveil d'un processus
- suspension et reprise d'un processus, grâce à l'ordonnanceur de processus (scheduler)
- demande de mémoire supplémentaire ou restitution de mémoire inutilisée
- attendre la fin d'un processus fils
- remplacer son propre code par celui d'un programme différent

- échanges de messages avec d'autres processus
- spécification d'actions à entreprendre en fonction d'événements extérieurs asynchrones
- modifier la priorité d'un processus

Dans une entité logique unique, généralement un mot, le SE regroupe des informations-clés sur le fonctionnement du processeur : c'est le mot d'état du processeur (Processor Status Word, PSW). Il comporte généralement :

- la valeur du compteur ordinal
- des informations sur les interruptions (masquées ou non)
- le privilège du processeur (mode maître ou esclave)
- etc.... (format spécifique à un processeur)

A chaque instant, un processus est caractérisé par son état courant : c'est l'ensemble des informations nécessaires à la poursuite de son exécution (valeur du compteur ordinal, contenu des différents registres, informations sur l'utilisation des ressources). A cet effet, à tout processus, on associe un bloc de contrôle de processus (BCP). Il comprend généralement :

- une copie du PSW au moment de la dernière interruption du processus
- l'état du processus : prêt à être exécuté, en attente, suspendu, ...
- des informations sur les ressources utilisées
- mémoire principale
- temps d'exécution
- périphériques d'E/S en attente
- files d'attente dans lesquelles le processus est inclus, etc...
- et toutes les informations nécessaires pour assurer la reprise du processus en cas d'interruption

Les BCP sont rangés dans une table en mémoire centrale à cause de leur manipulation fréquente.

2.2 Les interruptions

Une interruption est une commutation du mot d'état provoquée par un signal généré par le matériel. Ce signal est la conséquence d'un événement interne au processus, résultant de son exécution, ou bien extérieur et indépendant de son exécution. Le signal va modifier la valeur d'un indicateur qui est consulté par le SE. Celui-ci est ainsi informé de l'arrivée de l'interruption et de son origine. A chaque cause d'interruption est associé un niveau d'interruption. On distingue au moins 3 niveaux d'interruption :

- les interruptions externes : panne, intervention de l'opérateur, ...
- les déroutements qui proviennent d'une situation exceptionnelle ou d'une erreur liée à l'instruction en cours d'exécution (division par 0, débordement, ...)
- les appels système

Le chargement d'un nouveau mot d'état provoque l'exécution d'un autre processus, appelé le traitant de l'interruption. Le traitant réalise la sauvegarde du contexte du processus interrompu (compteur ordinal, registres, indicateurs,....). Puis le traitant accomplit les opérations liées à l'interruption concernée et restaure le contexte et donne un nouveau contenu au mot d'état : c'est l'acquittement de l'interruption.

Généralement un numéro de priorité est affecté à un niveau d'interruption pour déterminer l'ordre de traitement lorsque plusieurs interruptions sont positionnées. Il est important de pouvoir retarder, voire annuler la prise en compte d'un signal d'interruption. Les techniques que l'on utilise sont le masquage et le désarmement des niveaux d'interruption :
- le masquage d'un niveau retarde la prise en compte des interruptions de ce niveau. Pour cela, on positionne un indicateur spécifique dans le mot d'état du processeur. Puisqu'une interruption modifie le mot d'état, on peut masquer les interruptions d'autres niveaux pendant l'exécution du traitant d'un niveau. Lorsque le traitant se termine par un acquittement, on peut alors démasquer des niveaux qui avaient été précédemment masqués. Les interruptions intervenues pendant l'exécution du traitant peuvent alors être prises en compte

- le désarmement d'un niveau permet de supprimer la prise en compte de ce niveau par action sur le mot d'état. Pour réactiver la prise en compte, on réarme le niveau. Il est évident qu'un déroutement ne peut être masqué; il peut toutefois être désarmé.

2.3 Les ressources

On appelle ressource tout ce qui est nécessaire à l'avancement d'un processus (continuation ou progression de l'exécution) : processeur, mémoire, périphérique, bus, réseau, compilateur, fichier, message d'un autre processus, etc... Un défaut de ressource peut provoquer la mise en attente d'un processus.

Un processus demande au SE l'accès à une ressource. Certaines demandes sont implicites ou permanentes (la ressource processeur). Le SE alloue une ressource à un processus. Une fois une ressource allouée, le processus a le droit de l'utiliser jusqu'à ce qu'il libère la ressource ou jusqu'à ce que le SE reprenne la ressource (on parle en ce cas de ressource préemptible, de préemption).

On dit qu'une ressource est en mode d'accès exclusif si elle ne peut être allouée à plus d'un processus à la fois. Sinon, on parle de mode d'accès partagé. Un processus possédant une ressource peut dans certains cas en modifier le mode d'accès.
Exemple : un disque est une ressource à accès exclusif (un seul accès simultané), une zone mémoire peut être à accès partagé.

2.4 L'ordonnancement

On appelle ordonnancement la stratégie d'attribution des ressources aux processus qui en font la demande. Différents critères peuvent être pris en compte :

- temps moyen d'exécution minimal

- temps de réponse borné pour les systèmes interactifs
- taux d'utilisation élevé de l'UC
- respect de la date d'exécution au plus tard, pour le temps réel, etc...

2.5 Le système de gestion de fichiers

Une des fonctions d'un SE est de masquer les spécificités des disques et des autres périphériques d'E/S et d'offrir au programmeur un modèle de manipulation des fichiers agréable et indépendant du matériel utilisé.

Les appels système permettent de créer des fichiers, de les supprimer, de lire et d'écrire dans un fichier. Il faut également ouvrir un fichier avant de l'utiliser, le fermer ultérieurement. Les fichiers sont regroupés en répertoires arborescents; ils sont accessibles en énonçant leur chemin d'accès (chemin d'accès absolu à partir de la racine ou bien chemin d'accès relatif dans le cadre du répertoire de travail courant).

Le SE gère également la protection des fichiers.