Appunti di Sistemi Operativi

di:

Facchini Luca

Corso tenuto dal prof. Sistemi Operativi

Università degli Studi di Trento

A.A. 2024/2025

Autore:
Facchini Luca
Mat. 245965
Email:luca.facchini-1@studenti.unitn.it
luca@fc-software.it

Corso:
Sistemi Operativi [146065]
CdL: Laurea Triennale in Informatica
Prof. Crispo Bruno
Email: bruno.crispo@unitn.it

Sommario

Appunti del corso di Sistemi Operativi, tenuto dal prof. Crispo Bruno presso l'Università degli Studi di Trento. Corso seguito nell'anno accademico 2024/2025.
Dove non specificato diversamente, le immagini e i contenuti sono tratti dalle slide del corso del prof. Crispo Bruno (bruno.crispo@unitn.it)

Indice

1 Definizioni e Storia **1**

2 Componenti di un sistema operativo **2**
2.1 Le Componenti in generale . . . 2
2.2 Come usare i servizi dei sistemi operativi . . . 3
2.2.1 Interprete dei comandi . . . 3
2.2.2 L'interfaccia grafica . . . 3
2.2.3 *System calls* . . . 3

3 Architettura di un Sistema Operativo **6**
3.1 Tipi di architetture . . . 6
3.2 Implementazione di un `SO` . . . 9

4 Processi e *Thread* **10**
4.1 Processi . . . 10
4.1.1 Stato di un processo . . . 10
4.1.2 Operazioni sui processi . . . 11
4.1.3 Gestione dei processi del `SO` . . . 12
4.2 *Thread* . . . 13
4.2.1 Implementazione dei *thread* . . . 13
4.2.2 Esempio di libreria - `pthreads` . . . 14

5 Comunicazione tra processi **16**
5.1 `IPC` - *Message Passing* . . . 16
5.1.1 Nominazione . . . 17
5.1.2 Sincronizzazione . . . 18
5.2 `IPC` - Memoria Condivisa . . . 18

6 *Scheduling* della `CPU` **20**
6.1 Concetto di *Scheduling* . . . 20
6.2 Tipi di *Scheduling* . . . 20
6.3 Scheduling della `CPU` . . . 21
6.3.1 Algoritmi di *scheduling* . . . 23
6.3.2 Valutazione degli algoritmi . . . 27

7 Sincronizzazione dei processi **29**
7.1 Problema della sezione critica . . . 29
7.1.1 Soluzioni al problema della sezione critica . . . 30
7.1.2 Soluzioni *hardware* . . . 32
7.2 Semafori . . . 34

8 Deadlock **38**
8.1 Prevenzione del deadlock . . . 38
8.1.1 Prevenzione Statica . . . 38
8.1.2 Prevenzione Dinamica . . . 39
8.1.3 Rilevamento del *deadlock* & ripristino . . . 41
8.1.4 Conclusioni . . . 42

9 Gestione della memoria **43**
9.1 Introduzione . . . 43
9.1.1 Dal programma al processo . . . 43
9.1.2 Spazi di indirizzamento . . . 45
9.2 Schemi di gestione della memoria . . . 45
9.2.1 Allocazione contigua . . . 45
9.2.2 Paginazione . . . 47
9.3 Segmentazione . . . 48
9.4 Segmentazione con paginazione . . . 49

10 Memoria Virtuale **51**
10.1 Paginazione su domanda . . . 51
10.2 Algoritmi di rimpiazzo delle pagine . . . 53
10.3 Allocazione dei frame . . . 55

11 Gestione della memoria secondaria **57**
11.1 Tipologia della memoria secondaria . . . 57
11.2 *Scheduling* degli accessi al disco . . . 58
11.3 Gestione del disco . . . 59

12 File System **61**
12.1 L'interfaccia del *file system* . . . 61
12.2 Struttura delle *directory* . . . 62
12.3 Implementazione del File system . . . 63
12.4 Gestione dello spazio libero . . . 65
12.5 Efficienza e Prestazioni . . . 66

13 RAID **67**

14 Il sottosistema di I/O **70**
14.1 Hardware di `I/O` . . . 70
14.2 Interfacce di `I/O` . . . 72
14.3 Software di `I/O` . . . 72

Capitolo 1

Definizioni e Storia

Capitolo 2

Componenti di un sistema operativo

Dopo aver definito cosa sia un sistema operativo, vediamo ora quali siano le sue componenti principali, a partire dalla gestione dei processi e della memoria (primaria e secondaria), per poi passare alla gestione dei dell `I/O` e dei file fino ad arrivare alla protezione, la gestione della rete e l'interprete dei comandi.

2.1 Le Componenti in generale

Gestione dei Processi

Definizione 2.1 (Processo)**.** *Un **processo** è un programma in esecuzione che necessita di **risorse** per poter funzionare. Questo inoltre è eseguito in modo **sequenziale** ed **una istruzione alla volta**, infine è possibile che un processo sia del `SO` o dell'utente.*

In materia di gestione dei processi il sistema operativo è responsabile nella loro creazione e distruzione, nella loro sospensione e ripresa e deve fornire dei meccanismi per la sincronizzazione e la comunicazione tra i processi stessi.

Gestione della memoria primaria

Definizione 2.2 (Memoria primaria)**.** *La **memoria primaria** è la memoria principale del computer che conserva dati condivisi dalla `CPU` e dai dispositivi `I/O` questa è direttamente accessibile dalla `CPU`, per essere eseguito un programma deve essere caricato in memoria.*

La gestione della memoria primaria richiede la gestione dello spazi di memoria oltre alla decisione su quale processo debba essere caricato in memoria e quale debba essere rimosso. Inoltre il sistema operativo deve fornire dei meccanismi allocare e de-allocare la memoria.

Gestione della memoria secondaria

Definizione 2.3 (Memoria secondaria)**.** *La **memoria secondaria** è una memoria **non volatile** ed **grande** rispetto alla memoria primaria, questa è utilizzata per memorizzare i dati e i programmi in modo **permanente**.*

Questa memoria consiste di uno o più dischi (magnetici) ed il sistema operativo deve fornire dei meccanismi per la gestione dello spazio libero, l'allocazione dello spazio ed lo *scheduling* degli accessi ai dischi.

Gestione dell'`I/O`

Il `SO` nasconde la complessità dell'`I/O` ai programmi utente, fornendo un'astrazione dell'`I/O` e fornendo dei meccanismi per: accumulare gli accessi ai dispositivi (*buffering*), fornire una interfaccia generica per i dispositivi e fornire dei *driver* specifici (scritti in `C`, `C++` o *assembly*).

Gestione dei file

Definizione 2.4 (File)**.** *Un **file** è una sequenza di byte memorizzata in un qualsiasi supporto fisico controllato da driver del sistema operativo.*

Un file è dunque un'astrazione logica per rendere più semplice la memorizzazione e l'uso della memoria **non volatile**. Il sistema operativo deve fornire dei meccanismi per la creazione, la cancellazione, la lettura e la scrittura di file e *directory* oltre a fornire delle primitive (copia, sposta, rinomina) per la gestione dei file.

Protezione

Il sistema operativo deve fornire dei meccanismi per controllare l'accesso a tutte le risorse da parte di processi e utenti, inoltre l'`SO` è responsabile della definizione di accessi autorizzati e non autorizzati, oltre a definire i controlli necessari ed a fornire dei meccanismi per verificare le politiche di accesso definite.

2.2 Come usare i servizi dei sistemi operativi

Il sistema operativo metta a disposizione le sue interface tramite delle *system call* che sono delle chiamate a funzione che permettono di accedere ai servizi del sistema operativo precedentemente descritti. Queste chiamate a funzione sono utilizzate per eseguire operazioni che richiedono privilegi di sistema, come ad esempio la gestione dei processi, della memoria, dell'`I/O` e dei file.

2.2.1 Interprete dei comandi

Un esempio di utilizzo delle *system call* è l'interazione con l'interprete dei comandi, che permette di eseguire comandi e programmi tramite una interfaccia testuale. Questo interprete tramuta i comandi in *system call* che vengono poi eseguite dal sistema operativo. Questo permette di creare e gestire processi, gestire `I/O`, disco, memoria e file oltre alla gestione delle protezione e della rete.
Nel `SO` esistono dei comandi predefiniti che possono essere chiamati direttamente per il loro nome, questi sono implementati con una semantica specifica e possono essere utilizzati per eseguire operazioni di base, nel caso di comandi non predefiniti è possibile scrivere dei programmi che vengono eseguiti dall'interprete dei comandi.

2.2.2 L'interfaccia grafica

Un'altra interfaccia che permette di interagire con il sistema operativo è l'interfaccia grafica, che permette di interagire con il sistema operativo tramite il *mouse* e la tastiera. Questa interfaccia più intuitiva e facile da usare rispetto all'interprete dei comandi, permette di interagire con il `SO` tramite icone e finestre. Questa interfaccia, anche se più semplice, non è per forza più veloce dell'interprete dei comandi, in quanto l'interfaccia grafica è più lenta e richiede più risorse rispetto all'interprete dei comandi.

2.2.3 *System calls*

I processi non usano le *shell* per eseguire le *system call*, ma usano delle `API` (*Application Programming Interface*) che permettono di accedere ai servizi del sistema operativo. Queste `API` sono delle librerie di funzioni ad alto livello che permettono di accedere ai servizi del sistema operativo. Queste librerie sono scritte in `C` o `C++` e permettono di accedere ai servizi del sistema operativo in modo più semplice e più sicuro rispetto all'uso diretto delle *system call*.

Esempio di `API` Un esempio di `API` è la `Win32`, prendiamo in esame la funzione `ReadFile` che permette di leggere un file:

```
BOOL ReadFile (
    HANDLE file,
    LPVOID buffer,
    DWORD bytes to read,
    LPDWORD bytes read,
    LPOVERLAPPED ovl
);
```

Questa funzione ritorna un valore booleano che indica se la funzione è andata a buon fine o meno, inoltre questa funzione prende in input il file da leggere, il buffer in cui scrivere i dati letti, il numero di byte da leggere, il numero di byte letti e un puntatore a una struttura `OVERLAPPED` che permette di specificare un offset per la lettura.

Le `API` nei diversi `SO`

Le 2 `API` più comuni per Windows sono: `Win32` e `Win64` mentre per Linux sono: `POSIX` (*Portable Operating-System Interface*) che includono le *system call* per tutte le versioni di `UNIX`, *Linux* e *Mac OS X*, o tutte le distribuzioni `POSIX`-*compliant*.

Windows* su *Linux Per eseguire programmi *Windows* su *Linux* è possibile usare `Wine` che è un *emulatore* il quale traduce le chiamate `API` di *Windows* in chiamate `API` di *Linux on-the-fly*, ovvero durante l'esecuzione del programma. Questo permette di eseguire programmi *Windows* su *Linux* senza dover riscrivere il codice del programma.

Implementazione delle *System Call*

Ad ogni *system call* è associato un numero univoco, che permette al sistema operativo di identificare la *system call* richiesta. È compito dell'interfaccia tenere traccia dei numeri associati alle *system call* e di passare i parametri alla *system call* richiesta. Questa interfaccia invoca la *system call* nel *kernel* del sistema operativo, che esegue la *system call* e ritorna il risultato al chiamante. Questo meccanismo permette al chiamante di non dover conoscere i dettagli di implementazione della *system call* ma solo la sua interfaccia.

Esecuzione delle *system calls* Per eseguire una *system call* dopo che il processo ha eseguito la chiamata all'interfaccia del `SO` il quale conoscendo il numero della *system call* controlla dove questa è implementata tramite la *system call table* (una tabella che contiene i puntatori alla implementazione delle *system call*). Una volta trovata la *system call* il `SO` esegue la *system call* e ritorna il risultato al chiamante.

Opzioni per il passaggio dei parametri I parametri di una *system call* possono essere passati in diversi modi. I più comuni sono: passaggio tramite registri, passaggio tramite lo `stack` e passaggio tramite puntatori. Il passaggio tramite registri è il più veloce ma permette di passare pochi parametri e di piccola dimensione, il passaggio tramite lo `stack` permette di passare più parametri e di dimensioni maggiori, infine il passaggio tramite puntatori permette di passare parametri di dimensioni maggiori e di passare parametri complessi, ma và passata una tabella di parametri che deve essete passata tramite `stack` o registri.

Parametri tramite `stack`Il passaggio dei parametri tramite `stack` avviene in questo modo:

1-3 Salvataggio parametri sullo `stack`

4 Chiamata della funzione di libreria

5 Caricamento del numero della *system call* su un registro `Rx`

6 Esecuzione `TRAP` (Passaggio in *kernel mode*)

7-8 Esecuzione della *system call*

9 Ritorno al chiamante

10-11 Ritorno al codice utente ed incremento dello `stack pointer`

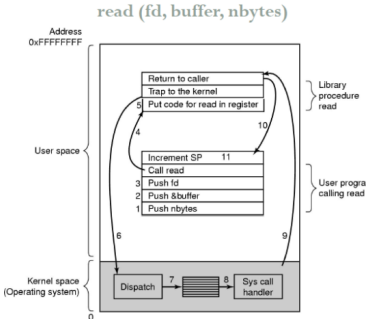


Figura 2.1: Passaggio dei parametri tramite `stack`

Passaggio di parametri tramite tabella Come anticipato il passaggio di parametri tramite tabella viene utilizzato per passare parametri complessi o di dimensioni maggiori andando a passare un puntatore alla tabella che contiene i parametri. Questo metodo permette di passare un numero maggiore di parametri e di dimensioni maggiori in quanto i parametri sono passati per riferimento alla memoria primaria.

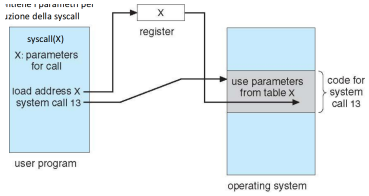


Figura 2.2: Passaggio dei parametri tramite tabella

Capitolo 3

Architettura di un Sistema Operativo

In un sistema operativo è molto importante separare le *policy* dai *meccanismi*. I meccanismi sono le funzionalità che il sistema operativo mette a disposizione, mentre le *policy* sono le regole che il sistema operativo segue per decidere come utilizzare i meccanismi.

Principi di progettazione Il principio di progettazione di un sistema operativo è quello di `KISS` (*Keep It Small and Simple*) usato per ottimizzare al meglio le *performance* implementando solo lo stretto necessario. Altro principio è il `POLP` (*Principle of Least Privilege*), ovvero dare il minimo dei privilegi necessari ad ogni componente per svolgere il proprio compito. Quest'ultimo principio è molto importante per garantire affidabilità e sicurezza.

3.1 Tipi di architetture

Sistemi monoblocco

Nei sistemi monoblocco non è presente una gerarchia tra i vari livelli del sistema operativo. Questo tipo di architettura è molto semplice e consiste in un unico strato *software* tra l'utente ed l'*hardware* del sistema. Le componenti sono dunque tutte allo stesso livello permettendo una comunicazione diretta tra l'utente e l'*hardware*. Questo tipo di architettura è molto semplice e veloce, ma il codice risulta interamente dipendente dall'architettura ed è distribuito su tutto il sistema operativo. Inoltre per testare ed eseguire il *debugging* di un singolo componente è necessario analizzare l'intero sistema operativo.

Sistemi a struttura semplice

Nei sistemi a struttura semplice è presente una piccola gerarchia, molto flessibile, tra i vari livelli del sistema operativo. Questo tipo di architettura mira ad una riduzione dei costi di sviluppo ed di manutenzione del sistema operativo. Non avendo una struttura ben definita, i componenti possono comunicare tra loro in modo diretto. Questo tipo di architettura è molto flessibile e permette di avere un sistema operativo molto piccolo e veloce come `MS-DOS` o `UNIX` originale.

`MS-DOS` Il sistema operativo `MS-DOS` è un sistema operativo a struttura semplice, molto piccolo e veloce. Questo sistema operativo è pensato per fornire il maggior numero di funzionalità in uno spazio ridotto. Infatti non sussistono suddivisioni in moduli, ed le interfacce e livelli non sono ben definiti. È infatti possibile accedere direttamente alle *routine* del sistema operativo ed non è prevista la *dual mode*.

`UNIX` **(Originale)** Struttura semplice limitata dalle poche funzionalità disponibili all'epoca in materia di *hardware*, con un *kernel* molto piccolo e veloce il quale scopo è risiedere tra l'interfaccia delle *system call* e l'*hardware*. Questo sistema operativo è stato progettato per essere molto flessibile e fornisce: *File System*, *Scheduling* della `CPU`, gestione della memoria e molto altro.

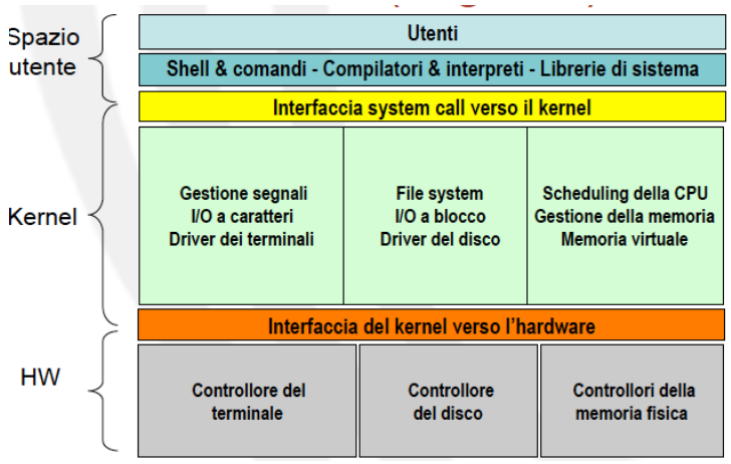


Figura 3.1: Struttura di `UNIX` originale

Sistema a livelli

Nei sistemi operativi organizzati a livelli gerarchici l'interfaccia utente risiede al livello più altro mentre l'*hardware* dal lato opposto. Ogni livello intermedio può solo usare funzioni fornite dal livello inferiore ed offrire funzionalità al livello superiore. Principale vantaggio di questa architettura è la modularità, infatti ogni livello può essere sviluppato e testato indipendentemente dagli altri. Questo tipo di architettura, d'altronde non è priva di svantaggi, infatti diventa difficile definire in modo approssimato gli strati, l'efficienza decresce in quanto ogni singolo strato aggiunge un costo di *overhead* ed le funzionalità dipendenti dal'*hardware* sono sparse su più livelli.

`THE` Il sistema operativo `THE` è un sistema d'uso accademico ed è il primo sistema operativo a struttura a livelli. Questo `SO` consiste in un insieme di processo che cooperano tra di loro usando la tecnica dei "semafori" per la sincronizzazione.

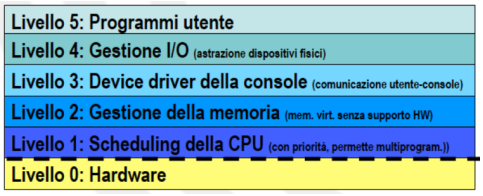


Figura 3.2: Struttura di `THE`

Sistemi basati su *Kernel*

I sistemi di questo genere hanno due soli livelli: i servizi *kernel* e quelli non-*kernel* (o *utente*). Il *file system* è un esempio di servizio non-*kernel*. Questo tipo di architettura è molto diffuso in quanto il ridotto e ben definito numero di livelli ne permette una facile implementazione e manutenzione, spesso però questo sistema può risultare troppo rigido e non adatto a tutti i tipi di applicazioni, oltre alla totale assenza di regole organizzative per le parti del `SO` al di fuori del *kernel*.

Micro-kenrel

Questo tipo di *kernel* è molto piccolo e fornisce solo i servizi essenziali per il funzionamento del sistema operativo. Tutte le altre funzionalità sono implementate come processi utente. Un esempio di ciò è `seL4` un *kernel open source* che implementa un *micro-kernel* e fornisce un'interfaccia per la gestione della memoria, dei processi e della comunicazione tra processi. `SeL4` è matematicamente verificato e privo di bug rispetto alle sue specifiche di forte sicurezza

Virtual Machine

L'architettura a VM è una estremizzazione dell'approccio a più livelli di IBM (1972), questo è pensato per offrire un sistema di *timesharing* "multiplo" dove il sistema operativo viene eseguito su una VM ed questa dà illusione di processi multipli, ma nella realtà ognuno di questi è in esecuzione sul proprio HW. In questo paradigma sono possibili più SO in una unica macchina.

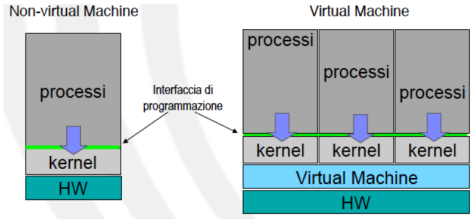


Figura 3.3: Differenze tra una macchina senza e con VM

Come è possibile notare ogni singolo processo è separato ed possiede il proprio *kernel*. Vengono quindi separata la multiprogrammazione ed la presentazione.

Tipo di *Hypervisor*

- **Tipo 1** (*Bare Metal*): Questo tipo di *Hypervisor* è installato direttamente sul'*hardware* e non necessita di un sistema operativo ospite. Questo tipo di *Hypervisor* è molto veloce e sicuro, ma è molto complesso da installare e configurare.
- **Tipo 2** (*Hosted*): Questo tipo di *Hypervisor* è installato sopra un sistema operativo ospite. Questo tipo di *Hypervisor* è molto più semplice da installare e configurare rispetto al tipo 1, ma è più lento e meno sicuro, inoltre è possibile avere problemi di compatibilità tra il sistema operativo ospite e il *Hypervisor*.

***Monolithic* vs *Micro-kernel* VM** Prima di fare una distinzione tra i due tipi di VM è necessario dire che entrambi rientrano nel tipo 1 di *Hypervisor* e dunque tutti i SO sono eseguiti direttamente sul'*hardware* virtualizzato.

- ***Monolithic***: Questo tipo di VM è molto simile ad un sistema operativo tradizionale, infatti ogni *VM* è un processo separato che esegue il proprio *kernel*. Questo tipo di VM è molto veloce, ma è molto complesso da implementare e mantenere.
- ***Micro-kernel***: Questo tipo di VM è molto simile ad un sistema operativo a *micro-kernel*, infatti il *kernel* della VM fornisce solo i servizi essenziali per il funzionamento del sistema operativo. Tutte le altre funzionalità sono implementate come processi utente. Questo tipo di VM è molto più semplice da implementare e mantenere rispetto al tipo 1, ma è più lento e meno sicuro.

Vantaggi - Svantaggi Principale vantaggio di questo tipo di architettura è la completa protezione del sistema, infatti ogni SO è separato e non può accedere alle risorse degli altri SO. Inoltre è possibile avere più SO in una sola macchina andando ad ottimizzare le risorse e ridurre i costi di sviluppo di un sistema operativo, oltre ad aumentare la portabilità delle applicazioni. Principale svantaggio riguardano le prestazioni del sistema, infatti ogni SO è eseguito su una VM e questo può portare ad un aumento dei tempi di esecuzione delle applicazioni. Inoltre è necessario avere gestire una *dual-mode* virtuale e non è possibile avere un sistema operativo in tempo reale, inoltre il fatto che una VM non possa accedere alle altre VM può portare ad un aumento dei costi di sviluppo e manutenzione del sistema.

Sistemi *client-server*

Poco diffusi ai giorni nostri, i sistemi *client-server* sono basati su un'architettura a due livelli: il *client* e il *server*. Questo sistema di basa sull'idea che il codice del sistema operativo vada portato sul livello superiore (il *client*) e il *server* rimanga molto piccolo e veloce andando solo a fornire i servizi essenziali per il funzionamento del sistema operativo ed la comunicazione tra il *client* e l'*hardware*. Questo tipo di architettura si presta bene per sistemi distribuiti.

3.2 Implementazione di un SO

I sistemi operativi sono tradizionalmente scritti in linguaggio *assembler* anche se è possibile scriverli in linguaggi di alto livello, come `C` o `C++`. La scrittura di un sistema operativo in linguaggio di alto livello permette di avere una implementazione molto rapida oltre ad aumentarne la compattezza e la mantienebilità. Inoltre è possibile avere una maggiore portabilità del sistema operativo, in quanto è possibile compilare il codice sorgente su più architetture. Tuttavia la scrittura di un sistema operativo in linguaggio di alto livello può portare ad un aumento dei tempi di esecuzione delle applicazioni e ad un aumento dei costi di sviluppo e manutenzione del sistema operativo.

Capitolo 4

Processi e *Thread*

In questo capitolo vedremo cosa sono i processi e i *thread* capendone le differenze e le somiglianze, vedremo come vengono gestiti e come vengono eseguiti. Infine vedremo come vengono gestiti i processi dal sistema operativo e come vengono eseguiti i processi dal sistema operativo.

4.1 Processi

Un processo è l'istanza di un programma in esecuzione, quando il programma viene eseguito e quindi caricato nella memoria primaria (`RAM`) diventa un processo. Mentre un programma è la parte statica di un software, il processo è la parte dinamica. Un processo viene eseguito in maniera sequenziale, ovvero un'istruzione alla volta, ma nei sistemi operativi moderni un processo può essere eseguito in maniera concorrente, ovvero più processi possono essere eseguiti in parallelo.

Immagine in memoria

Un processo quando viene caricato in memoria viene caricato in una zona di memoria chiamata *spazio degli indirizzi* (*address space*). Questo spazio è diviso in varie sezioni (da indirizzi alti ad indirizzi bassi):

- **Dati**: contiene le variabili globali e statiche del programma.
- ***Stack***: contiene le variabili locali e i parametri delle funzioni.
- eventuale memoria dinamica allocata durante l'esecuzione.
- ***Heap***: contiene la memoria dinamica allocata durante
- **Codice**: contiene il codice del programma.
- **Attributi del processo**: contiene informazioni sul processo.

4.1.1 Stato di un processo

Un processo durante la sua creazione ed esecuzione può trovarsi in diversi stati:

- **Nuovo**: il processo è stato creato ma non è ancora in esecuzione.
- **Pronto**: il processo è pronto per essere eseguito, ma non è ancora in esecuzione. (oppure è stato messo in attesa dalla `CPU`).
- **In esecuzione**: il processo è in esecuzione sulla `CPU`.
- **In attesa**: il processo è in attesa di un evento (es. `I/O`).
- **Terminato**: il processo è terminato.

Per la gestione di questi stati il sistema operativo usa un *dispatcher* il quale compito è quello di passare tra i processi e cambiare il loro stato. Per questo motivo il *dispatcher* è chiamato anche *scheduler*.

Scheduling

Lo *scheduling* è il processo di selezione del processo da eseguire sulla `CPU`. Esistono vari tipi di *scheduler*:

- ***Long time scheduler***: decide quali processi devono essere caricati in memoria. (Nella coda dei processi pronti).
- ***Short time scheduler***: decide quale processo deve essere eseguito sulla `CPU`. (Seleziona i processi dalla coda dei processi pronti).

Mentre lo *short-term* scheduler è un processo molto veloce in quanto viene chiamato molto spesso (ogni $10 - 100ms$), il *long-term* scheduler è un processo più lento in quanto viene chiamato molto raramente (ogni $1 - 10s$ o anche di più), questo però è responsabile del grado di multiprogrammazione del sistema.

Accantonamento L'accantonamento è il processo per il quale i processi pronti ad essere eseguiti vengono messi in una coda di attesa. Quando la `CPU` è pronta per eseguire un processo, il processo viene preso dalla coda e viene eseguito, nel caso nel quale il processo richieda un'operazione di `I/O` il processo viene messo in richiesta ed quando l'operazione di `I/O` (caratterizzata a sua volta da una coda per ogni dispositivo connesso) è completata il processo viene rimesso nella coda dei processi pronti.
Può anche succedere che il tempo per l'esecuzione di un processo sia scaduto, in questo caso il processo viene rimesso nella coda dei processi pronti. Se poi il processo generi dei processi figli, questi dopo la loro inizializzazione vengono messi nella coda dei processi pronti e vengono eseguiti, se il padre necessita che il processo figlio termini prima di lui, il padre viene messo in attesa che il figlio termini, altrimenti anche il padre viene messo nella coda dei processi pronti. Infine se un processo necessita di un segnale da parte di un altro processo, il processo viene messo in attesa finché non riceve il segnale (dal sistema o da un altro processo).

`I/O` vs `CPU` *bound* Un processo può essere `I/O` bound o `CPU` *bound*. Un processo `I/O` *bound* è un processo che richiede molte operazioni di `I/O` e poche operazioni sulla `CPU`, mentre un processo `CPU` *bound* è un processo che richiede molte operazioni sulla `CPU` e poche operazioni di `I/O`. Non è possibile stabilire a priori se un processo è `I/O` *bound* o `CPU` *bound*, ma è possibile stabilirlo solo durante l'esecuzione del processo analizzando quanta `CPU` usa e se richiede molte operazioni di `I/O`, sulla base di questo il processo viene classificato come `I/O` *bound* o `CPU` *bound*.

Operazione di *dispatch*

Quando si deve passare da un processo ad un altro si deve fare un'operazione di *dispatch*. Questa operazione consiste nel:

1. Cambiare il contesto (salvare lo stato del processo corrente (`PCB`) e caricare lo stato del processo successivo (`PCB`)).
2. Passare alla modalità utente (quando viene eseguito il *context switch* il sistema operativo è in modalità *kernel*, mentre il processo deve essere eseguito in modalità utente).
3. Salto alla prossima istruzione da eseguire del processo successivo.

Questa operazione è molto costosa in termini di tempo, in particolare l'operazione di *context switch* richiede risorse che rallentano il sistema senza eseguire nessuna operazione utile, la durata di ciò è strettamente dipendente dall'architettura del processore e dal sistema operativo.

4.1.2 Operazioni sui processi

Nella quasi totalità dei sistemi operativi moderni è possibile eseguire più processi in parallelo, per fare ciò il sistema operativo deve fornire delle operazioni per la gestione dei processi oltre ad un modo per l'identificazione dei processi. Di seguito vediamo quali sono le operazioni possibili sui processi.

Creazione di un processo

Un processo, come già detto, può creare altri processi, questi processi creati sono detti processi figli. Un processo padre può creare più processi figli, questi processi figli possono creare a loro volta altri processi figli e così via. Ai processi normalmente viene associato un *PID* (*Process IDentifier*) che è un numero

univoco che identifica il processo all'interno del sistema operativo.
Il processo figlio può ottenere le risorse necessarie per la sua esecuzione in due modi:

- Ereditando le risorse del processo padre (*sharing*)
- Ottenendo nuove risorse dal sistema operativo (*partitioning*)

Inoltre il processo figlio può essere eseguito in parallelo in maniera sincrona rispetto al processo padre (il processo padre aspetta che il processo figlio termini) o asincrona (il processo figlio viene eseguito in parallelo al processo padre).

Nei sistemi `UNIX` Nei sistemi `UNIX` esistono diverse *system call* per la creazione di processi, la principale è `fork()` che crea un processo figlio identico al processo padre, la differenza tra i due processi è il *PID* e il *PPID* (*Parent Process IDentifier*). Il processo figlio eredita tutte le risorse del processo padre, inoltre il processo figlio può modificare le risorse ereditate dal processo padre. Altra chiamata di sistema è `exec()` che permette di caricare un nuovo programma in un processo figlio, in questo caso il programma tra il processo padre e il processo figlio è differente. Infine la chiamata di sistema `wait()` permette l'esecuzione sincrona di un processo figlio rispetto al processo padre.

Terminazione di un processo

Un processo può terminare in tre modi:

- **Normalmente**: il processo termina la sua esecuzione invocando la *system call* `exit()` (con eventualmente un codice di uscita).
- **Forzatamente dal processo padre**: il processo padre può terminare il processo figlio invocando la *system call* `kill()`, oppure nel caso di un eccessivo uso di risorse, oppure a sua volta il processo padre termina anormalmente.
- **Forzatamente dal sistema operativo**: il sistema operativo può terminare un processo nel caso di un errore di esecuzione, oppure nel caso nel quale l'utente chiuda l'applicazione.

Nota come nel primo caso non sia esclusa la possibilità che il processo termini in maniera anomala, ad esempio per un errore di esecuzione gestito dal processo stesso, infatti quando il codice di uscita è diverso da `0` si intende che il processo è terminato in maniera anomala, ogni codice diverso da `0` ha un significato diverso.
Quando un processo termina il sistema operativo si occupa di liberare le risorse utilizzate dal processo come la memoria allocata, i file aperti, le connessioni di rete, o altre risorse.

4.1.3 Gestione dei processi del `SO`

Di fatto il sistema operativo non è altro che un programma a tutti gli effetti, e dunque la sua esecuzione è un processo come un altro. Questo non significa però che il sistema operativo non essere gestito separatamente dagli altri processi, infetti esistono diverse opzioni l'esecuzione del *kernel*:

- Il *kernel* viene eseguito completamente in maniera separata dagli altri processi.
- Il *kernel* viene eseguito all'interno di un processo utente.
- Il *kernel* viene eseguito come un processo separato.

***Kernel* separato** In questo caso il *kernel* è eseguito al di fuori degli altri processi, questo gli permette di avere uno spazio in memoria ben definito e riservato oltre ad avere il totale controllo del sistema ed a essere eseguito in modalità *kernel* (ovvero con privilegi elevati). I processi sono dunque solo propri all'utente ed un processo non potrà mai essere eseguito in modalità *kernel*.

***Kernel* nel processo utente** In questo caso il *kernel* è eseguito all'interno di un processo utente, questo permette ai programmi utente di chiamare qualunque servizio del sistema operativo, ma tramite una modalità protetta (*kernel mode*) che permette al sistema operativo di controllare le chiamate e di evitare che un processo utente possa fare danni al sistema. Dato che il *kernel* è un processo a tutti gli effetti la sua immagine in memoria sarà composta dal "*kernel stack*" per la gestione delle chiamate di

sistema e dal "*kernel code*" che consiste nei dati e codice del SO condiviso tra tutti i processi.
Questo approccio porta ad una riduzione del tempo di *context switch* in quanto è necessario solo la *mode switch* e non l'intero *context switch* lasciando però intatte le possibilità di riattivazione del processo utente o di eseguire un altro processo eseguendo un *context switch* completo.

***Kernel* come processo separato** In questo caso ogni servizio del sistema operativo è eseguito come un processo separato in modalità protetta. L'unica parte del *kernel* che deve essere eseguita separatamente è lo *scheduler* in quanto deve essere eseguito in modalità *kernel*. Questo approccio è molto vantaggioso per sistemi multiprocessore in quanto permette di eseguire i servizi del sistema operativo in parallelo ed in un processore designato.

4.2 *Thread*

Un *thread* è l'unità di base d'uso della CPU, un processo può contenere uno o più *thread* che condividono lo stesso codice, dati e file aperti, ma ognuno ha un suo *stack*, lo stato del *program counter* e dei registri ed un numero identificativo.
Dunque le risorse e lo spazio di indirizzamento sono propri del processo, mentre lo stato della CPU è proprio del *thread* assieme al *program counter* e ai registri.
Classicamente un processo è composto da un solo *thread*, la capacità di avere più *thread* in un processo è chiamata *multithreading*. Questo permette di avere un processo con più *thread* separando il flusso di esecuzione e lo spazio di indirizzamento, ma condividendo le risorse del processo.

Vantaggi I vantaggi del *multithreading* sono:

- **Risposta più veloce**: Se sono necessari molti calcoli o operazioni di I/O è possibile eseguire queste operazioni in parallelo.
- **Condivisione delle risorse**: I *thread* possono condividere le risorse del processo, mentre processi separati devono usare meccanismi di comunicazione.
- **Economia**: Creare un *thread* è più veloce e meno costoso di creare un processo.
- **Scalabilità**: I *thread* possono essere eseguiti in parallelo su più processori o su più core.

4.2.1 Implementazione dei *thread*

Vediamo ora come sono implementati i *thread* nei sistemi operativi.

Stato dei *thread*

Un *thread*, come un processo, può trovarsi in diversi stati:

- **Pronto**: il *thread* è pronto per essere eseguito.
- **In esecuzione**: il *thread* è in esecuzione sulla CPU.
- **In attesa**: il *thread* è in attesa di un evento.

Un *thread* può essere in uno di questi stati, ma il processo può non essere nello stesso stato di un *thread* in quanto un processo può contenere più *thread* e quindi un processo può essere in uno stato diverso da quello dei suoi *thread*.
Un classico problema degli stati dei *thread* è la questione di cosa fare quando un *thread* è in attesa di un evento, questa "attesa" deve bloccare l'intero processo o solo il *thread* in attesa? Ciò dipende dall'implementazione dei *thread* nel sistema operativo.

Implementazione dei *thread*

Esistono due principali implementazioni dei *thread*:

- ***User-level threads***: I *thread* sono implementati a livello utente, il sistema operativo non è a conoscenza dei *thread* e non li gestisce. Le funzionalità sono implementate in una libreria che gestisce i *thread* e le chiamate di sistema.

- ***Kernel-level threads***: I *thread* sono implementati a livello del *kernel*, il sistema operativo è a conoscenza dei *thread* e li gestisce.

- ***Hybrid threads***: I *thread* sono implementati a livello del *kernel*, ma il sistema operativo permette di creare *thread* a livello utente. (es. `SOLARIS`)

User-level threads Se si opta per l'implementazione dei *thread* a livello utente, il sistema operativo non è a conoscenza dei *thread* e non li gestisce e dunque non è necessario passare in modalità *kernel* per la gestione dei *thread* risparmiando due *context switch*. Ogni applicazione deve però implementate lo *scheduler* dei *thread* e la gestione degli stati dei *thread*. Quanto detto garantisce una maggiore portabilità delle applicazioni senza dover riscrivere il codice per ogni sistema operativo, ma allo stesso tempo non permette di sfruttare appieno le potenzialità del sistema operativo. Se però un *thread* necessita di un'operazione di `I/O` o di un'operazione che richiede l'intervento del sistema operativo, tutti i *thread* del processo vengono bloccati in quanto il sistema operativo non è a conoscenza dei *thread* e non può gestire i *thread* in maniera indipendente. (es. *Green threads (`JDK1.1`)*, `GNU` *Portable Threads*, `POSIX` *Pthreads*)

Kernel-level threads Se si opta per l'implementazione dei *thread* a livello del *kernel*, il sistema operativo è a conoscenza dei *thread* e li gestisce, dunque il sistema operativo può gestire i *thread* in maniera indipendente e può sfruttare appieno le potenzialità del sistema operativo. Ogni *thread* è un processo a tutti gli effetti, dunque ogni *thread* ha il proprio *PCB* e il proprio spazio di indirizzamento. Questo permette di sfruttare appieno le potenzialità del sistema operativo, ma allo stesso tempo richiede due *context switch* per passare da un *thread* all'altro. (es. `Windows`, `Linux`, *Native Threads (`JDK1.2`)*)

Hybrid threads Se si opta per l'implementazione dei *thread* ibridi, il sistema operativo permette di creare *thread* a livello utente, ma i *thread* sono implementati a livello del *kernel*. Questo permette di sfruttare appieno le potenzialità del sistema operativo, ma allo stesso tempo permette di creare *thread* a livello utente. (es. `SOLARIS`)

4.2.2 Esempio di libreria - `pthreads`

Nel caso di implementazione dei *thread* a livello utente, il sistema operativo non è a conoscenza dei *thread* e dunque non li gestisce, ma è necessario utilizzare una libreria che gestisca i *thread*. Un esempio di libreria per la gestione dei *thread* è `pthreads` (`POSIX` *Threads*).
`pthreads` è una libreria standard per la gestione dei *thread* in sistemi `UNIX` e sistemi `UNIX-like`. La libreria fornisce un'interfaccia standard per la creazione, la sincronizzazione e la terminazione dei *thread* nel linguaggio `C`. La libreria fornisce la possibilità di caratterizzare i *thread* sulla base della priorità (influenza lo *scheduling*) e della dimensione dello *stack* (stabilisce quante risorse può utilizzare il *thread*). Gli attributi di un *thread* sono contenuti nell'oggetto di tipo `pthread_attr_t` e tramite la funzione `pthread_attr_init()` si inizializzano gli attributi del *thread*. Una volta inizializzati gli attributi tramite la funzione `pthread_create()` si crea il *thread* passando come argomenti:

1. Una variabile del tipo `pthread_t` che conterrà l'identificativo del *thread*.

2. Un oggetto del tipo `pthread_attr_t` che conterrà gli attributi del *thread*.

3. Un puntatore alla funzione che il *thread* dovrà eseguire.

4. Un puntatore agli argomenti della funzione.

Una volta creato il *thread* questo terminerà quando il codice della funzione terminerà, oppure quando nel codice della funzione verrà invocata la funzione `pthread_exit()` con parametro `value_ptr` che conterrà il valore di uscita del *thread*. Se invece il *thread* deve essere sospeso in attesa di un altro *thread* si può utilizzare la funzione `pthread_join()` con parametri:

1. Un oggetto del tipo `pthread_t` che identifica il *thread* da attendere.

2. Un puntatore alla variabile che conterrà il valore di uscita del *thread* atteso.

```
#include <pthread.h>
#include <stdio.h>
void *tbody(void *arg)
    {
    int j;
    printf("ciao sono un thread, mi hanno appena creato\n");
    *(int *)arg = 10;
    sleep(2) /* faccio aspettare un po il mio creatore poi termino */
    pthread_exit((int *)50); /* oppure return ((int *)50); */
}
main(int argc, char **argv)
{
    int i;
    pthread_t mythread;
    void *result;
    printf("sono il primo thread, ora ne creo un altro \n");
    pthread_create(&mythread, NULL, tbody, (void *) &i);
    printf("ora aspetto la terminazione del thread che ho creato \n");
    pthread_join(mythread, &result);
    printf("Il thread creato ha assegnato %d ad i\n",i);
    printf("Il thread ha restituito %d \n",result);
}
```

In questo esempio la variabile "mythread" assume dei valori corrispondenti all'identificativo del *thread* creato, mentre la variabile "result" assume il valore di uscita del *thread* creato. La funzione "tbody" è la funzione che il *thread* dovrà eseguire, mentre la variabile "i" è un argomento passato alla funzione. La funzione "pthread_exit()" termina il *thread* e restituisce il valore passato come argomento, mentre la funzione "pthread_join()" sospende il *thread* corrente in attesa del *thread* passato come argomento e restituisce il valore di uscita del *thread* atteso.

Condivisione dello spazio logico

Come già anticipato i *thread* condividono lo stesso spazio logico, questo significa che i *thread* possono accedere alle stesse variabili globali e statiche e se un *thread* modifica una variabile globale, la modifica sarà visibile a tutti gli altri *thread*. Questo può portare a problemi di sincronizzazione tra i *thread* e dunque è necessario utilizzare meccanismi di sincronizzazione per evitare problemi di accesso concorrente alle variabili globali. Possono esistere variabili locali ai *thread* che sono visibili solo al *thread* che le ha dichiarate, ma non sono visibili agli altri *thread*, ciò usando la classe `thread_specific_data`.

Per la sincronizzazione Per la sincronizzazione tra i *thread* si possono utilizzare o gli strumenti direttamente forniti dalla libreria `pthreads` (come i semafori) oppure si possono utilizzare le primitive di sincronizzazione fornite dal sistema operativo (come `sleep(n)` che sospende il *thread* corrente per n secondi). Per tenere traccia del tempo trascorso nella funzione possono essere usati due metodi:

- Un *interrupt Request* (`IRQ`) che viene generato ad intervalli regolari e che incrementa un contatore. Il `SO` controlla se ci sono delle `sleep` scadute e se ci sono le risveglia.
- Riconfigurazione delle `IRQ` in modo che avvenga una `IRQ` quando la prima `sleep` scade, e una seconda `IRQ` quando la seconda `sleep` scade e così via. Ciò comporta a migliore precisione ma alto *overhead* per la riconfigurazione delle `IRQ` ad ogni `sleep`.

Capitolo 5

Comunicazione tra processi

Normalmente i processi si dividono in processi indipendenti, ovvero quei processi la cui esecuzione è indipendente da quella degli altri processi ed non condivide i dati, e processi cooperanti, ovvero quei processi che condividono i dati e devono comunicare tra loro, la loro esecuzione non è deterministica e non è riproducibile.

In generale Esistono diversi motivi per cui i processi devono comunicare tra loro, tra cui:

- **Scambio di informazioni**: i processi devono scambiarsi informazioni per cooperare tra loro.
- **Accelerazione del calcolo**: i processi possono cooperare per eseguire un calcolo più velocemente.
- **Modularità**: i processi possono essere scritti in modo indipendente e comunicare tra loro per cooperare.
- **Convenienza**: è più semplice scrivere processi separati che cooperano tra loro piuttosto che scrivere un unico processo.

per ottenere una comunicazione tra processi è necessario che i processi condividano un canale di comunicazione, esistono due tipi di canali di comunicazione:

scambio di messaggi i processi comunicano scambiandosi messaggi che vengono inviati attraverso un canale di comunicazione tra il *kernel* e i processi, i messaggi possono essere inviati in modo sincrono o asincrono.

memoria condivisa i processi comunicano condividendo una regione di memoria, i processi possono leggere e scrivere nella memoria condivisa, la memoria condivisa è un canale di comunicazione molto più veloce rispetto allo scambio di messaggi, ma è più difficile da gestire.

Il primo risulta più sicuro in quanto i processi non possono accedere direttamente alla memoria degli altri processi ed il messaggio viene verificato dal *kernel* prima di essere inviato, mentre il secondo è più veloce in quanto non richiede l'intervento del *kernel* per la comunicazione.
Tutti i meccanismi di comunicazione tra processi sono implementati dal *kernel* del sistema operativo racchiusi nei protocolli di comunicazione tra processi (`IPC` - *Inter-Process Communication*).

5.1 `IPC` - *Message Passing*

Il protocollo `ICP` racchiude un insieme di meccanismi che permettono la comunicazione tra processi, tra i quali vi è il *message passing*, ovvero un meccanismo che permette ai processi di comunicare scambiandosi messaggi e senza condividere delle variabili e/o memoria. Le operazioni di base che ogni `SO` deve fornire per il *message passing* sono:

- `send` - invia un messaggio ad un processo. (con lunghezza fissa o variabile)
- `receive` - riceve un messaggio da un processo.

Prima ancora che i processi possano comunicare tra loro è necessario che essi siano in grado di identificarsi e stabilire un canale di comunicazione, per fare ciò è necessario che i processi abbiano un identificativo univoco, ovvero un *PID* (*Process IDentifier*).
L'implementazione di questo canale di comunicazione può essere realizzata in due modi:

livello fisico i messaggi vengono inviati attraverso un canale di comunicazione fisico, come ad esempio una rete o un bus.

livello logico i messaggi vengono inviati attraverso un canale di comunicazione logico, come ad esempio una coda di messaggi.

Le scelte di uno o dell'altro canale di comunicazione dipendono dalle esigenze del sistema e dalle prestazioni richieste. Fattori che influenzano la scelta sono:

- Come vengono stabiliti i canali
- Se un canale può essere utilizzato da più processi contemporaneamente
- Quanti canali possono essere aperti contemporaneamente tra una stessa coppia di processi
- La lunghezza massima del canale
- La lunghezza (fissa/variabile) massima dei messaggi
- Se il canale è *simplex*, *half-duplex* o *full-duplex*

5.1.1 Nominazione

A livello di nominazione, ovvero come i processi si identificano, esiste la comunicazione diretta e la comunicazione indiretta:

Comunicazione Diretta

Nella comunicazione diretta i processi si identificano direttamente, ovvero il mittente conosce l'identificativo del destinatario e viceversa, in questo modo il mittente può inviare il messaggio direttamente al destinatario. Questo metodo è molto veloce, ma presenta dei problemi:

- Il mittente deve conoscere l'identificativo del destinatario
- Il destinatario deve essere in esecuzione
- Nel caso in cui il destinatario o il ricevente cambi identificativo, il mittente deve essere aggiornato

La comunicazione diretta può a sua volta essere simmetrica o asimmetrica:

Simmetrica Il mittente e il destinatario si conoscono a priori e possono comunicare tra loro. Sia per l'invio che per la ricezione dei messaggi è necessario conoscere l'identificativo del processo con cui si vuole comunicare.

Asimmetrica Il mittente e il destinatario non si conoscono a priori. Solo il mittente conosce l'identificativo del destinatario, il destinatario non conosce l'identificativo del mittente ed ascolta qualsiasi messaggio che arriva.

Comunicazione Indiretta

Nella comunicazione indiretta i messaggi vengono inviati ad un canale di comunicazione comune detto *mailbox* (o porte), ognuna di queste *mailbox* ha associato un numero identificativo univoco, i processi possono inviare e ricevere messaggi da queste *mailbox* senza dover conoscere l'identificativo del destinatario, ma devono condividere una *mailbox* comune.

Flusso di una *mailbox* Prima di poter inviare un messaggio ad una *mailbox* è necessario che questa venga creata, una volta creata la *mailbox* il processo può inviare un messaggio ad essa, il messaggio viene inserito in una coda di messaggi associata alla *mailbox*, il destinatario può ricevere il messaggio dalla *mailbox* e leggerlo, una volta letto il messaggio viene rimosso dalla coda. Quando la *mailbox* non è più necessaria può essere eliminata.

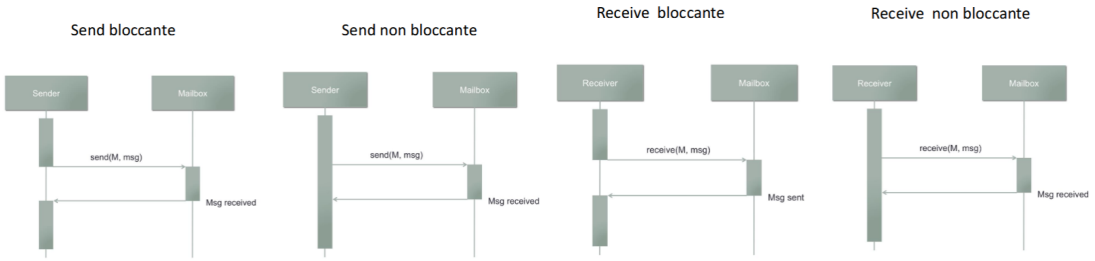
Invio e ricezione Per inviare e ricevere messaggi da una *mailbox* è necessario conoscere l'identificativo della *mailbox*, una volta conosciuto l'identificativo il processo può inviare e ricevere messaggi dalla *mailbox*.

Proprietà del canale Come già detto, il canale di comunicazione viene stabilito solo se i processi condividono una *mailbox*, ma una *mailbox* può essere associata a molti processi ed una stessa coppia di processi può avere più *mailbox* associate, inoltre una *mailbox* può essere o meno bi-direzionale.

Problema riceventi multipli Un problema che si può presentare è quello dei riceventi multipli, ovvero quando un mittente invia un messaggio ad una *mailbox* e ci sono più processi che ricevono i messaggi da quella *mailbox*, in questo caso il *SO* deve permettere solo ad uno dei processi di ricevere il messaggio, e questo viene fatto in maniera arbitraria.

5.1.2 Sincronizzazione

Uno scambio di messaggi può essere "bloccante" (sincrono) o "non bloccante" (asincrono), ovvero il mittente può continuare ad eseguire il proprio codice dopo aver inviato il messaggio o deve attendere che il destinatario riceva il messaggio.
Se il canale di comunicazione è bloccante, il mittente deve attendere che il destinatario riceva il messaggio ed assicurarsi che il messaggio sia stato ricevuto, se il canale di comunicazione è non bloccante il mittente può continuare ad eseguire il proprio codice dopo aver inviato il messaggio, senza dover attendere che il destinatario riceva il messaggio, in questo caso il mittente non può sapere se il messaggio è stato ricevuto o meno.



(a) Invio bloccante/non bloccante (b) Ricezione bloccante/non bloccante

Figura 5.1: Invio e ricezione bloccante/non bloccante

5.2 IPC - Memoria Condivisa

Un altro meccanismo di comunicazione tra processi è la memoria condivisa, ovvero un'area di memoria condivisa tra più processi, i processi possono leggere e scrivere nella memoria condivisa, la memoria condivisa è un canale di comunicazione molto più veloce rispetto allo scambio di messaggi, ma è più difficile da gestire, inoltre il *kernel* non può controllare l'accesso alla memoria condivisa, quindi è necessario che i processi si sincronizzino tra loro per evitare problemi di accesso concorrente.

Flusso di POSIX

Prendiamo come esempio la memoria condivisa in POSIX, per poter utilizzare la memoria condivisa è necessario che uno dei processi crei la memoria condivisa, una volta creata la memoria condivisa l'altro processo deve "attaccarsi" al segmento di memoria condivisa, una volta "attaccato" il processo può avere il permesso di leggere e scrivere oppure solo di leggere, una volta terminato il processo deve "staccarsi" dalla memoria condivisa. Il processo che ha creato la memoria condivisa deve rimuoverla una volta terminato.

Le *pipe*

Un altro meccanismo di comunicazione tra processi è la *pipe*, ovvero un canale di comunicazione tra processi.

La *pipe* è un canale di comunicazione che permette di inviare e ricevere messaggi tra processi, la *pipe* è un canale di comunicazione unidirezionale, ovvero i messaggi possono essere inviati in una sola direzione, ma è possibile creare due *pipe* per permettere la comunicazione in entrambe le direzioni. Distinguiamo tra *pipe* ordinarie e *pipe* con nome:

***Pipe* ordinarie** Le *pipe* ordinarie permettono la comunicazione in uno stile "Produttore-Consumatore" dove il produttore scrive nella *pipe* e il consumatore legge dalla *pipe*, questa tipologia richiede una relazione tra processi, queste infatti possono essere aperte solo da processi padri verso i processi figli.

***Pipe* con nome** Le *pipe* con nome sono simili alle *pipe* ordinarie, ma permettono la comunicazione bidirezionale, non è richiesta la relazione tra processi e più processi possono usare la stessa *pipe* per comunicare tra loro, ma è necessario che i processi si sincronizzino tra loro per evitare problemi di accesso concorrente. Le *pipe* con nome sono disponibili sia in sistemi `UNIX` che in sistemi `Windows`, ma in quest'ultimo caso non sono implementate come file di tipo *FIFO* ma come file temporanei.

Capitolo 6

Scheduling della CPU

Andremo an analizzare in questo capitolo lo *Scheduling* della CPU, ovvero il modo in cui il sistema operativo decide quale processo eseguire in un dato momento. Lo *Scheduling* è una parte fondamentale del sistema operativo, poiché influisce direttamente sulle prestazioni e sull'efficienza del sistema. Distingueremo inoltre i vari tipi di *Scheduling* (breve, medio e lungo termine) e i vari algoritmi di *Scheduling* (FIFO, SJF, Round Robin, ecc.).

6.1 Concetto di *Scheduling*

Lo *Scheduling* è il processo di assegnazione di attività nel tempo, l'uso della multiprogrammazione permette di eseguire più processi in parallelo, ma il sistema operativo deve decidere quale processo eseguire in un dato momento, visto che la CPU può eseguire solo un processo alla volta. Bisogna quindi decretare se un programma può essere ammesso nella memoria e quale processo deve essere eseguito in un dato momento.
Come visto nella sezione 4.1.1, un processo può trovarsi in uno dei seguenti stati:

- **New**: il processo è stato creato, ma non è ancora pronto per essere eseguito.
- **Ready**: il processo è pronto per essere eseguito, ma non ha ancora ottenuto l'accesso alla CPU.
- **Running**: il processo sta attualmente eseguendo sulla CPU.
- **Waiting**: il processo è in attesa di un evento esterno (ad esempio, l'input dell'utente o la disponibilità di una risorsa).
- **Terminated**: il processo ha completato la sua esecuzione e sta per essere rimosso dalla memoria.

Inoltre esistono diverse code di Ready e di Waiting, a seconda del tipo di processo. Ad esempio, i processi in attesa di I/O potrebbero essere in una coda separata rispetto ai processi in attesa di un semaforo.

Implementazione delle Code

A livello pratico le code di Ready e di Waiting sono implementate come liste collegate (*linked list*) o come array.
Ogni coda ha un *queue header* che contiene informazioni sulla coda stessa, come il puntatore al primo elemento della coda ed il puntatore all'ultimo elemento della coda. Ogni processo ha un *process control block* (PCB) che contiene informazioni sul processo stesso, come il suo stato, il contenuto dei registri quando il processo era in esecuzione, il puntatore al prossimo processo da eseguire, ecc...
Una coda può essere o la coda di *ready*, dove i processi pronti per essere eseguiti sono in attesa di essere assegnati alla CPU, oppure una delle code di *waiting*, dove i processi sono in attesa di un evento esterno ed ogni coda rappresenta un evento diverso.

6.2 Tipi di *Scheduling*

Per la gestione dei processi si possono distinguere tre tipi di *scheduling*, di cui due principali ed uno secondario:

- ***Long-term scheduling*** - Pianificazione a lungo termine
- ***Short-term scheduling*** - Pianificazione a breve termine
- ***Medium-term scheduling*** - Pianificazione a medio termine

Pianificazione a lungo termine - (*job-scheduler*)

La pianificazione a lungo termine è il processo di selezione dei processi da ammettere nella memoria principale. Questo tipo di pianificazione è responsabile della creazione di nuovi processi e della loro ammissione nella memoria, e dunque nella coda *ready*, per essere eseguiti. La pianificazione a lungo termine determina il grado di multiprogrammazione del sistema, ovvero il numero di processi che possono essere eseguiti contemporaneamente. Se il grado di multiprogrammazione è troppo alto, il sistema potrebbe diventare instabile e i processi potrebbero non ricevere le risorse necessarie per essere eseguiti. Se il grado di multiprogrammazione è troppo basso, la CPU potrebbe rimanere inattiva per lunghi periodi di tempo, riducendo l'efficienza del sistema. Inoltre il *long-term scheduler* è responsabile della determinazione del tipo di processo da eseguire, ovvero se questo è CPU *bound* oppure I/O *bound*.

Frequenza La pianificazione a lungo termine viene eseguita con frequenza dell'ordine del secondo o di pochi secondi, poiché la creazione di nuovi processi e la loro ammissione nella memoria sono operazioni relativamente costose in termini di tempo e risorse.
Questo sistema è opzionale e può essere assente.

Pianificazione a breve termine - *CPU-scheduler*

La pianificazione a breve termine è il processo di selezione del processo da eseguire sulla CPU in un dato momento. Questo tipo di pianificazione è responsabile della gestione dei processi in esecuzione e della loro assegnazione alla CPU. La pianificazione a breve termine determina quale processo deve essere eseguito in un dato momento, in base a diversi criteri, come la priorità del processo, il tempo di attesa e il tempo di completamento. La pianificazione a breve termine è responsabile della gestione della CPU e della sua assegnazione ai processi in esecuzione.

Frequenza La pianificazione a breve termine viene eseguita con frequenza dell'ordine dei millisecondi, dunque deve essere una operazione molto veloce, infatti se il tempo di processo è $100ms$ ed il tempo di *scheduling* è $10ms$, il tempo di *scheduling* incide per il 9% sul tempo totale di esecuzione del processo. Se il tempo di *scheduling* è troppo alto, il sistema potrebbe diventare instabile e i processi potrebbero non ricevere le risorse necessarie per essere eseguiti. Se il tempo di *scheduling* è troppo basso, la CPU potrebbe rimanere inattiva per lunghi periodi di tempo, riducendo l'efficienza del sistema.
Questo sistema è sempre presente e non può essere assente, poiché è necessario per la gestione della CPU e dei processi in esecuzione.

Pianificazione a medio termine - *medium-term scheduler*

La pianificazione a medio termine è un processo intermedio tra la pianificazione a lungo termine e la pianificazione a breve termine. Questo è presente se e solo se il sistema operativo supporta la *swapping*, ovvero il trasferimento di processi dalla memoria principale alla memoria secondaria (disco) e viceversa. La porzione di disco usata per questo scopo è chiamata *swap space* ed sostanzialmente è della RAM virtuale che viene usata per memorizzare i processi che non sono attualmente in esecuzione. La pianificazione a medio termine è responsabile della gestione della memoria e della sua assegnazione ai processi in esecuzione. Questo tipo di pianificazione è responsabile dell'immagazzinamento dei processi nella memoria secondaria e del loro trasferimento nella memoria principale quando necessario. Questo processo viene eseguito con tutti i processi che escono dalla CPU per rientrare nella *ready queue* ma non può avvenire se il processo esce dalla CPU per inserirsi nella *waiting queue*.

6.3 Scheduling della CPU

Scheduler Lo *scheduler* della CPU è, a livello logico, il modulo del SO che decide quale processo eseguire in un dato momento, vista la frequenza di chiamate a funzioni di *Scheduling* e la velocità con cui i processi passano da uno stato all'altro, lo *scheduler* deve essere molto veloce.

Dispatcher Il *dispatcher* è il modulo del SO che effettivamente esegue il passaggio di controllo tra i processi, ovvero il passaggio da un processo all'altro. Il *dispatcher* è responsabile di:

- *Switch* del contesto: salva il contesto del processo corrente e carica il contesto del processo successivo.
- Passaggio alla modalità utente: il SO deve passare dalla modalità kernel alla modalità utente per eseguire il processo.
- Salto alla opportuna locazione nel codice: il `dispatcher` deve saltare alla locazione corretta nel codice del processo.

La latenza di un *dispatcher* consiste nel tempo necessario per eseguire queste operazioni, ovvero fermare il processo corrente e passare al successivo. La latenza del *dispatcher* è molto importante, poiché influisce sulle prestazioni del sistema. Un *dispatcher* veloce può migliorare le prestazioni del sistema, mentre un *dispatcher* lento può causare un degrado delle prestazioni.

Modello astratto del sistama

Quando parliamo di un processo a livello astratto consideriamo che questo possa essere o in CPU *burst* oppure in I/O *burst*

Distribuzione dei CPU *burst* Solitamente i processi hanno una distribuzione dei CPU *burst* che segue una distribuzione esponenziale, ovvero la maggior parte dei processi ha un CPU *burst* breve, mentre pochi processi hanno un CPU *burst* lungo. Questo è dovuto al fatto che i processi brevi sono più comuni rispetto ai processi lunghi. Per questo motivo è stato implementato il processo di prelazione (*preemption*)

Prelazione Come detto in precedenza, i processi brevi sono più comuni rispetto ai processi lunghi, per questo motivo è stato implementato il processo di prelazione (*preemption*), ovvero la possibilità di interrompere un processo in esecuzione per dare la precedenza ad un altro processo. Esistono dunque in circolazione due tipi di *scheduler*:

- **Non preemptive**: il processo in esecuzione non può essere interrotto, ma deve terminare la sua esecuzione prima di passare al successivo.
- **Preemptive**: il processo in esecuzione può essere interrotto in qualsiasi momento per dare la precedenza ad un altro processo.

La prelazione è utile per garantire che i processi brevi vengano eseguiti il prima possibile, evitando che i processi lunghi occupino la CPU per troppo tempo. Tuttavia, la prelazione può anche causare un aumento della latenza del *dispatcher*, poiché il *dispatcher* deve eseguire il passaggio di controllo tra i processi più frequentemente.

Metriche di *scheduling*

Esistono diverse metriche sulle quali scegliere un algoritmo di *scheduling* piuttosto che un altro, le più comuni sono:

- **Utilizzo della CPU** (*CPU Utilization*): percentuale di tempo in cui la CPU è occupata ad eseguire processi. Un utilizzo della CPU del 100% è l'ideale, ma è difficile da raggiungere.
- ***Throughput***: numero di processi completati in un dato intervallo di tempo. Un *throughput* elevato è desiderabile, poiché indica che il sistema sta eseguendo molti processi in un breve periodo di tempo.
- **Tempo di attesa** (*Waiting Time*): tempo medio che un processo trascorre in attesa di essere eseguito. Un tempo di attesa basso è desiderabile, poiché indica che i processi vengono eseguiti rapidamente.
- **Tempo di completamento** (*Turnaround Time*): tempo medio che un processo trascorre nel sistema, dalla sua creazione alla sua terminazione. Un tempo di completamento basso è desiderabile, poiché indica che i processi vengono eseguiti rapidamente.

- **Tempo di risposta** (*Response Time*): tempo medio che intercorre tra l'invio di una richiesta e la ricezione della risposta. Un tempo di risposta basso è desiderabile, poiché indica che il sistema risponde rapidamente alle richieste degli utenti.

Il compito di un algoritmo di *scheduling* è quello di massimizzare l'utilizzo della CPU e il *throughput*, minimizzando il tempo di attesa, il tempo di completamento e il tempo di risposta. Tuttavia, non è sempre possibile ottimizzare tutte queste metriche contemporaneamente, poiché spesso ci sono compromessi tra di esse. Ad esempio, un algoritmo che massimizza l'utilizzo della CPU potrebbe aumentare il tempo di attesa dei processi, mentre un algoritmo che minimizza il tempo di attesa potrebbe ridurre l'utilizzo della CPU.

6.3.1 Algoritmi di *scheduling*

Andiamo ora ad analizzare i vari algoritmi di *scheduling* della CPU, partendo da quelli più semplici e passando a quelli più complessi.

First-Come, First-Served (FCFS)

L'algoritmo FCFS è il più semplice degli algoritmi di *scheduling*, i processi vengono eseguiti nell'ordine in cui arrivano nella coda di Ready. Questo algoritmo è semplice da implementare e non richiede alcun calcolo complesso. Tuttavia, ha alcuni svantaggi:

- Non tiene conto della lunghezza dei processi, quindi i processi lunghi possono bloccare l'esecuzione dei processi brevi.
- Può causare un aumento del tempo di attesa e del tempo di completamento per i processi brevi.

L'algoritmo FCFS è un algoritmo non preemptive, poiché un processo in esecuzione non può essere interrotto fino al suo completamento. Questo può portare a una bassa efficienza del sistema, poiché i processi brevi possono rimanere in attesa per lungo tempo.

Esempio consideriamo questi processi:

Processo	**Tempo di arrivo**	***CPU burst***
P1	0	24
P2	2	3
P3	4	3

Allora i tempi di attesa, completamento e di risposta sono:

Processo	T_r	T_w	T_t
P1	0	0	24
P2	24	22	25
P3	27	23	30

Dunque il tempo medio di attesa è:

$$T_{w,med} = \frac{T_{w,P1} + T_{w,P2} + T_{w,P3}}{3} = \frac{0 + 22 + 23}{3} = 15$$

Il tempo medio di completamento è:

$$T_{t,med} = \frac{T_{t,P1} + T_{t,P2} + T_{t,P3}}{3} = \frac{24 + 25 + 30}{3} = 26$$

Se però cambiano i tempi di arrivo dei processi, ad esempio:

Processo	**Tempo di arrivo**	***CPU burst***
P1	4	24
P2	0	3
P3	2	3

Allora i tempi di attesa, completamento e di risposta sono:

Processo	T_r	T_w	T_t
P1	2	2	26
P2	0	0	3
P3	1	1	4

Dunque il tempo medio di attesa è:

$$T_{w,med} = \frac{T_{w,P1} + T_{w,P2} + T_{w,P3}}{3} = \frac{2+0+1}{3} = 1$$

Il che è molto più veloce rispetto al caso precedente, nonostante il processo P1 sia più lungo. Questo è dovuto al fatto che i processi brevi sono stati eseguiti prima di P1, riducendo il tempo di attesa per P1.

Shortest Job First (SJF)

L'algoritmo SJF è un algoritmo di *scheduling* che assegna la CPU al processo con il *CPU burst* più breve. Questo algoritmo è in grado di ridurre il tempo di attesa e il tempo di completamento dei processi, poiché i processi brevi vengono eseguiti per primi. Questo algoritmo può essere implementato sia in modo *preemptive* che in modo non *preemptive*. Se è implementato in modo *preemptive*, il processo in esecuzione può essere interrotto se arriva un processo con un *CPU burst* più breve rispetto al CPU *burst* rimanente del processo in esecuzione (*Shortest-Remaining-Time-First* - SRTF). Se è implementato in modo non *preemptive*, il processo in esecuzione non può essere interrotto fino al suo completamento.

Esempio consideriamo questi processi:

Processo	**Tempo di arrivo**	***CPU burst***
P1	0	7
P2	2	4
P3	4	1
P4	5	4

Allora i processi verranno eseguiti in questo ordine:

- P1 (0-7)
- P3 (7-8)
- P2 (8-12)
- P4 (12-16)

Creando quindi i seguenti tempi di attesa, completamento e di risposta:

Processo	T_r	T_w	T_t
P1	0	0	7
P2	6	6	10
P3	3	3	4
P4	7	7	11

Dunque il tempo medio di attesa è:

$$T_{w,med} = \frac{T_{w,P1} + T_{w,P2} + T_{w,P3} + T_{w,P4}}{4} = \frac{0+6+3+7}{4} = 4$$

Se gli stessi processi arrivano nello stesso ordine ma in un sistema *preemptive*, i processi verranno eseguiti in questo ordine:

- P1 (0-2) - 5 rimasti
- P2 (2-4) - 2 rimasti

- P3 (4-5)
- P2 (5-7)
- P4 (7-11)
- P1 (11-16)

Creando quindi i seguenti tempi di attesa, completamento e di risposta:

Processo	T_r	T_w	T_t
P1	0	0	16
P2	5	5	10
P3	4	4	5
P4	7	7	11

Dunque il tempo medio di attesa è:

$$T_{w,med} = \frac{T_{w,P1} + T_{w,P2} + T_{w,P3} + T_{w,P4}}{4} = \frac{0 + 5 + 4 + 7}{4} = 4$$

Il che è lo stesso del caso non *preemptive*, ma in questo caso il tempo di attesa è più basso per i processi brevi, mentre il tempo di attesa per i processi lunghi è più alto.
Il principale problema di questo algoritmo è quello che è impossibile determinare con precisione il *CPU burst* di un processo, poiché questo dipende da molti fattori esterni. Viene dunque usata una media esponenziale (*exponential average*) per stimare il *CPU burst* di un processo. La media esponenziale è una media che dà più peso ai valori recenti rispetto ai valori più vecchi. La formula per calcolare la media esponenziale è:

$$T_{n+1} = \alpha T_n + (1 - \alpha) T_{n-1}$$

dove:

- T_{n+1} è il nuovo valore della media esponenziale
- T_n è il valore corrente del *CPU burst*
- T_{n-1} è il valore precedente del *CPU burst*
- α è un valore compreso tra 0 e 1 che determina il peso dei valori recenti rispetto ai valori più vecchi. Un valore di α vicino a 1 dà più peso ai valori recenti, mentre un valore di α vicino a 0 dà più peso ai valori più vecchi.

Scheduling a priorità

Nello *scheduling* a priorità ad ogni processo viene associata una priorità, i processi con priorità più alta vengono eseguiti per primi. Questo algoritmo può essere implementato sia in modo *preemptive* che in modo non *preemptive*. Un esempio di *scheduling* con priorità è il comando `nice` di `Linux`, che permette di modificare la priorità di un processo.

Politiche di assegnamento priorità L'assegnamento di un livello di priorità rispetto ad un altro può essere influenzato da fattori interni o esterni al `SO`. Come fattori interni troviamo ad esempio: limiti di tempo, requisiti di memoria, numero di file richiesti, ...
Possono anche essere influenzati da fattori esterni, come ad esempio l'importanza (umana) del processo, se per quel determinato processo si ha un guadagno economico o anche motivi politici, ...

Problemi Il principale problema dei `SO` con *scheduling* a priorità è la *starvation*, ovvero certi processi a bassa priorità potrebbero non essere mai eseguiti in quanto ci sono sempre processi con priorità più alta in attesa di essere eseguiti. Questo problema può essere risolto utilizzando una tecnica chiamata *aging*, che consiste nell'aumentare gradualmente la priorità dei processi a bassa priorità man mano che trascorrono del tempo in attesa. In questo modo, anche i processi a bassa priorità avranno la possibilità di essere eseguiti, evitando la *starvation*.

***Higher Response Ratio Next* - HRRN** L'HRRN è un algoritmo di *scheduling* a priorità, sempre non-*preentive*. In questo algoritmo la priorità di un processo viene calcolata in base al suo tempo di attesa e al suo *CPU burst*. La formula per calcolare la priorità di un processo è:

$$R = \frac{T_w + T_{CPU}}{T_{CPU}}$$

il maggiore valore di R avrà la priorità più alta. Questo algoritmo è in grado di ridurre il tempo di attesa e il tempo di completamento dei processi, poiché i processi brevi vengono eseguiti per primi. Inoltre se un processo è in attesa per lungo tempo, la sua priorità aumenta, evitando la *starvation* e la priorità è dunque dinamica. Se alla fine di un processo è arrivato un altro processo allora la priorità dei processi in attesa viene ricalcolata, andando a modificare l'ordine di esecuzione dei processi se necessario, oppure può anche essere ricalcolata alla fine di ogni CPU *burst* indipendentemente dal fatto che sia arrivato un nuovo processo o meno (dipende dalle implementazioni).

Esempio consideriamo questi processi:

Processo	**Tempo di arrivo**	***CPU burst***
P1	1	10
P2	0	2
P3	2	2
P4	2	1
P5	1	5

Allora il calcolo della priorità dei processi è il seguente:

Processo	$t = 0$	$t = 2$	$t = 7$	$t = 8$
P1	-	1+1/10	1+6/10	1+7/10
P2	1	-	-	-
P3	-	1 + 0/2	1 + 5/2	1 + 6/2
P4	-	1 + 0/1	1 + 5/1	-
P5	-	1 + 1/5	-	-

[1]

***Round Robin* (RR)**

L'algoritmo RR è un algoritmo *scheduling preemptive* che assegna un piccolo tempo di CPU chiamato *quantum* ($10 - 100$ ms) ad ogni processo in coda. Quando il tempo di CPU di un processo scade, il processo viene interrotto e messo in coda, e il *dispatcher* passa al processo successivo. Questo algoritmo è in grado di garantire una buona risposta per i processi interattivi, poiché i processi brevi vengono eseguiti rapidamente. Tuttavia bisogna fare una scelta sul valore del *quantum*, se questo è molto grande allora è esattamente come l'algoritmo *First Come First Serve*, se invece è molto piccolo allora il tempo di *scheduling* aumenta, poiché ad ogni passaggio di processo il *dispatcher* deve eseguire il *context switch* e passare alla modalità utente, il valore ottimale per il tempo q deve essere scelto in modo che il tempo q sia minore dell'80% dei *CPU burst*

Quanto-*context-switch* Esiste una relazione direttamente inversa tra il numero di *context switch* e il tempo di CPU *burst*, ovvero più è lungo il *CPU burst* e meno *context switch* ci sono. Questo è dovuto al fatto che ogni volta che si esegue un *context switch* il sistema deve salvare lo stato del processo corrente e caricare lo stato del processo successivo, il che richiede tempo e risorse. Se i processi hanno un *CPU burst* lungo, ci saranno meno *context switch* e quindi meno tempo sprecato.

Quanto-Tempo di attesa A differenza del numero di *context-switch*, il tempo di attesa non è legato direttamente al tempo di *quantum*, ma è legato al numero di processi in coda e ai tempi di esecuzione dei processi stessi. Se ci sono molti processi in coda, il tempo di attesa per ciascun processo aumenta, poiché

[1] Nota come nel seguente esempio il calcolo della priorità è stato fatto anche per i tempi $t = 7$ e $t = 8$, anche se non sono arrivati nuovi processi.

ogni processo deve attendere il completamento del *quantum* degli altri processi prima di essere eseguito nuovamente. Tuttavia, un *quantum* troppo piccolo può aumentare il tempo di attesa complessivo a causa dell'overhead introdotto dai frequenti *context-switch*.

Code multi-livello

Le code multi-livello sono una suddivisione della *ready queue* in più code, ognuna con un ruolo specifico ed un algoritmo di *scheduling* specifico. Ad esempio si potrebbe avere una coda per i processi interattivi, una coda per i processi batch e una coda per i processi in background, queste potrebbero essere gestite con RR, FCFS e SJF rispettivamente. In questo modo si può garantire una buona risposta per i processi interattivi e una buona efficienza per i processi batch. Questo però al costo di dover gestire lo *scheduling* tra le varie code. Quest'ultimo solitamente è gestito o come *time slice*, ovvero ogni coda ha una percentuale di tempo di CPU da utilizzare, oppure tramite priorità fissa, ovvero ogni coda ha una priorità fissa e la coda con priorità più alta viene eseguita, e liberata, prima delle altre.

Code multi-livello a *feedback* Mentre in una coda multi-livello tradizionale quando un processo viene assegnato ad una coda specifica non può cambiare coda, in una coda multi-livello a *feedback* i processi possono cambiare coda sulla base delle proprie caratteristiche di esecuzione, spesso ciò viene fatto per implementare l'*aging* e per evitare la *starvation*. I parametri sui quali si può agire per la configurazione dello *scheduler* sono: il numero delle code, l'algoritmo usato, i criteri di promozione e/o retrocessione dei processi ed i criteri per definire la coda iniziale di un processo.

Scheduling fair share

Dato che un programmatore sapendo che in alcuni casi se pianifica molti *threads* allora il suo processo potrebbe ottenere più tempo di CPU rispetto ad un altro processo, è stato implementato lo *scheduling fair share*, che lavora per applicazione e non per processo. Ad ogni applicazione viene assegnato una percentuale del tempo di CPU e i processi di quell'applicazione possono usare solo quella percentuale del tempo di CPU. Questo algoritmo è in grado di garantire che ogni applicazione riceva una quantità equa di tempo di CPU, evitando che un'applicazione monopolizzi le risorse del sistema. Tuttavia, questo algoritmo può causare un aumento del tempo di attesa per i processi ed un aumento del tempo di completamento, poiché i processi devono attendere che il loro turno arrivi. Inoltre, questo algoritmo può essere più complesso da implementare rispetto ad altri algoritmi di *scheduling*, poiché richiede la gestione delle percentuali di tempo di CPU per ogni applicazione e la loro assegnazione ai processi.

Contesto reale

Solitamente nei sistemi operativi moderni viene usato un mix di algoritmi di *scheduling*, ad esempio il Linux usa un algoritmo *completely fair scheduler* (CFS) che è una combinazione di RR e SJF. Questo algoritmo è in grado di garantire una buona risposta per i processi interattivi e una buona efficienza per i processi batch. Inoltre, questo algoritmo è in grado di adattarsi alle diverse condizioni del sistema, modificando dinamicamente le priorità dei processi in base al loro comportamento.

6.3.2 Valutazione degli algoritmi

Esistono diversi metodi per valutare le prestazioni degli algoritmi di *scheduling*, noi affrontiamo il modello deterministico ed il modello a reti di code.

Modello deterministico (Analitico)

Il modello deterministico è un metodo di valutazione degli algoritmi di *scheduling* che si basa su un insieme di processi con tempi di arrivo e tempi di esecuzione noti.[2] Questo metodo consente di calcolare le prestazioni degli algoritmi di *scheduling* in modo preciso e dettagliato, poiché i tempi di arrivo e i tempi di esecuzione sono noti in anticipo. Tuttavia, questo metodo non tiene conto delle variazioni nei tempi di arrivo e nei tempi di esecuzione dei processi, il che può portare a risultati poco realistici. Inoltre, questo metodo richiede una conoscenza dettagliata dei processi e delle loro caratteristiche, il che può essere difficile da ottenere in un sistema reale. Per questo motivo, il modello deterministico è più adatto per la valutazione di algoritmi di *scheduling* in ambienti controllati e non in ambienti reali.

[2] Come è stato fatto in precedenza con i vari esempi.

Modello a reti di code

Il modello a reti di code è un metodo di valutazione degli algoritmi di *scheduling* che si basa su un preciso numero di processi sempre uguali ed con tempi di `CPU` *burst*, `I/O` *burst* ed arrivo basati su una distribuzione di Poisson della quale si varia il parametro λ. Da questo modello è possibile calcolare il tempo medio di attesa, il tempo medio di completamento e il tempo medio di risposta per ogni algoritmo di *scheduling*. Questo metodo consente di valutare le prestazioni degli algoritmi di *scheduling* in modo più realistico rispetto al modello deterministico, poiché tiene conto delle variazioni nei tempi di arrivo e nei tempi di esecuzione dei processi. Tuttavia, questo metodo richiede una conoscenza dettagliata delle distribuzioni dei processi e delle loro caratteristiche, il che può essere difficile da ottenere in un sistema reale. Per questo motivo, il modello a reti di code è più adatto per la valutazione di algoritmi di *scheduling* in ambienti reali e non in ambienti controllati.

Simulazione

La simulazione è un metodo di valutazione degli algoritmi di *scheduling* che si basa sull'esecuzione di un insieme di processi in un ambiente simulato. Questo metodo consente di valutare le prestazioni degli algoritmi di *scheduling* in modo realistico, poiché tiene conto delle variazioni nei tempi di arrivo e nei tempi di esecuzione dei processi. Inoltre, questo metodo consente di testare gli algoritmi di *scheduling* in condizioni controllate ma pseudo-reali, il che può essere utile per la valutazione di algoritmi di *scheduling* in ambienti reali. Tuttavia, questo metodo richiede che il modello del sistema sia già disponibile, inoltre il suo uso seppur garantendo una buona valutazione delle prestazioni degli algoritmi di *scheduling* può essere costoso in termini di tempo e risorse.

Implementazione

L'implementazione di un algoritmo di *scheduling* è l'unico metodo certo per valutare le prestazioni di un algoritmo di *scheduling* in un sistema reale. Questo metodo consente di testare gli algoritmi di *scheduling* in condizioni reali e concrete e di valutare le loro prestazioni in un ambiente reale. Tuttavia, questo metodo richiede che questo algoritmo sia già codificato, inserito nel `SO` e solo dopo è possibile verificarne l'effettiva efficienza. Inoltre, questo metodo può essere costoso in termini di tempo e risorse, poiché richiede la modifica del `SO` e la sua re-installazione. Tuttavia, l'implementazione di un algoritmo di *scheduling* è il metodo più preciso e affidabile per valutare le prestazioni degli algoritmi di *scheduling* in un sistema reale.

Capitolo 7

Sincronizzazione dei processi

In questo capitolo andremo ad affrontare come gestire la sincronizzazione dei processi, ovvero come evitare che più processi accedano contemporaneamente a risorse condivise, causando inconsistenze nei dati o comportamenti imprevisti. La sincronizzazione è fondamentale in un sistema operativo per garantire che le operazioni sui dati condivisi siano eseguite in modo sicuro e prevedibile. Si parlerà di varie primitive di sincronizzazione ancora più complesse del meccanismi di `join` e `fork` già visti in precedenza. In particolare, ci concentreremo su mutex, semafori e variabili di condizione. Questi strumenti sono essenziali per la programmazione concorrente e ci permettono di gestire l'accesso alle risorse condivise in modo sicuro e controllato. Inoltre, esploreremo le problematiche legate alla sincronizzazione, come il deadlock e la starvation, e come evitarle attraverso tecniche di progettazione adeguate.

Modello astratto Il modello astratto di un processo è quello di produttore-consumatore dove un processo produce dati e un altro li consuma. Deve essere quindi garantita l'esecuzione concorrente di più processi, in modo che il produttore possa aggiungere ad un *buffer* condiviso e il consumatore possa prelevare dati da esso contemporaneamente. Questo *buffer* ha comunque dei vincoli, non deve essere permessa la scrittura se questo è pieno e se è vuoto non deve essere permesso il prelievo.

Buffer* `P/C`: Modello *software

Il *buffer* viene visto in maniera circolare con due puntatori, `in` ed `out` dove, `in` punta alla prossima posizione libera e `out` punta alla prossima posizione da prelevare. Il *buffer* vuoto ha `in` == `out` e il *buffer* pieno ha `out` == $(\texttt{in} + 1)\%n$. Nel corso per semplicità usiamo un contatore `counter` che indica il numero di elementi presenti nel *buffer* e quindi il *buffer* è vuoto se `counter == 0` e pieno se `counter == n`. Dunque con l'uso del contatore il processo produttore aumenta il contatore di uno e il processo consumatore lo diminuisce di uno, il problema di ciò è che l'istruzione `counter++` e `counter--` vengono divise in tre istruzioni assembly differenti:

```
mov eax, [counter] ; carica il contatore in eax
add eax, 1 ; incrementa il contatore
mov [counter], eax ; salva il contatore
```

Se due processi eseguono in parallelo il contatore potrebbe essere incrementato due volte o decrementato due volte, portando a risultati errati. Per evitare questo problema è necessario utilizzare un meccanismo di sincronizzazione che garantisca l'accesso esclusivo alla variabile `counter` durante l'operazione di incremento o decremento. Abbiamo appena visto un esempio di sezione critica costituita dalla lettura e scrittura della variabile `counter`.

7.1 Problema della sezione critica

La sezione critica è una porzione di codice che accede a una risorsa condivisa e deve essere eseguita in modo esclusivo da un solo processo alla volta. Per garantire che solo un processo alla volta possa eseguire la sezione critica. La soluzione deve soddisfare le seguenti proprietà:

- **Mutua esclusione**: Solo un processo alla volta può essere nella sezione critica.

- **Progresso**: Se nessun processo è nella sezione critica e ci sono processi in attesa, uno di essi deve essere in grado di entrare nella sezione critica. La decisione non può essere rimandata indefinitamente.
- **Attesa limitata**: Deve esistere un numero massimo di volte per cui un processo può essere bloccato in attesa di entrare nella sezione critica. Non deve essere possibile che un processo rimanga in attesa indefinitamente.

Struttura generica di un processo

La struttura generica di un processo che accede a una sezione critica è la seguente:

```
while (true) {
    // Sezione non critica
    // ... codice non critico ...

    // Sezione di entrata
    // ... codice per entrare nella sezione critica ...
    // Sezione critica
    // Sezione di uscita
    // ... codice per uscire dalla sezione critica ...
    // Sezione non critica
}
```

La sezione di entrata è il codice che consente al processo di entrare nella sezione critica, mentre la sezione di uscita è il codice che consente al processo di uscire dalla sezione critica. La sezione non critica è il codice che può essere eseguito in parallelo con altri processi senza problemi di sincronizzazione.

7.1.1 Soluzioni al problema della sezione critica

Quando si prova a risolvere il problema della sezione critica, è importante considerare le varie soluzioni e i loro vantaggi e svantaggi. Assumiamo di prima istanza che la sincronizzazione sia in ambiente globale, ovvero che esistono celle di memoria condivise tra i processi. In questo caso, possiamo sfruttare delle soluzioni *software* le quali richiedono solo un aggiunta di codice alle applicazioni esistenti, ma ciò non sfrutta nessun supporto da parte dell'*hardware* e/o dal sistema operativo. Le soluzioni *hardware* invece richiedono un supporto da parte dell'*hardware* e/o del sistema operativo, ma richiedono molte meno modifiche al codice delle applicazioni. Le soluzioni *hardware* sono più veloci e più efficienti rispetto a quelle *software*, ma richiedono un maggiore sforzo di implementazione e possono essere più complesse da gestire.

Algoritmo 1

```
PROCESS i;
int turn; /* Se turn == i processo i entra nella sezione critica */
while(1){
    while(turn != i); /* Attesa attiva */
    // Sezione critica
    turn = j; /* Passa il turno al processo j */
    // Sezione non critica
}
```

In questo modo si garantisce che solo un processo alla volta possa entrare nella sezione critica. Tuttavia, questo algoritmo presenta alcuni problemi, infatti se uno dei due processi termina, l'altro processo rimarrà bloccato in attesa dopo che questo ha passato il suo turno (non viene rispettato il progresso). Inoltre, questo algoritmo richiede una stretta alternanza tra i processi, infatti finché entrambi i processi non vogliono entrare nella sezione critica, non è possibile che uno dei due possa entrare. Infine, questo algoritmo non è adatto per più di due processi, poiché richiede una variabile `turn` per ogni coppia di processi.

Algoritmo 2

```
PROCESS i;
bool flag[2]; /* flag[i] == true se il processo i vuole entrare*/
while(true){
    flag[i] = true; /* Indica che il processo i vuole entrare */
    while(flag[j]); /* Attesa attiva */
    // Sezione critica
    flag[i] = false; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

In questo algoritmo nel momento nel quale un processo vuole entra nella sezione critica, imposta il proprio flag a `true` e attende che l'altro processo imposti il proprio flag a `false`. In questo modo si garantisce che solo un processo alla volta possa entrare nella sezione critica. Qui il problema del progresso è risolto, ma questa soluzione presenta un problema di "stallo" (starvation), infatti se un processo imposta il proprio flag a `true` e poi avviene un timeout ed il processo viene messo in attesa, l'altro processo non potrà mai entrare nella sezione critica pur impostando il proprio flag a `true`.
Se l'impostazione del flag avvenisse dopo l'attesa attiva, allora non si presenterebbe il problema dello stallo, ma si presenterebbe un problema di mutua esclusione.

Algoritmo 3

```
PROCESS i;
int turn; /* Se turn == i processo i entra nella sezione critica */
bool flag[2]; /* flag[i] == true se il processo i vuole entrare*/
while(true){
    flag[i] = true; /* Indica che il processo i vuole entrare */
    turn = j; /* Passa il turno al processo j */
    while(flag[j] && turn == j); /* Attesa attiva */
    // Sezione critica
    flag[i] = false; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

Questo algoritmo risolve il problema dello stallo, e garantisce la mutua esclusione, questo grazie alla doppia condizione di attesa. Infatti, se il processo i vuole entrare nella sezione critica, imposta il proprio flag a `true` e poi passa il turno al processo j. Se il processo j è in attesa e ha il turno, il processo i non può entrare nella sezione critica, se invece il processo j non vuole entrare nella sezione critica il suo flag sarà `false` e il processo i potrà entrare nella sezione critica.

Algoritmo del fornaio

L'idea di questo algoritmo è quella che quando un processo vorrebbe entrare nella sezione critica, questo deve prima scegliere un numero, il processo con il numero più basso entra nella sezione critica. Se due processi scelgono lo stesso numero, il processo con l'identificativo più basso entra per primo. Questo algoritmo se implementato correttamente garantisce tre proprietà fondamentali:

```
PROCESS i;
int number[N]; /* number[i] == numero scelto dal processo i */
while(true){
    number[i] = Max(number[0], number[1], ..., number[N-1]) + 1; /* Sceglie un numero */
    for (j = 0; j < N; j++){
        while(number[j] != 0 && number[j] < number[i]); /* Attesa attiva */
    }
    // Sezione critica
    number[i] = 0; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

Se si fà particolare attenzione alla condizione di attesa, si può notare che il processo i entra nella sezione critica solo se il numero scelto è il più basso tra tutti i processi. Inoltre, se due processi scelgono lo stesso numero, il processo con l'identificativo più basso entra per primo. Questo algoritmo è molto semplice e intuitivo, ma presenta alcuni problemi di correttezza, infatti con questa implementazione non è garantita la mutua esclusione, in quanto due processi potrebbero scegliere lo stesso numero e entrare nella sezione critica contemporaneamente.

Algoritmo del fornaio v0.2

```
PROCESS i;
int number[N]; /* number[i] == numero scelto dal processo i */
int turn; /* Se turn == i processo i entra nella sezione critica */
bool choosing[N]; /* choosing[i] == true se il processo i sta scegliendo un numero */
while(true){
    choosing[i] = true; /* Indica che il processo i sta scegliendo un numero */
    number[i] = Max(number[0], number[1], ..., number[N-1]) + 1; /* Sceglie un numero */
    choosing[i] = false; /* Indica che il processo i ha scelto un numero */
    for (j = 0; j < N; j++){
        while(choosing[j]); /* Attesa attiva */
        while(number[j] != 0 && number[j] < number[i]); /* Attesa attiva */
    }
    // Sezione critica
    number[i] = 0; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

In questo algoritmo, il processo i prima di scegliere un numero imposta il proprio flag `choosing` a `true`, in questo modo gli altri processi sanno che il processo i sta scegliendo un numero e non devono entrare nella sezione critica. Dopo aver scelto un numero, il processo i imposta il proprio flag `choosing` a `false`, in questo modo gli altri processi sanno che il processo i ha scelto un numero e possono entrare nella sezione critica. Questo algoritmo garantisce la mutua esclusione, il progresso e l'attesa limitata.

Algoritmo del fornaio v1.0

Per ovviare ai problemi di correttezza dell'algoritmo del fornaio, è possibile utilizzare un algoritmo che utilizza una variabile `int number[N]` che contiene l'ultimo processo scelto, in questo modo si garantisce che l'ultimo processo scelto non vada a ri-occupare subito la sezione critica.

```
PROCESS i;
bool choosing[N]; /* choosing[i] == true se il processo i sta scegliendo un numero */
int number[N]; /* ultimo numero scelto dal processo i */
while(1){
    choosing[i] = true; /* Indica che il processo i sta scegliendo un numero */
    number[i] = Max(number[0], number[1], ..., number[N-1]) + 1; /* Sceglie un numero */
    choosing[i] = false; /* Indica che il processo i ha scelto un numero */
    for (j = 0; j < N; j++){
        while(choosing[j]); /* Attesa che il processo j ha scelto un numero */
        while(number[j] != 0 && (
            number[j] < number[i]
                ||
            number[j] == number[i]
            && i < j)); /* Attesa attiva */
    }
    // Sezione critica
    number[i] = 0; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

In questo algoritmo, viene risolto il problema della mutua esclusione, del progresso e dell'attesa limitata.

7.1.2 Soluzioni *hardware*

Le soluzioni *hardware* sono più veloci e più efficienti rispetto a quelle *software*, ma richiedono un maggiore sforzo di implementazione e possono essere più complesse da gestire. Un metodo "semplice" come gestione *hardware* è quello di disabilitare gli *interrupt* durante l'esecuzione della sezione critica. Questo metodo è semplice da implementare, ma presenta alcuni problemi, infatti se il test per l'accesso alla sezione critica viene eseguito in molto tempo gli *interrupt* dovrebbero essere disabilitati per molto tempo, causando un degrado delle prestazioni del sistema. Inoltre, questo metodo non è adatto per sistemi multiprocessore, poiché gli *interrupt* possono essere disabilitati solo per il processore corrente e non per gli altri processori. Una alternativa è quello di rendere l'operazione di accesso e scrittura alla variabile atomica, ovvero che impieghi un unico ciclo di *clock*. Esempio di questo è l'istruzione `test&set` oppure lo `compare&swap`.

Test&Set

L'istruzione `test&set` è un'istruzione atomica che consente di testare e impostare una variabile in un'unica operazione. Questa istruzione è utilizzata per implementare la mutua esclusione nei sistemi operativi. La sintassi dell'istruzione `test&set` è la seguente:

```
bool test_and_set(bool &var){
    bool temp;
    temp = var; /* Salva il valore corrente di var in temp */
    var = true; /* Imposta var a true */
    return temp; /* Restituisce il valore precedente di var */
}
```

Tutte e tre le operazioni sono eseguite in un'unica istruzione atomica, quindi non possono essere interrotte da altri processi. Quando la funzione viene chiamata, il valore corrente di `var` viene salvato in `temp`, quindi `var` viene impostato a `true` e infine il valore precedente di `var` viene restituito. Se il valore precedente di `var` era `false`, significa che la sezione critica è libera e il processo può entrarvi. Se il valore precedente di `var` era `true`, significa che la sezione critica è occupata e il processo deve attendere.

```
PROCESS i;
bool lock; /* lock == true se la sezione critica e' occupata */
while(true){
    while(test_and_set(lock)); /* Attesa attiva */
    // Sezione critica
    lock = false; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

In questo esempio, il processo i utilizza l'istruzione `test&set` per testare e impostare la variabile `lock`. Se la sezione critica è occupata, il processo i rimane in attesa finché non diventa libera. Dato che l'istruzione `test&set` è atomica, non ci sono problemi di mutua esclusione e il processo i può entrare nella sezione critica in modo sicuro. Inoltre solo il primo processo che vede `lock == false` può entrare nella sezione critica, gli altri processi rimarranno in attesa finché non diventa libera. Una volta che il processo i esce dalla sezione critica, imposta `lock` a `false` per indicare che la sezione critica è ora libera. Il problema di questo algoritmo è che l'attesa finita non è garantita, infatti non sono presenti meccanismi per evitare che un processo rimanga in attesa indefinitamente in quanto un processo potrebbe entrare ed uscire dalla sezione critica più volte prima che il processo i possa entrarvi.

Swap

Uso dell'istruzione `test&set` L'istruzione `swap` è un'altra istruzione atomica che consente di scambiare il valore di due variabili in un'unica operazione. Questa istruzione è utilizzata per implementare la mutua esclusione nei sistemi operativi. La sintassi dell'istruzione `swap` è la seguente:

```
void swap(bool &a, bool &b){
    bool temp;
    temp = a; /* Salva il valore corrente di a in temp */
    a = b; /* Imposta a al valore di b */
    b = temp; /* Imposta b al valore di temp */
}
```

Tutte e tre le operazioni sono eseguite in un'unica istruzione atomica, quindi non possono essere interrotte da altri processi. Quando la funzione viene chiamata, il valore corrente di `a` viene salvato in `temp`, quindi `a` viene impostato al valore di `b` e infine `b` viene impostato al valore di `temp`. In questo modo, i valori di `a` e `b` vengono scambiati in un'unica operazione atomica.

```
PROCESS i;
bool lock; /* lock == true se la sezione critica e' occupata */
while(true){
    bool dummy = true; /* Dummy per evitare di passare un riferimento */
    do
        swap(lock, dummy); /* Attesa attiva */
    while(dummy); /* Dummy == false se la sezione critica e' occupata */
    // Sezione critica
    lock = false; /* Indica che il processo i e' uscito */
    // Sezione non critica
}
```

In questo esempio, il processo i utilizza l'istruzione `swap` per scambiare il valore di `lock` con un valore `dummy`. Se la sezione critica è occupata, il processo i rimane in attesa finché non diventa libera. Dato

che l'istruzione `swap` è atomica, non ci sono problemi di mutua esclusione e il processo i può entrare nella sezione critica in modo sicuro. Inoltre solo il primo processo che vede `lock == false` può entrare nella sezione critica, gli altri processi rimarranno in attesa finché non diventa libera. Una volta che il processo i esce dalla sezione critica, imposta `lock` a `false` per indicare che la sezione critica è ora libera. Il problema di questo algoritmo è che l'attesa finita non è garantita, infatti non sono presenti meccanismi per evitare che un processo rimanga in attesa indefinitamente in quanto un processo potrebbe entrare ed uscire dalla sezione critica più volte prima che il processo i possa entrarvi.

Test&Set con attesa limitata

```
PROCESS i;
bool waiting[N]; /* waiting[i] == true se il processo i sta aspettando */
bool lock; /* lock == true se la sezione critica e' occupata */
while(1){
    waiting[i] = true; /* Indica che il processo i sta aspettando */
    bool key = true;
    while(waiting[i] && key){ /* Attesa attiva */
        key = test_and_set(lock); /* Prova ad entrare */
    }
    waiting[i] = false; /* Indica che il processo sta per entrare */
    // Sezione critica
    int j = (i+1) % N; /* Prendo in considerazione il prossimo processo */
    while(j != i && !waiting[j]){ // Controllo se il processo j sta aspettando
        j = (j+1) % N; /* Prendo in considerazione il prossimo processo */
    }
    if(j == i){ /* Se non ci sono altri processi in attesa */
        lock = false; /* Indica che la sezione critica e' libera */
    }else{
        waiting[j] = false; /* Indica che il processo j puo' entrare */
    }
    // Sezione non critica
}
```

In questo esempio, il processo i utilizza l'istruzione `test&set` per testare e impostare la variabile `lock`. Se la sezione critica è occupata, il processo i rimane in attesa finché non diventa libera. Dato che l'istruzione `test&set` è atomica, non ci sono problemi di mutua esclusione e il processo i può entrare nella sezione critica in modo sicuro. Inoltre solo il primo processo che vede `lock == false` può entrare nella sezione critica, gli altri processi rimarranno in attesa finché non diventa libera, oppure il processo che è appena uscito dalla sezione critica gli passi il turno. Se nessun processo è in attesa allora l'ultimo processo che è uscito dalla sezione critica libera la sezione critica impostando `lock` a `false`. In questo modo si garantisce che solo un processo alla volta possa entrare nella sezione critica e che l'attesa sia limitata.

7.2 Semafori

Uso dell'istruzione `swap` I semafori sono una primitiva di sincronizzazione la quale è un numero intero non negativo, che può essere incrementato e decrementato. Per eseguire queste operazioni, i semafori utilizzano due operazioni fondamentali: `wait` - `P` e `signal` - `V`. La prima operazione decrementa il valore del semaforo, se questo è maggiore di zero, altrimenti il processo viene messo in attesa. La seconda operazione incrementa il valore del semaforo e, se ci sono processi in attesa, uno di essi viene risvegliato. Esistono due tipi principali di semafori, i semafori binari e i semafori generici. I semafori binari possono assumere solo i valori 0 e 1, mentre i semafori generici possono assumere qualsiasi valore intero non negativo. I semafori binari possono essere implementati tramite i semafori generici, ma non viceversa.

Semafori binari

L'implementazione concettuale delle funzioni `wait` e `signal` è la seguente:

```
P(semaphore s){
    while(s == false){ /* Attesa attiva */
        // Non fa nulla
    }
    s = false; /* Imposta il semaforo a false */
}
V(semaphore s){
    s = true; /* Imposta il semaforo a true */
}
```

In questo esempio, la funzione `P` attende che il semaforo sia libero (ovvero che il suo valore sia `true`) e poi lo imposta a `false`. La funzione `V` imposta il semaforo a `true`, indicando che la sezione critica è ora libera.

Semafori generici

L'implementazione concettuale delle funzioni `wait` e `signal` è la seguente:

```
P(semaphore s){
    while(s <= 0){ /* Attesa attiva */
        // Non fa nulla
    }
    s--; /* Decrementa il semaforo */
}
V(semaphore s){
    s++; /* Incrementa il semaforo */
}
```

In questo esempio, la funzione `P` attende che il semaforo sia maggiore di zero e poi lo decrementa. La funzione `V` incrementa il semaforo, indicando che la sezione critica è ora libera.

Note

Per essere efficienti le funzioni `P` e `V` devono essere implementate in modo atomico, in modo che non possano essere interrotte da altri processi. Come abbiamo visto però l'implementazione atomica delle funzioni `P` e `V` non è garantita. Andiamo dunque a vedere come garantire l'atomicità delle funzioni `P` e `V` utilizzando i semafori.

```
/* Inizializzato s a true */
P(bool &s){
    bool key = false;
    do{
        swap(s, key); /* Attesa attiva */
    }while(key == false);
    s = false; /* Imposta il semaforo a false */
}
V(bool &s){
    s = true; /* Imposta il semaforo a true */
}
```

In questo esempio, la funzione `P` utilizza l'istruzione `swap` per testare e impostare il semaforo. Se il semaforo è occupato, il processo rimane in attesa finché non diventa libero. La funzione `V` imposta il semaforo a `true`, indicando che la sezione critica è ora libera. In questo modo si garantisce che le funzioni `P` e `V` siano atomiche e che non possano essere interrotte da altri processi.

Implementazione semafori interi

Coi semafori interi l'implementazione si complica, dobbiamo comunque garantire l'atomicità delle funzioni `P` e `V`, ma dobbiamo anche garantire che il semaforo non possa assumere valori negativi ed che sia possibile incrementarlo di uno senza che durante questo tempo un altro processo possa decrementarlo. Per fare ciò possiamo usiamo due semafori binari, `mutex` e `delay` questi vengono usati rispettivamente per garantire la mutua esclusione e per garantire che le operazioni siano eseguite in modo atomico. La funzione `P` viene implementata come segue:

```
P(semaphore s){
    P(mutex); /* Acquisisce il semaforo mutex */
    s = s - 1; /* Decrementa il semaforo */
    if(s < 0){ /* Se il semaforo e' negativo */
        V(mutex); /* Rilascia il semaforo mutex */
        P(delay); /* Attende il semaforo delay */
    }else{
        V(mutex); /* Rilascia il semaforo mutex */
    }
}
```

In questo esempio, la funzione `P` acquisisce il semaforo `mutex` per garantire la mutua esclusione. Poi decrementa il semaforo e controlla se il suo valore è negativo. Se il valore è negativo, rilascia il semaforo `mutex` e attende il semaforo `delay` (che è Inizializzato a `false`). Se il valore è maggiore o uguale a zero, rilascia il semaforo `mutex` e termina. La funzione `V` viene implementata come segue:

```
V(semaphore s){
    P(mutex); /* Acquisisce il semaforo mutex */
    s = s + 1; /* Incrementa il semaforo */
    if(s <= 0){ /* Se il semaforo e' negativo */
        V(delay); /* Rilascia il semaforo delay */
    }
    V(mutex); /* Rilascia il semaforo mutex */
}
```

In questo esempio, la funzione `V` acquisisce il semaforo `mutex` per garantire la mutua esclusione. Poi incrementa il semaforo e controlla se il suo valore è negativo. Se il valore è negativo, rilascia il semaforo `delay` (che è Inizializzato a `false`). Se il valore è maggiore o uguale a zero, rilascia il semaforo `mutex` e termina. Nota come acquisire il semaforo `mutex` prima di modificare il semaforo `s` e rilasciarlo dopo averlo modificato, dato che questo è inizializzato a `true`, serva per garantire che ogni operazione sul semaforo `s` sia eseguita in modo atomico.
Notiamo come nell'implementazione, anche se risolviamo il problema della sezione critica, non risolviamo il problema del *busy-waiting* ovvero l'attesa attiva. Infatti, se un processo entra nella sezione critica quando il suo tempo di `CPU`-*burst* termina e il processo viene messo in attesa, il semaforo `mutex` rimarrà occupato fino a quando il processo non verrà risvegliato andando a far sprecare al processo corrente tempo di `CPU`.

Implementazione senza *busy-waiting*

Per evitare il problema del *busy-waiting* possiamo utilizzare una lista `s.list` che contiene la lista dei processi in attesa. La funzione `P` viene implementata come segue:

```
P(semaphore s){
    P(mutex); /* Acquisisce il semaforo mutex */
    s.value = s.value - 1; /* Decrementa il semaforo */
    if(s.value < 0){ /* Se il semaforo e' negativo */
        append(process I, s.List); /* Aggiunge il processo alla lista */
        sleep() & V(mutex); /* Rilascia il semaforo mutex */
    }else{
        V(mutex); /* Rilascia il semaforo mutex */
    }
}
```

In questo esempio, la funzione `P` acquisisce il semaforo `mutex` per garantire la mutua esclusione. Poi decrementa il semaforo e controlla se il suo valore è negativo. Se il valore è negativo, aggiunge il processo alla lista dei processi in attesa e lo mette in attesa. Se il valore è maggiore o uguale a zero, rilascia il semaforo `mutex` e termina. La funzione `V` viene implementata come segue:

```
V(semaphore s){
    P(mutex); /* Acquisisce il semaforo mutex */
    s.value = s.value + 1; /* Incrementa il semaforo */
    if(s.value <= 0){ /* Se il semaforo e' negativo */
        PCB *p = remove(s.List); /* Rimuove il primo processo dalla lista */
        wakeup(p) & V(mutex); /* Rilascia il semaforo mutex */
    }else{
        V(mutex); /* Rilascia il semaforo mutex */
    }
}
```

In questo esempio, la funzione `V` acquisisce il semaforo `mutex` per garantire la mutua esclusione. Poi incrementa il semaforo e controlla se il suo valore è negativo. Se il valore è negativo, rimuove il primo processo dalla lista dei processi in attesa e lo risveglia. Se il valore è maggiore o uguale a zero, rilascia il semaforo `mutex` e termina. Notare come non abbiamo discusso l'assenza di *busy-waiting* sui semafori binari `mutex` e `delay`, ma generalmente il processo non rimane in attesa attiva, ma gli viene inviato un segnale di *sleep* e viene messo in attesa. Questo algoritmo è più efficiente rispetto all'implementazione precedente, poiché non richiede l'attesa attiva e consente di risparmiare tempo di `CPU`. Tuttavia, questo algoritmo richiede un maggiore sforzo di implementazione e può essere più complesso da gestire.
Per gestire l'implementazione senza *busy-waiting* per i semafori booleani `mutex` e `delay` dobbiamo far conto sul `SO`, infatti questo può gestire questi o disabilitando gli `interrupt`, oppure ignorando il *busy-waiting* su `mutex` in quanto questo è un semaforo binario ed il suo cambio è molto veloce.
Per garantire l'assenza di *starvation* la lista deve essere implementata tramite `FIFO`

Applicazioni dei semafori

Nel caso di voler eseguire in sequenza i processi A e poi B allora scriveremo il seguente codice:

```
/* S=0 */
PROCESS A;
    // Esecuzione del processo A
    V(S); /* Indica che il processo A e' uscito */
PROCESS B;
    // Attesa del processo A
    P(S); /* Attende il processo A */
    // Esecuzione del processo B
```

In ogni caso col semaforo `S` inizializzato a 0, il processo A eseguirà per primo e poi il processo B, sia che il tempo in `CPU` venga assegnato prima ad A e poi a B, sia che il tempo in `CPU` venga assegnato prima a B e poi a A. Infatti, il processo B non potrà entrare nella sezione critica finché il processo A non avrà rilasciato il semaforo `S`.
Se invece vogliamo eseguire in sequenza i processi $A, B, A, B, \ldots$ allora scriveremo il seguente codice:

```
/* S=0 S1=1 */
PROCESS A;
while(1){
    P(S1);
    // Esecuzione del processo A
    V(S); /* Indica che il processo A e' uscito */
}
PROCESS B;
while(1){
    P(S);
    // Esecuzione del processo B
    V(S1); /* Indica che il processo B e' uscito */
}
```

In questo caso, il processo A e il processo B si alternano nell'esecuzione. Infatti, il processo A entra nella sezione critica e poi rilascia il semaforo `S` per consentire al processo B di entrare nella sezione critica. Una volta che il processo B esce dalla sezione critica, rilascia il semaforo `S1` per consentire al processo A di entrare nella sezione critica. Questo ciclo continua fino a quando i processi non terminano.[1]

[1] Ulteriori esempi ed applicazioni trattate a lezione sono state omesse, in quanto anche se trattate approfonditamente a lezione, queste sono la merita applicazione del concetto di semaforo.

Capitolo 8

Deadlock

Quando l'uso contemporaneo delle risorse da parte di più processi porta a una situazione in cui nessun processo può proseguire senza aver accesso a una risorsa, si verifica un *deadlock*.
Un classico esempio di *deadlock* è quello di due processi devono acquisire due risorse per completare il loro lavoro. Se il processo A acquisisce la risorsa 1 e il processo B acquisisce la risorsa 2, entrambi i processi non possono proseguire perché ciascuno aspetta che l'altro rilasci la risorsa di cui ha bisogno. Generalizzando sono necessarie quattro condizioni affinché si verifichi un *deadlock*:

1. **Mutua esclusione**: almeno una risorsa deve essere assegnata in modo esclusivo a un processo. Se una risorsa è assegnata a un processo, nessun altro processo può accedervi.
2. ***Hold & Wait***: un processo che detiene almeno una risorsa sta aspettando di acquisire altre risorse.
3. **Nessuna *preemption***: una risorsa non può essere forzatamente rimossa da un processo che la detiene. La risorsa può essere rilasciata solo volontariamente dal processo che la detiene.
4. **Attesa circolare**: esiste un insieme di processi P1, P2, ..., Pn tali che P1 sta aspettando una risorsa detenuta da P2, P2 sta aspettando una risorsa detenuta da P3, ..., e Pn sta aspettando una risorsa detenuta da P1.

Per analizzare come una risorsa viene allocata ai processi, è utile costruire un `RAG` (*Resource Allocation Graph*), un grafo diretto in cui i nodi rappresentano processi e risorse. Le risorse sono rappresentate da cerchi, mentre i processi sono rappresentati da quadrati. Un arco diretto da un processo a una risorsa indica che il processo sta aspettando la risorsa, mentre un arco diretto da una risorsa a un processo indica che la risorsa è attualmente assegnata al processo. Se esiste un ciclo nel grafo, si verifica un *deadlock*, se ci sono più istanze di una risorsa ed esiste un ciclo allora non è detto che ci sia un *deadlock*.

8.1 Prevenzione del deadlock

Abbiamo diverse strategie per prevenire il *deadlock*:

- Prevenzione Statica
- Prevenzione Dinamica
- Rilevamento (*Detection*) e Recupero (*Recovery*)
- Struzzo - Non preoccuparsi del *deadlock*

8.1.1 Prevenzione Statica

La prevenzione statica del *deadlock* implica l'uso della scrittura di codice che eviti che si verifichi una delle quattro condizioni necessarie per il *deadlock*.

Mutua esclusione

Non è possibile evitare la mutua esclusione, poiché è una condizione necessaria per l'uso delle risorse.

Hold & Wait

Soluzioni Per evitare questa condizione, un processo deve acquisire tutte le risorse di cui ha bisogno prima di iniziare a eseguire. Inoltre un processo può acquisire una risorsa solo se non detiene già altre risorse.

Problemi Questo approccio può portare a un utilizzo inefficiente delle risorse e a un aumento del tempo di attesa.

No preemption

Soluzioni Per evitare questa condizione, un processo che richiedere una risorsa non disponibile deve rilasciare tutte le risorse che detiene, oppure può cedere la risorsa che detiene su richiesta di un altro processo.

Problemi Questo approccio è fattibile solo per risorse digitali come CPU e memoria, ma non per risorse fisiche come stampanti o dischi.

Attesa circolare

Soluzioni Per evitare questa condizione, è possibile assegnare una priorità ad ogni risorsa, in modo che un processo possa acquisire solo risorse con priorità superiore a quelle già detenute. In questo modo si evita la formazione di cicli di attesa in quanto un processo non può acquisire una risorsa con priorità inferiore a quelle già detenute.

Problemi Solitamente è difficile definire una priorità per le risorse, e questo approccio può portare a un utilizzo inefficiente delle risorse.

8.1.2 Prevenzione Dinamica

Dato che la prevenzione statica del *deadlock* può portare a un utilizzo inefficiente delle risorse, in quanto queste tecniche impostano dei vincoli rigidi sul modo in cui i processi possono acquisire le risorse, la prevenzione dinamica punta a basare la prevenzione del *deadlock* sulla base delle richieste delle risorse da parte dei processi. In questo modo, i processi possono acquisire le risorse in modo più flessibile, evitando il *deadlock* senza compromettere l'efficienza delle risorse.

Prerequisito Per utilizzare la prevenzione dinamica del *deadlock*, è necessario che il sistema operativo conosca a priori il caso peggiore, ovvero il numero massimo di risorse che ogni processo richiederà.

Il mentenimento del *safe state*

Definizione Lo stato di assegnazione delle risorse, calcolato come il numero di istanze di risorse disponibili e il numero di istanze di risorse assegnate a ciascun processo su le richieste massime, è definito *safe state* se esiste una *safe sequence* di processi.

Safe sequence Una sequenza di processi $(P_1, \ldots, P_n)$ è definita *safe sequence* se ogni processo P_i, le risorse che richiede P_i possono essere soddisfatte usando le risorse disponibili e le risorse rilasciate dai processi $P_1, \ldots, P_{i-1}$.
Se non esiste una *safe sequence* di processi, il sistema è in uno stato non sicuro (*unsafe state*) e si potrebbe verificare un *deadlock*, ma non è garantito che si verifichi.

La prevenzione

Per mantenere il sistema in uno stato sicuro, il sistema operativo attua degli algoritmi per decidere se una richiesta di risorse da parte di un processo può essere soddisfatta senza portare il sistema in uno stato non sicuro. Dunque ad ogni richiesta di risorse da parte di un processo, il sistema operativo attua questa verifica.

Svantaggi La prevenzione dinamica del *deadlock* porta comunque ad una riduzione dell'efficienza del sistema, poiché il sistema operativo deve eseguire calcoli complessi per determinare se una richiesta di risorse può essere soddisfatta senza portare il sistema in uno stato non sicuro. Inoltre, la prevenzione dinamica riduce l'uso delle risorse rispetto ad un non-uso della prevenzione del *deadlock*, poiché i processi devono attendere che il sistema operativo calcoli se la richiesta di risorse può essere soddisfatta.

Implenentazione

L'implementazione viene effettuata solitamente tramite o l'uso del `RAG`, oppure tramite l'uso dell' "algoritmo del banchiere".

Uso di `RAG` L'algoritmo che prevede l'uso di un `RAG` funziona solo se si ha una istanza per ogni risorsa. Il `RAG` viene esteso con archi di "rivendicazione", ovvero archi che rappresentano una "possibilità di richiesta" di una risorsa da parte di un processo, $P_i \to R_j$ se P_i in futuro potrebbe richiedere R_j. Questo comporta che ogni processo quando viene creato deve dichiarare le risorse che potrebbe richiedere in futuro.
In un secondo momento, il sistema operativo allocherà la risorsa R_j al processo P_i se l'arco $R_j \to P_i$ non crei un ciclo nel grafo. Se si verifica un ciclo, il sistema operativo non assegnerà la risorsa al processo P_i in quanto porterebbe il sistema in uno stato non sicuro. Se il grafo non ha cicli, il sistema operativo assegnerà la risorsa al processo P_i e rimuoverà l'arco di rivendicazione $P_i \to R_j$.

Algoritmo del banchiere L'algoritmo del banchiere è un algoritmo meno efficiente rispetto all'uso del `RAG`, ma funziona con qualunque numero di istanze di risorse.
Ogni processo è un cliente che possono richiedere del credito dalla banca (ovvero il sistema operativo) ogni risorsa allocabile è vista come del denaro. Ovviamente il `SO` non permetterà a tutti i clienti di raggiungere il massimo limite personale contemporaneamente. Ciò infatti porterebbe ad una situazione di *deadlock*.
Anche in questo algoritmo ogni processo dichiarare il numero massimo di risorse che andrà ad richiedere, ad ogni richiesta di allocazione si verifica se questa mantiene lo stato *safe*. Per fare ciò si sfruttano due algoritmi differenti, uno per l'allocazione ed uno per la verifica dello stato.

Listing 8.1: Algoritmo del banchiere

```
int available[m]// numero di istanze di R_i disponibili
int max[n][m]   // matrice delle richieste massime
int alloc[n][m] // matrice delle risorse allocate
int need[n][m]  // matrice delle richieste residue
                // (need[i][j] = max[i][j] - alloc[i][j])

void request(int req_vec[]){
    if(req_vec[i] > need[i][j])
        error("Richiesta maggiore del massimo richiesto");
    if(req_vec[i] > available[j])
        wait();
    available[] = available[] - req_vec[];
    alloc[i][] = alloc[i][] + req_vec[];
    need[i][] = need[i][] - req_vec[];
    if(!state_safe()){
        available[] = available[] + req_vec[];
        alloc[i][] = alloc[i][] - req_vec[];
        need[i][] = need[i][] + req_vec[];
        wait();
    }
}

bool state_safe(){
    int work[m] = available[];
    bool finish[n] = {false, false, ..., false};
    int i;
    while(finish != {true, true, ..., true}){
        // cerca P_i che non e' terminato
        // e che puo' terminare
```

```
        for(i=0; i<n && (finish[i] || need[i][] > work[]); i++);
        if(i == n) // non esiste P_i che puo' terminare
            return false; // non e' in uno stato sicuro
        work[] = work[] + alloc[i][]; // rilascia le risorse di P_i
        finish[i] = true; // P_i e' terminato
    }
    return true; // e' in uno stato sicuro
}
```

Notare come questo algoritmo sia molto dispendioso e che deve essere eseguito ogni volta che un processo richiede una risorsa.[1]

8.1.3 Rilevamento del *deadlock* & ripristino

Gli algoritmi di rilevamento del *deadlock* sono progettati per rilevare la presenza di un *deadlock* nel sistema e per ripristinare il sistema a uno stato sicuro. Questi algoritmi sono meno restrittivi rispetto alla prevenzione (statico o dinamico) e consentono ai processi di acquisire le risorse in modo più flessibile. Tuttavia, richiedono un monitoraggio costante del sistema per rilevare la presenza di un *deadlock* ed eventualmente correggere la situazione, questi possono comportare un sovraccarico significativo.
Esistono due approcci principali per il rilevamento del *deadlock*: il rilevamento basato su `RAG` di attesa e l'"algoritmo di rilevazione".

Rilevamento basato su `RAG`

Anche in questo caso il `RAG` funziona solo se si ha una istanza per ogni risorsa. Il grafo di attesa è un grafo composto da soli processi, ogni arco nel grafo ($P_i \to P_j$) indica che il processo P_i sta aspettando una risorsa detenuta dal processo P_j. Se esiste un ciclo nel grafo, si verifica un *deadlock*. Il sistema operativo può utilizzare questo grafo per rilevare la presenza di un *deadlock* e per determinare quali processi sono coinvolti nel *deadlock*. Tuttavia, questo approccio richiede che il sistema operativo monitori costantemente il grafo

Algoritmo di rilevamento

L'algoritmo di rilevamento esplora ogni possibile sequenza di allocazione dei processi che ancora non hanno terminato, se la sequenza và a buon fine, allora non c'è *deadlock* altrimenti potrebbe esserci.
L'algoritmo è il seguente:

Listing 8.2: Algoritmo di rilevamento

```
int available[m]// numero di istanze di R_i disponibili
int alloc[n][m] // matrice delle risorse allocate
int req_vec[n][m] // matrice delle richieste

void check(){
    int work[m] = available[]; // risorse disponibili
    bool finish = {false, false, ..., false}; // processi terminati
    bool found = true; // trovato un processo che puo' terminare
    while(found){
        found = false; // resetta la variabile
        for(int i=0; i<n && !found; i++){
            // cerca P_i che non e' terminato
            // e che puo' terminare
            if(!finish[i] && req_vec[i][] <= work[]){
                // Assumo che P_i possa terminare senza deadlock
                // dunque rilascia le risorse senza deadlock
                work[] = work[] + alloc[i][]; // rilascia le risorse di P_i
                finish[i] = true; // P_i terminera' senza problemi
                found = true; // trovato un processo che puo' terminare
            }
        }
    }
```

[1] Nota dell'autore: È utile imparare questo algoritmo a memoria, in quanto verrà richiesto in sede di esame.

```
    // Se finish[i] == false, per uno o piu' P_i
    // allora i processi P_i sono in deadlock
}
```

Ripristino

Prima di procedere al ripristino del *deadlock*, bisogna capire ogni quando eseguire l'algoritmo di rilevamento del *deadlock*, solitamente questo viene fatto o dopo ogni richiesta, o dopo N secondi, oppure quando l'uso della `CPU` scende sotto una soglia T preimpostata.
Una volta che il *deadlock* è stato rilevato, il sistema operativo deve decidere come ripristinare il sistema a uno stato sicuro, ci sono diversi approcci per il ripristino del *deadlock*.

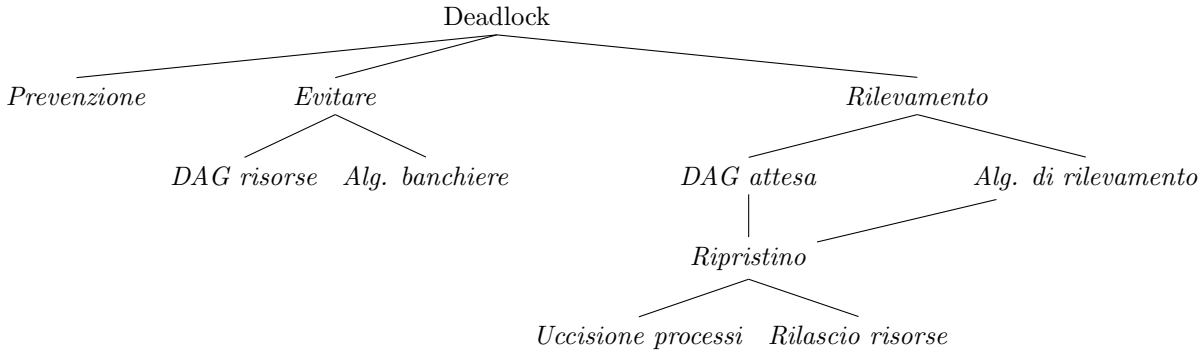
Uccisione di processi La soluzione più comune per il ripristino del *deadlock* è quella di uccidere uno o più processi coinvolti nel *deadlock*. Ci sono due approcci principali per uccidere i processi:

- Uccidere tutti i processi coinvolti nel ciclo - Soluzione molto costosa e drastica, ma è la più semplice da implementare.
- Uccidere selettivamente i processi coinvolti nel ciclo - Questa soluzione è più complessa da implementare, ma può essere più efficiente in termini di utilizzo delle risorse, ma non dal punto di uso in quanto deve essere ri-eseguita la rilevazione dopo ogni uccisione.

Rilascio di risorse Un'altra soluzione per il ripristino del *deadlock* è quella di rilasciare alcune risorse detenute dai processi coinvolti nel *deadlock*. Questa soluzione presente un problema: i processi che vengono forzati a rilasciare le risorse potrebbero non essere in grado di completare il loro lavoro normalmente, portando a un aumento del tempo di attesa e a una riduzione dell'efficienza del sistema. Inoltre se in situazione di *deadlock* rilascio sempre le risorse coinvolte in uno stesso processo, si potrebbe verificare una situazione di *starvation*, per ovviare questo si deve tener conto dei *rollback* effettuati in precedenza.

8.1.4 Conclusioni

Concludendo possiamo riassumere le varie soluzioni al *deadlock* del seguente diagramma:



Ognuno dei metodi ha i suoi vantaggi e svantaggi, e la scelta del metodo dipende dalle esigenze specifiche del sistema e dalle risorse disponibili. Principalmente però nessuno dei metodi di gestione è ottimale e delle soluzioni combinate di più metodi possono essere più efficaci se dividiamo le risorse in classi quali: risorse interne, memoria, risorse di processo (eg. file) e *swap* applicando ad ognuna di queste la soluzione più efficiente: ordinamento di risorse per la prima, prelazione per la memoria, prevenzione dinamica per i file, e prevenzione con pre-allocazione (*hold&wait*) per la *swap*.
In linea principale però la soluzione più semplice ed adottata dalla maggior parte dei `SO` è quella dello "struzzo" ovvero ignoriamo il problema del *deadlock* in quanto si verificano raramente e tutte le soluzioni sono costose o con algoritmi sbagliati.

Capitolo 9

Gestione della memoria

In questo capitolo si parlerà della gestione della memoria nei sistemi operativi, in particolare si analizzeranno gli spazi di indirizzamento, l'allocazione contigua, la paginazione, la segmentazione e la segmentazione con paginazione. Si analizzeranno anche le problematiche legate alla gestione della memoria, e le tecniche di allocazione della memoria. Si parlerà anche della memoria virtuale e delle tecniche di swapping, e si analizzeranno i problemi legati alla frammentazione della memoria. Infine si parlerà della gestione della memoria nei sistemi operativi moderni, e delle tecniche di allocazione della memoria nei sistemi operativi a *micro-kernel*.

9.1 Introduzione

Per rendere un sistema operativo efficiente la condivisione della memoria deve essere gestita in modo da evitare conflitti tra i processi ed efficienti.

Problematiche Le problematiche che un buon sistema operativo deve affrontare in merito alla gestione della memoria sono:

- Allocazione della memoria ai singoli *job*
- Protezione dello spazio di indirizzamento
- Condivisione dello spazio di indirizzamento
- Nei sistemi operativi moderni: Gestione della memoria virtuale (*swap*)

Tutti questi problemi devono obbligatoriamente essere affrontati in modo, oltre che efficiente, anche sicuro per evitare che un processo possa accedere alla memoria di un altro processo o di un dispositivo e compromettere la stabilità del sistema e/o la sicurezza dei dati.

9.1.1 Dal programma al processo

Ogni programma prima di essere eseguito deve essere caricato in memoria e trasformato in un processo. Il sistema operativo deve successivamente prelevare le istruzioni in memoria sulla base del `PC` (*Program Counter*) e caricare i dati necessari per l'esecuzione del programma. Quando questo è terminato, il sistema operativo deve liberare la memoria occupata dal processo e restituirla al sistema. La creazione di un processo è quindi un'operazione complessa che richiede l'allocazione di risorse e la gestione della memoria.
Prima di trasformare un programma in un processo, ci sono diverse operazioni da eseguire, in ognuna di queste operazioni possono esserci diverse semantiche degli indirizzi (spazio logico / spazio fisico), solitamente quando si scrive il codice sorgente si fa riferimento a indirizzi logici, mentre quando il codice viene eseguito si deve fare riferimento a indirizzi fisici.
Il compilatore traduce gli indirizzi logici in indirizzi simbolici ri-locabili, il linker o il loader si occupano di tradurre gli indirizzi ri-locabili in indirizzi assoluti, infine il sistema operativo si occupa di tradurre gli indirizzi assoluti in indirizzi fisici.
L'insieme delle operazioni che vengono eseguite per trasformare gli indirizzi simbolici in indirizzi assoluti è detto ***binding***.

Binding & Indirizzi

Il *binding* può avvenire in diversi momenti, e a seconda di quando avviene il *binding* si può dividere in *binding* statico e *binding* dinamico:

- ***Binding* a *compile time***: il *binding* avviene durante la compilazione del programma, gli indirizzi simbolici vengono tradotti in indirizzi assoluti e il programma viene caricato in memoria con gli indirizzi assoluti. Questo tipo di *binding* è molto veloce, ma non permette di modificare il programma una volta compilato. Questo tipo di *binding* è della categoria del *binding* statico.
- ***Binding* a *load time***: il *binding* avviene durante il caricamento del programma in memoria, gli indirizzi simbolici vengono tradotti in indirizzi assoluti e il programma viene caricato in memoria con gli indirizzi assoluti. Questo tipo di binding richiede che gli indirizzi siano ri-locabili, e in questo caso vengono usati indirizzi relativi alla posizione di memoria in cui il programma viene caricato. Questo tipo di *binding* è della categoria del *binding* statico in quanto se il programma viene caricato in una posizione di memoria diversa da quella prevista, il sistema operativo deve modificare gli indirizzi assoluti in indirizzi relativi alla nuova posizione di memoria.
- ***Binding* a *run time***: il *binding* avviene durante l'esecuzione del programma, gli indirizzi simbolici vengono tradotti in indirizzi assoluti ed il programma può essere caricato in memoria in qualsiasi posizione e spostato in memoria durante l'esecuzione. Questo tipo di *binding* è della categoria del *binding* dinamico ma richiede del supporto *hardware* aggiuntivo.

Collegamento (*linking*)

Un altro passaggio prerequisito alla creazione di un processo è il *linking*, ovvero l'associazione di un programma a tutte le librerie e i moduli necessari. Questo passaggio può avvenire in due modi:

- ***Linking* statico**: il *linking* avviene durante la compilazione del programma, le librerie e i moduli necessari vengono inclusi per intero nel programma e il programma viene caricato in memoria con tutte le librerie e i moduli necessari. Questo tipo di *linking* è molto veloce, ma aumenta la dimensione del programma e richiede più memoria. Inoltre se una libreria o un modulo viene aggiornato, il programma deve essere ricompilato per utilizzare la nuova versione della libreria o del modulo, oltre al fatto che se non si usano tutte le funzioni di una libreria, il programma occupa più memoria del necessario.
- ***Linking* dinamico**: il *linking* avviene durante l'esecuzione del programma, le librerie e i moduli necessari vengono caricati in memoria solo quando sono necessari e il programma viene caricato in memoria con i riferimenti alle librerie e ai moduli necessari. Questo tipo di *linking* è più lento, ma riduce la dimensione del programma e richiede meno memoria. Inoltre se una libreria o un modulo viene aggiornato, il programma non deve essere ricompilato per utilizzare la nuova versione della libreria o del modulo.

Il *linking* statico è più veloce, ma richiede più memoria e non permette di aggiornare le librerie e i moduli senza ricompilare il programma. Il *linking* dinamico è più lento, ma richiede meno memoria e permette di aggiornare le librerie e i moduli senza ricompilare il programma.

Caricamento (*loading*)

Il caricamento è l'operazione che permette di caricare un programma in memoria e prepararlo per l'esecuzione. Il caricamento può avvenire in due modi:

- ***Loading* statico**: l'intero programma viene caricato in memoria in un'unica operazione, il programma viene caricato in memoria con gli indirizzi assoluti e il programma viene eseguito. Questo tipo di *loading* è molto veloce, ma richiede che il programma sia di dimensioni fisse e non permette di modificare il programma una volta caricato in memoria.
- ***Loading* dinamico**: il caricamento di alcuni moduli avviene in modo dinamico, ovvero il programma viene caricato in memoria in più operazioni quando sono necessari. Questo tipo di *loading* è più lento, ma permette di caricare solo i moduli necessari e di modificare il programma una volta caricato in memoria. Inoltre se una porzione di codice non viene mai eseguita, non viene caricata in memoria e quindi non occupa spazio in memoria.

Il *loading* statico è più veloce, ma richiede che il programma sia di dimensioni fisse e non permette di modificare il programma una volta caricato in memoria. Il *loading* dinamico è più lento, ma permette di caricare solo i moduli necessari e di modificare il programma una volta caricato in memoria.

9.1.2 Spazi di indirizzamento

Come accennato in precedenza esistono due tipi di indirizzamento: l'indirizzamento logico, gestito dalla `CPU` e l'indirizzamento fisico, gestito dalla memoria. L'associazione tra i due indirizzamenti nel caso del *binding* statico è semplice, in quanto gli indirizzi logici e fisici coincidono. Nel caso del *binding* dinamico invece, l'associazione tra i due indirizzamenti è più complessa, in quanto gli indirizzi logici e fisici non coincidono e devono essere tradotti. Per gestire questa traduzione viene usato un processore detto `MMU` (*Memory Management Unit*), che si occupa di tradurre gli indirizzi logici in indirizzi fisici. Questo componente agisce a *run time* e a partite da un indirizzo base, pre-impostato per il processo in esecuzione, lo somma all'indirizzo logico per ottenere l'indirizzo fisico. Questo processo è detto **relocazione dinamica** e permette di eseguire più processi in memoria senza conflitti.

Considerazioni Bisogna considerare che in un sistema multi-programmato non è possibile conoscere in anticipo dove un processo può essere posizionato in memoria, e quindi è necessario che il sistema operativo gestisca la memoria in modo da evitare conflitti tra i processi. Inoltre l'esigenza di avere lo *swap* impedisce di poter usare indirizzi ri-locati in modo statico. Ne consegue che la ri-locazione dinamica viene usato per sistemi più "complessi" e la gestione è eseguita dal `SO`, mentre la ri-locazione statica viene usata solo per applicazioni specifiche ed il `SO` non può fare granché in materia di gestione della memoria.

9.2 Schemi di gestione della memoria

La gestione della memoria è un'altra funzione fondamentale del sistema operativo, che deve garantire l'allocazione della memoria ai processi in modo da migliorare l'uso della memoria e garantire la protezione dei processi.
Esistono diversi schemi di gestione della memoria, ognuno con i propri vantaggi e svantaggi. I principali schemi di gestione della memoria sono:

- Allocazione contigua
- Paginazione
- Segmentazione
- Segmentazione con paginazione

Anche se nelle soluzioni reali viene usata della memoria virtuale

9.2.1 Allocazione contigua

L'allocazione contigua è lo schema di gestione della memoria più semplice il quale prevede che la memoria sia suddivisa in partizioni, le quali possono essere o fisse, o variabili. Se la dimensione dell'immagine di un processo occupa $10Kb$ allora questo occuperà $10Kb$ consecutivi.

Allocazione con partizioni fisse Quando si usa l'allocazione con partizioni fisse, la memoria viene suddivisa in partizioni di dimensioni fisse (solitamente usando partizioni con dimensioni di potenze di 2) e ogni processo viene caricato in una partizione di dimensioni fisse. Questo tipo di allocazione è semplice e veloce, ma può portare a problemi di assegnazione di memoria a diversi *job*, se ad esempio un processo richiede $10Kb$ e la partizione più grande disponibile è di $8Kb$, il processo non può essere caricato in memoria finché non viene liberata una partizione di dimensioni sufficienti. Inoltre se un processo richiede meno memoria di quella disponibile in una partizione, la memoria rimanente non può essere utilizzata da altri processi, portando a problemi di frammentazione interna.

***Scheduling* a lungo termine**In questo caso lo *scheduling* viene eseguito o con più code (una per ogni partizione) oppure con una coda semplice. Nel primo caso ogni processo viene associato alla partizione più piccola che lo può contenere e viene eseguito in modo da non superare la dimensione della partizione. Nel secondo caso il processo che viene allocato potrebbe o essere il primo della coda (`FCFS`) che và ad occupare la partizione più piccola disponibile, oppure viene eseguita una scansione della coda e una determinata partizione viene assegnato o al processo che richiede la memoria più simile alla dimensione della partizione (*best-fit-only*) oppure viene assegnato il primo *job* che può stare nella partizione (*first-available-fit*).
In tutti i casi abbiamo diverse problematiche, nelle code per ogni partizione dopo che un *job* è entrato in coda non può essere spostato in un'altra coda, e quindi non può essere spostato in una partizione più piccola o più grande se questa è libera e non ci sono altri processi in attesa. Problema simile riguarda il `FCFS` in quanto se non è disponibile una partizione abbastanza grande per il primo processo, allora tutti i processi attendono anche se ci sono delle partizioni libere. Mentre nel caso del *best-fit-only* e del *first-available-fit* bisogna considerare che ogni volta bisogna analizzare l'intera coda.
In ogni caso nel caso di allocazione con partizioni fisse il grado di multiprogrammazione è limitato dal numero di partizioni disponibili, e quindi il numero di processi che possono essere eseguiti contemporaneamente è limitato dal numero di partizioni disponibili. Inoltre esiste un grande problema di frammentazione sia interna che esterna, in quanto se un processo richiede meno memoria di quella disponibile in una partizione, la memoria rimanente non può essere utilizzata da altri processi, portando a problemi di frammentazione interna. Inoltre se un processo richiede più memoria di quella disponibile in una partizione, il processo non può essere caricato in memoria finché non viene liberata una partizione di dimensioni sufficienti, portando a problemi di frammentazione esterna.

Allocazione con partizioni variabili Quando si usa l'allocazione con partizioni variabili, la memoria viene suddivisa in partizioni di dimensioni variabili e ogni processo viene caricato in una partizione di dimensioni variabili. Questo tipo di allocazione è più flessibile rispetto all'allocazione con partizioni fisse, ma può portare a problemi di assegnazione di memoria a diversi *job*, e può portare a problemi di frammentazione esterna.
In questo caso il `SO` deve tener conto oltre alle partizioni allocate anche delle *buche* ovvero delle aree di memoria libere. Quando arriva un processo viene allocato a questo la prima buca che lo può contenere, e se la buca è più grande del necessario viene creata una nuova buca. Se invece la buca è più piccola del necessario, bisogna provvedere a deframmentare la memoria, ovvero spostare i processi in modo da creare una buca più grande.

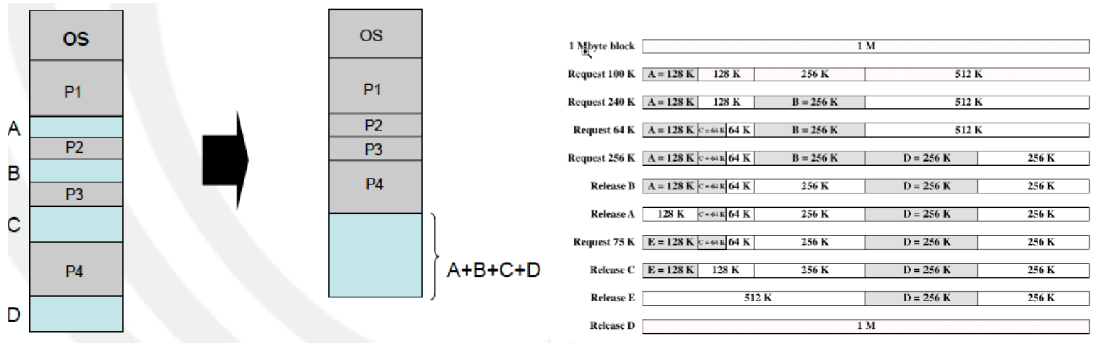
Stategie di *scheduling*Per l'allocazione con partizioni variabili esistono diverse strategie di *scheduling*:

- *First-fit*: viene allocata la prima buca che può contenere il processo, e se la buca è più grande del necessario viene creata una nuova buca.
- *Best-fit*: viene allocata la buca più piccola che può contenere il processo, e se la buca è più grande del necessario viene creata una nuova buca.
- *Worst-fit*: viene allocata la buca più grande che può contenere il processo, e se la buca è più grande del necessario viene creata una nuova buca.

Tipicamente *first-fit* è la migliore.

Tecniche di deframmentazione La deframmentazione è l'operazione che permette di spostare i processi in memoria in modo da creare buche più grandi che possono contenere processi più grandi. Esistono diverse tecniche di deframmentazione, le principali sono:

- **Compattazione**: ogni processo viene spostato in memoria in modo da creare buche più grandi, e le buche vengono unite in un'unica buca. Questo tipo di deframmentazione è molto veloce, ma richiede che i processi siano spostati in memoria e comunque richiede tempo.
- ***Buddy-system***: la memoria viene suddivisa in blocchi di dimensioni potenze di 2. Ogni volta che arriva un processo si procede a dividere la memoria in due finché una ulteriore divisione non permetterebbe di allocare il processo. Quando viene liberato un blocco di memoria, il `SO` verifica se il blocco può essere unito con un altro blocco adiacente, e se è possibile i due blocchi vengono uniti in un unico blocco. Questo tipo di deframmentazione è molto veloce, ma persiste il problema della frammentazione interna



(a) Compattazione (b) *Buddy-system*

Figura 9.1: Tecniche di deframmentazione

9.2.2 Paginazione

La paginazione è una tecnica per eliminare la frammentazione, sia esterna che interna, si basa sull'idea che lo spazio di indirizzamento di un processo può essere non-contiguo e che la memoria fisica può essere suddivisa in blocchi di dimensioni fisse detti *frame*, mentre la memoria logica viene divisa in blocchi di dimensioni fisse detti *page*.
Se vogliamo eseguire un programma il quale necessita n pagine allora bisogna trovare n *frame* liberi, indipendentemente che questi siano contigui o meno. Per trovare le pagine libere il `SO` mantiene una tabella delle pagine, la quale mappa ogni processo alle sue pagine associate agli *frame* fisici.

Tabulaizone degli indirizzi

Ogni indirizzo logico (generato dalla `CPU`) viene suddiviso in due parti:

- **Numero di pagina (p)**: il numero della pagina a cui appartiene l'indirizzo logico (se la pagina ha dimensione 2^n e la memoria ha dimensione 2^m allora il numero di pagina è di $m - n$ bit)
- **Offset (d)**: la posizione all'interno della pagina (se la pagina ha dimensione 2^n allora l'offset è di n bit)

L'indirizzo logico viene quindi rappresentato come una coppia di numeri (p, d), dove p è il numero della pagina e d è l'offset. Per ottenere l'indirizzo fisico, il `SO` deve tradurre il numero di pagina nell'indirizzo base della pagina fisica e sommarlo all'offset. L'indirizzo fisico viene quindi rappresentato come la coppia (f, d) dove f è il numero dell'*frame* e d è l'offset.

Implementazioni

Dato che l'efficienza è fondamentale in quanto la traduzione degli indirizzi deve essere eseguita in tempo reale, esistono diverse soluzioni per l'implementazione della traduzione degli indirizzi.

Implementazione tramite registri Tutte le *entry* della tabella sono archiviate in un registro, e ogni volta che viene generato un indirizzo logico, il `SO` deve accedere al registro per ottenere l'indirizzo fisico. Questo tipo di implementazione è molto veloce, ma questo al costo di un numero ridotto di *entry* ed un allungamento dei tempi di *context-switch*.

Implementazione in memoria Al posto di memorizzare l'intera tabella delle pagine in un registro, il `SO` memorizza solo un puntatore alla base della tabella delle pagine in un registro, ed un opzionale registro per la lunghezza della tabella delle pagine. Risulta quindi più veloce il *context-switch* in quanto la variazione riguarda solo il registro `PTBR` (*Page Table Base Register*) e il registro `PTLR` (*Page Table Length Register*). Questo però richiede due accessi alla memoria, uno per leggere la tabella delle pagine e uno per leggere l'*frame* fisico. Per ovviare a questo problema si usa una *cache* della tabella delle pagine, chiamata `TLB` (*Translation Look-aside Buffer*) che memorizza le traduzioni più recenti. La `TLB` è una memoria veloce che memorizza le traduzioni più recenti e permette di ridurre il numero di accessi alla memoria. Quando viene generato un indirizzo logico, il `SO` verifica se la traduzione è presente nella `TLB`, se è presente la traduzione viene eseguita in modo veloce ($< 10\%$ rispetto al *lookup* in memoria),

altrimenti viene eseguita la traduzione in memoria.
Il tempo di accesso effettivo medio alla memoria è dato dalla seguente formula:

$$T_{eff} = T_{TLB} + (1 - p) \cdot T_{mem} + p \cdot (T_{TLB} + T_{mem})$$

Dove T_{TLB} è il tempo di accesso alla `TLB`, T_{mem} è il tempo di accesso alla memoria e p è la probabilità che la traduzione sia presente nella `TLB`. Se la `TLB` è molto veloce e la probabilità che la traduzione sia presente è alta, il tempo di accesso effettivo alla memoria sarà molto basso.

Protezione

La protezione della memoria è implementata associando ad ogni *frame* un bit di protezione, che indica se il *frame* è accessibile o meno. Se un processo tenta di accedere a un *frame* non accessibile, il `SO` genera un'eccezione e il processo viene terminato. Esistono poi i bit di accesso che determinano se una pagina è modificabile o meno oppure se è eseguibile o meno.

Pagine condivise

La paginazione permette di avere più cope virtuali di una pagina in memoria ma una stessa copia fisica, ovvero più processi possono condividere la stessa pagina fisica. Questo permette condividere pagine (*read-only*) tra più processi mantenendo però la protezione della memoria.

Spazio di indirizzamento

Modernamente gli indirizzi sono o a `32` o a `64` bit, e quindi lo spazio di indirizzamento è di 2^{32} o 2^{64} byte, il quale è molto maggiore dello spazio di indirizzamento fisico. Per questo motivo vengono usati dei meccanismi per gestire il problema della dimensione della tabella delle pagine:

- Paginazione della tabella delle pagine: la tabella delle pagine viene suddivisa in pagine di dimensioni fisse, e ogni pagina della tabella delle pagine viene memorizzata in memoria. Questo permette di ridurre la dimensione della tabella delle pagine e di gestire lo spazio di indirizzamento in modo più efficiente.
- Tabella delle pagine invertita: la tabella delle pagine viene memorizzata in memoria e ogni *entry* della tabella delle pagine contiene l'indirizzo fisico della pagina. Questo permette di ridurre la dimensione della tabella delle pagine e di gestire lo spazio di indirizzamento in modo più efficiente.

La paginazione della tabella delle pagine è più complessa da implementare, ma permette di gestire lo spazio di indirizzamento in modo più efficiente. La tabella delle pagine invertita è più semplice da implementare, ma richiede più memoria e può portare a problemi di frammentazione.
La paginazione delle tabella delle pagine è una tecnica simile alla paginazione multi-livello, in quanto la tabella delle pagine viene suddivisa in più livelli di pagine. In questo caso la tabella delle pagine è suddivisa in più livelli di pagine, e ogni pagina della tabella delle pagine contiene un puntatore alla pagina successiva.

Paginazione con *hashing*

La paginazione con *hashing* è una tecnica di gestione della memoria che utilizza una funzione di *hashing* per restituire l'indirizzo fisico di una pagina. Riducendo il costo di ricerca delle pagine da $O(n)$ a $O(1)$, la paginazione con *hashing* è più veloce rispetto alla paginazione tradizionale.

9.3 Segmentazione

La segmentazione è una tecnica di gestione della memoria che permette di suddividere la memoria in segmenti di dimensioni variabili. Ogni segmento rappresenta un'unità logica del programma, come ad esempio una funzione o una variabile globale. La segmentazione permette di gestire la memoria in modo più flessibile rispetto alla paginazione, in quanto i segmenti possono avere dimensioni diverse e possono essere allocati in modo non contiguo.
La segmentazione è simile alla paginazione, ma invece di suddividere la memoria in pagine di dimensioni fisse, la segmentazione suddivide la memoria in segmenti di dimensioni variabili. L'indirizzo logico viene

quindi rappresentato come una coppia di numeri (s, d), dove s è il numero del segmento e d è l'offset. Per ottenere l'indirizzo fisico, il `SO` deve tradurre il numero del segmento nell'indirizzo base del segmento fisico e sommarlo all'offset. La tabella dei segmenti è simile alla tabella delle pagine, ma invece di memorizzare gli indirizzi fisici delle pagine, la tabella dei segmenti memorizza gli indirizzi fisici dei segmenti (base e limite). Anche per questa tabella esiste una `STBR` (*Segment Table Base Register*) e una `STLR` (*Segment Table Length Register*).
Prima di tradurre un indirizzo logico in un indirizzo fisico, il `SO` deve verificare se il numero del segmento è valido e se l'offset è compreso tra 0 e la dimensione del segmento. Se il numero del segmento è valido e l'offset è compreso tra 0 e la dimensione del segmento, il `SO` traduce l'indirizzo logico in un indirizzo fisico sommando l'indirizzo base del segmento all'offset. Se il numero del segmento non è valido o l'offset non è compreso tra 0 e la dimensione del segmento, il `SO` genera un'eccezione e il processo viene terminato.

Protezione Anche la segmentazione prevede l'uso di bit di protezione, che indicano se il segmento è accessibile o meno. Se un processo tenta di accedere a un segmento non accessibile, il `SO` genera un'eccezione e il processo viene terminato. Inoltre i segmenti possono essere protetti in modo da impedire la scrittura o la lettura da parte di altri processi.

Segmenti condivisi La segmentazione permette di avere più copie virtuali di un segmento in memoria ma una sola copia fisica, ovvero più processi possono condividere lo stesso segmento fisico. Questo permette di condividere segmenti (*read-only*) tra più processi mantenendo però la protezione della memoria.

Segmentazione e frammentazione Il sistema operativo deve allocare spazio in memoria per tutti i segmenti di un processo, e se un segmento richiede più memoria di quella disponibile, il processo non può essere caricato in memoria finché non viene liberato un segmento di dimensioni sufficienti. Questo porta a problemi di frammentazione esterna, in quanto i segmenti possono essere allocati in modo non contiguo e possono portare a problemi di frammentazione interna.

Paginazione v/s segmentazione

La paginazione e la segmentazione sono due tecniche di gestione della memoria che hanno vantaggi e svantaggi diversi. La paginazione permette una non esistenza di frammentazione (poca interna) ed l'allocazione dei frame non richiede uno specifico algoritmo. D'altro canto la segmentazione permette una maggiore consistenza tra vista fisica della memoria e vista logica della memoria, oltre alla associazione di protezione e condivisione a livello di segmento e non a livello di pagina. Tra gli svantaggi della segmentazione c'è la frammentazione esterna e la necessità di un algoritmo di allocazione per i segmenti dinamica, mentre tra gli svantaggi della paginazione c'è la separazione tra vista fisica e logica della memoria, la necessità di una tabella delle pagine e la frammentazione interna.

9.4 Segmentazione con paginazione

Una soluzione per eliminare i problemi di frammentazione esterna e interna è la segmentazione con paginazione combinando le due tecniche. In questo caso la memoria viene suddivisa in segmenti di dimensioni variabili e ogni segmento viene suddiviso in pagine di dimensioni fisse. Ogni segmento rappresenta un'unità logica del programma, come ad esempio una funzione o una variabile globale, mentre le pagine rappresentano l'unità fisica di allocazione della memoria.

`MULTICS` Il sistema operativo `MULTICS` è un esempio di sistema operativo che utilizza la segmentazione con paginazione. In questo caso la memoria viene suddivisa in segmenti di dimensioni variabili e ogni segmento viene suddiviso in pagine di dimensioni fisse. Ogni segmento rappresenta un'unità logica del programma, come ad esempio una funzione o una variabile globale, mentre le pagine rappresentano l'unità fisica di allocazione della memoria. La tabella dei segmenti è simile alla tabella delle pagine, ma invece di memorizzare gli indirizzi fisici delle pagine, la tabella dei segmenti memorizza gli indirizzi fisici dei segmenti (base e limite). Anche per questa tabella esiste una `STBR` (*Segment Table Base Register*) e una `STLR` (*Segment Table Length Register*).
La traduzione degli indirizzi logici in indirizzi fisici avviene in due passaggi:

- Il numero del segmento viene tradotto nell'indirizzo base del segmento fisico e l'offset viene sommato all'indirizzo base del segmento per ottenere l'indirizzo fisico della pagina.
- Il numero della pagina viene tradotto nell'indirizzo base della pagina fisica e l'offset viene sommato all'indirizzo base della pagina per ottenere l'indirizzo fisico.

In questo modo la segmentazione con paginazione permette di gestire la memoria in modo più flessibile rispetto alla paginazione e alla segmentazione, eliminando i problemi di frammentazione esterna e interna.

`ARMv7-A` Nella tecnologia `ARMv7-A` la segmentazione con paginazione è implementata divisa in quattro dimensioni:

- Super-Sezioni: $16MB$ di memoria fisica ($24bit$ di offset)
- Sezioni: $1MB$ di memoria fisica ($20bit$ di offset)
- Pagine grandi: $64KB$ di memoria fisica ($16bit$ di offset)
- Pagine piccole: $4KB$ di memoria fisica ($12bit$ di offset)

Ogni *entry* della tabella delle pagine può contenere o l'indirizzo direttamente di una sezione, oppure un puntatore al secondo livello della tabella delle pagine. Inoltre `ARM` ha due livelli di `TLB`:

- `L1 TLB`: piccola e veloce, memorizza le traduzioni più recenti. (32 *entry*)
- `L2 TLB`: più grande e lenta, memorizza le traduzioni meno recenti. (8+64 *entry*)

Capitolo 10

Memoria Virtuale

Andiamo ora ad analizzare la gestione di sistemi nei quali è possibile usare più memoria (`RAM`) di quella fisicamente installata. Questo è possibile grazie alla **memoria virtuale** o *swap*. La memoria virtuale è una tecnica che consente di utilizzare una porzione del disco rigido come se fosse memoria principale, permettendo così di eseguire più processi contemporaneamente anche quando la memoria fisica è insufficiente. In questo capitolo vedremo come funziona la memoria virtuale, i suoi vantaggi e svantaggi, e come viene implementata nei sistemi operativi moderni.

Concetti fondamentali Ogni pagina in memoria primaria puo essere "*swapped*" con una pagina in memoria secondaria, inoltre la memoria secondaria permette di avere un ulteriore livello di separazione tra il livello fisico e il livello logico. L'implementazione della memoria virtuale può essere realizzate o tramite paginazione su domanda (*demand paging*) o tramite segmentazione su domanda (*demand segmentation*).

10.1 Paginazione su domanda

Il principio della paginazione su domanda è quello di caricare in memoria solo le pagine necessarie per eseguire un processo. Quando un processo richiede una pagina che non è attualmente in memoria, si verifica un **page fault**, che è un'interruzione generata dal processore per indicare che la pagina richiesta non è presente in memoria.

Vantaggi Il principale vantaggio della paginazione su domanda è che richiede meno interazioni di `I/O` quando è necessario lo *swapping*, inoltre si una meno memoria fisica, poiché solo le pagine necessarie vengono caricate in memoria. Questo consente di eseguire più processi contemporaneamente e di utilizzare la memoria in modo più efficiente. Risulta necessario però conoscere lo stato di ogni pagina, se è in memoria o meno.

Stato delle pagine Per tenere traccia dello stato delle pagine, il sistema operativo aggiunge un bit di stato nella tabella delle pagine. Questo bit può assumere due valori: `0` se la pagina è nel disco e `1` se è in memoria. Quando un processo richiede una pagina, il sistema operativo controlla il bit di stato per determinare se la pagina è presente in memoria o meno. Se la pagina non è presente, il sistema operativo genera un *page fault* e carica la pagina richiesta dalla memoria primaria.

Gestione del page fault

Quando si verifica un *page fault*, il sistema operativo esegue una serie di operazioni per gestire la situazione. Queste operazioni possono variare a seconda del sistema operativo, ma in generale seguono questi passaggi:

1. `SO` verifica se la pagina è un riferimento valido, se non lo è viene generato un *segmentation fault* e il processo termina.
2. Viene caricato un frame vuoto
3. Viene eseguito lo *swapping* della pagina, ovvero viene copiata la pagina dal disco alla memoria fisica.

4. Modifica la tabella delle pagine per indicare che la pagina è ora presente in memoria.
5. Riprende l'esecuzione del processo, riprendendo l'istruzione che ha causato il *page fault*.

Tutti questi passaggi richiedono tempo e risorse, dunque il tempo effettivo di accesso alla memoria può essere calcolato come:

$$\texttt{EAT} = (1 - p) \cdot \texttt{T}_{\texttt{mem}} + p \cdot (\texttt{T}_{\texttt{page fault}}$$

Dove il `EAT` è il tempo medio di accesso alla memoria, `p` è la probabilità di un *page fault*, $\texttt{T}_{\texttt{mem}}$ è il tempo di accesso alla memoria calcolato come:

$$\texttt{T}_{\texttt{mem}} = (\texttt{T}_{\texttt{mem}} + \texttt{T}_{\texttt{TLB}}) \cdot \alpha + (2 \cdot \texttt{T}_{\texttt{mem}} + \texttt{T}_{\texttt{TLB}}) \cdot (1 - \alpha)$$

Dove $\texttt{T}_{\texttt{TLB}}$ è il tempo di accesso alla memoria secondaria e `alpha` è la probabilità di un *TLB hit*. La `TLB` è una cache che memorizza le traduzioni degli indirizzi virtuali in indirizzi fisici, riducendo così il numero di accessi alla tabella delle pagine.
Il tempo $\texttt{T}_{\texttt{page fault}}$ è dato da tre componenti principali:

- Il tempo necessario per l'interrupt
- Lo *swap* in lettura della pagina
- Il costo del riavvio del processo
- Successivamente il tempo di *swap* in scrittura della pagina

Rimpiazzo delle pagine Nel caso non ci siano *frame* liberi in memoria al momento della richiesta di una pagina, il sistema operativo deve prima cercare le pagine che sono in memoria e successivamente eseguire il rimpiazzo di una pagina. Il rimpiazzo delle pagine deve seguire un preciso algoritmo in modo da garantire la massima efficienza minimizzando il numero di *page fault*. Di seguito un esempio di algoritmo di rimpiazzo delle pagine:

1. Il sistema operativo verifica una tabella per determinare se questa è in memoria o meno.
2. Viene cercato un *frame* vuoto in memoria.
    a) Se presente salto al punto 4
    b) Se non presente viene eseguito l'algoritmo di rimpiazzo delle pagine.
3. Viene eseguito lo *swap* della pagina "vittima" sul disco.
4. Viene eseguito lo *swap* della pagina richiesta nel *frame* vuoto (o in un *frame* di una pagina "vittima").
5. Viene aggiornata la tabella delle pagine per indicare che la pagina è ora presente in memoria (e la pagina "vittima" non è più presente).
6. Viene ripresa l'esecuzione del processo, riprendendo l'istruzione che ha causato il *page fault*.

Come si può notare nel caso di un *page fault* senza *frame* liberi, il sistema operativo deve eseguire due accessi alla memoria per ogni *page fault*, uno per la pagina "vittima" e uno per la pagina richiesta. Questo raddoppia il tempo di esecuzione del processo di *page fault*, per ottimizzare il tempo di accesso alla memoria viene usato un `bit` che simboleggia se la pagina è stata modificata o meno. Se la pagina non è stata modificata, il sistema operativo può semplicemente scartarla senza eseguire lo *swap* sul disco. Questo riduce il numero di accessi alla memoria e migliora le prestazioni del sistema.

Problematiche della paginazione su domanda

La paginazione su domanda presenta alcune problematiche, tra cui:

- La scelta della pagina da rimpiazzare
- L'allocazione dei frame in memoria

per risolvere ciò sono stati sviluppati diversi algoritmi di rimpiazzo delle pagine e di allocazione dei frame.

10.2 Algoritmi di rimpiazzo delle pagine

Come anticipato l'obbiettivo principale degli algoritmi di rimpiazzo delle pagine è quello di minimizzare il numero di *page fault* e massimizzare l'efficienza del sistema. Valuteremo questi algoritmi in modo analitico andando a calcolare il numero di *page fault* su una stringa di indirizzi sapendo il numero di *frame* disponibili in memoria. In ogni caso il numero di *page fault* è inversamente proporzionale al numero di *frame* disponibili in memoria.

Algoritmo `FIFO`

L'algoritmo `FIFO` (*First In First Out*) è uno dei più semplici algoritmi di rimpiazzo delle pagine. In questo algoritmo, la pagina che è stata in memoria per più tempo viene rimpiazzata per prima. Questo algoritmo è semplice da implementare, ma può portare a situazioni in cui le pagine più frequentemente utilizzate vengono rimpiazzate, causando un aumento del numero di *page fault*. Di seguito un esempio di calcolo del numero di *page fault* usando l'algoritmo `FIFO`:
Assumiamo la stringa $7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1$ e 3 *frame* disponibili in memoria. Allora la tabella delle pagine nel tempo sarà:

Tempo	Pagina Richiesta	Frame 1	Frame 2	Frame 3
1	**7**	7	-	-
2	**0**	7	0	-
3	**1**	7	0	1
4	**2**	2	0	1
5	0	2	0	1
6	**3**	2	3	1
7	**0**	2	3	0
8	**4**	4	3	0
9	**2**	4	2	0
10	**3**	4	2	3
11	**0**	0	2	3
12	3	0	2	3
13	2	0	2	3
14	**1**	0	1	3
15	**2**	0	1	2
16	0	0	1	2
17	1	0	1	2
18	**7**	7	1	2
19	**0**	7	0	2
20	**1**	7	0	1

notiamo un totale di 12 *page fault*, notiamo come ad esempio quando nel punto 5 usiamo la pagina 0 questa e scattata al punto 6 e poi ri-caricata al punto 7 causando un *page fault* inutile in quanto rimpiazzando un altro *frame* non sarebbe successo. Questo algoritmo è semplice da implementare, ma è soggetto all'anomalia di *Belady*, questa anomalia si verifica quando aumentando il numero di *frame* disponibili in memoria, non è detto che il numero di *page fault* diminuisca. Questo è dovuto al fatto che l'algoritmo `FIFO` non tiene conto dell'uso delle pagine, ma solo del tempo in cui sono state caricate in memoria.

Algoritmo Ideale

L'algoritmo ideale è un algoritmo che dovrebbe prevedere il futuro e rimpiazzare la pagina che non verrà più utilizzata per il periodo di tempo più lungo. Questo algoritmo è teorico e non può essere implementato nella pratica ma nel caso di un programma compilato è possibile stimare la stringa di indirizzi e applicare l'algoritmo ideale. Di seguito un esempio di calcolo del numero di *page fault* usando l'algoritmo ideale e la stessa stringa di prima:
Assumiamo la stringa $7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1$ e 3 *frame* disponibili in memoria. Allora la tabella delle pagine nel tempo sarà:

Tempo	Pagina Richiesta	Frame 1	Frame 2	Frame 3
1	**7**	7	-	-
2	**0**	7	0	-
3	**1**	7	0	1
4	**2**	2	0	1
5	0	2	0	1
6	**3**	2	0	3
7	0	2	0	3
8	**4**	2	4	3
9	2	2	4	3
10	3	2	4	3
11	**0**	2	0	3
12	3	2	0	3
13	2	2	0	3
14	**1**	2	0	1
15	2	2	0	1
16	0	2	0	1
17	1	2	0	1
18	**7**	7	0	1
19	0	7	0	1
20	1	7	0	1

Notiamo come in questo caso il numero di *page fault* sia 9 rispetto ai 12 dell'algoritmo `FIFO`. Questo algoritmo è teorico e non può essere implementato nella pratica, ma può essere utilizzato come riferimento per valutare le prestazioni degli altri algoritmi di rimpiazzo delle pagine.

Algoritmo `LRU`

L'algoritmo `LRU` (*Least Recently Used*) è un algoritmo di rimpiazzo delle pagine che tiene conto dell'uso delle pagine. In questo algoritmo, la pagina che non è stata utilizzata per il periodo di tempo più lungo viene rimpiazzata per prima. Questo algoritmo è più efficiente rispetto all'algoritmo `FIFO`, poiché tiene conto dell'uso passato delle pagine e cerca di mantenere in memoria le pagine più frequentemente utilizzate. Ma comunque non è esente da problematiche, in quanto richiede una maggiore complessità di implementazione e può portare a un aumento del numero di accessi alla memoria per tenere traccia dell'uso delle pagine. Di seguito un esempio di calcolo del numero di *page fault*
Tralasciando l'esempio un problema di questo algoritmo è l'implementazione infatti richiede che venga tenuta traccia di tutte le pagine in memoria e del loro utilizzo, il che può richiedere una grande quantità di memoria aggiuntiva per memorizzare l'ultima referenza di ogni pagina in memoria oltre ad un tempo di ricerca dell'ultimo accesso, il che può portare a un aumento del numero di accessi alla memoria e quindi a un aumento del tempo di esecuzione del processo. Per questo motivo, l'algoritmo `LRU` viene approssimato in molti sistemi operativi moderni in quanto la sua reale implementazione porterebbe a più svantaggi che vantaggi.

Approssimazione dell'algoritmo `LRU` La tecnica principale di approssimazione dell'algoritmo `LRU` è quella di utilizzare un contatore per tenere traccia dell'uso delle pagine. In questo caso partiamo di algoritmo `LFU` (*Least Frequently Used*) che tiene traccia del numero di accessi a ciascuna pagina e rimpiazza la pagina meno frequentemente utilizzata. Questo algoritmo è più semplice da implementare rispetto all'algoritmo `LRU`, ma può portare a situazioni in cui le pagine più frequentemente utilizzate vengono rimpiazzate, causando un aumento del numero di *page fault*.

Approssimazione con `MFU` Un altro algoritmo di approssimazione dell'algoritmo `LRU` è l'algoritmo `MFU` (*Most Frequently Used*), che tiene traccia del numero di accessi a ciascuna pagina e rimpiazza la pagina più frequentemente utilizzata.

Approssimazion con timpiazzo *second chance* Un altro algoritmo di approssimazione dell'algoritmo `LRU` è l'algoritmo `second chance`, che sostanzialmente è una variante dell'algoritmo `FIFO` circolare con un bit di riferimento. In questo algoritmo è presente una "lancetta" che punta alla pagina successiva da rimpiazzare, questa se il `bit` di riferimento è 0 viene rimpiazzata, se è 1 viene azzerato e la lancetta

si sposta alla pagina successiva. Se una pagina in memoria viene richiesta, il sistema operativo imposta il `bit` di riferimento a 1. Questo algoritmo è più efficiente rispetto all'algoritmo `FIFO`, poiché tiene minimamente conto dell'uso delle pagine e cerca di mantenere in memoria le pagine più frequentemente utilizzate.

10.3 Allocazione dei frame

L'allocazione dei frame è un altro aspetto importante della gestione della memoria virtuale. Il sistema operativo deve decidere come allocare i frame in memoria per i processi in esecuzione. Bisogna tenere conto di diversi fattori, tri quali troviamo il fatto che ogni processo necessita di un numero minimo di pagine per essere eseguito, ogni istruzione interrotta per un *page fault* deve essere ripresa, ed il numero minimo di pagine è dato dal numero massimo di indirizzi specificabili in una istruzione.

Allocazione fissa L'allocazione fissa è un metodo di allocazione dei frame in cui il sistema operativo alloca in parti uguali i frame in memoria per ogni processo. Con n processi e m frame disponibili in memoria, ogni processo riceve $\frac{m}{n}$ frame. Oppure alloca un numero di frame proporzionalmente alla dimensione del processo. Questo metodo è semplice da implementare, ma può portare a una scarsa utilizzo della memoria se i processi hanno dimensioni molto diverse o se alcuni processi non utilizzano tutti i frame assegnati.

Allocazione variabile L'allocazione variabile è un metodo di allocazione dei frame in cui il sistema operativo alloca i frame in memoria in base alle esigenze dei processi. In questo metodo, il sistema operativo deve tenere traccia o calcolando il *working set* o calcolando il *page fault frequency* per determinare il numero di frame da allocare a ciascun processo. Questo metodo è più complesso da implementare rispetto all'allocazione fissa, ma può portare a un utilizzo più efficiente della memoria se i processi hanno dimensioni diverse o se alcuni processi non utilizzano tutti i frame assegnati.

Calcolo del *working set*

Il *working set* è un criterio per determinare il numero di frame da allocare a ciascun processo. Il *working set* viene calcolato sulla base del modello della località temporale, ovvero se e quando un processo accede a una pagina, è probabile che acceda a pagine vicine in un breve periodo di tempo, questo fenomeno viene calcolato analizzando gli accessi passati a una pagina. Il *working set* è definito come il numero delle pagine referenziate nell'intervallo di tempo $t - \Delta, t$, dove t è il tempo corrente e Δ è un intervallo di tempo definito dal sistema operativo. Anche qui la scelta del valore di Δ è critica, in quanto se troppo grande si rischia di non avere un numero sufficiente di frame per il processo, mentre se troppo piccolo si rischia di avere un numero eccessivo di frame per il processo. Per misurare il *working set* approssimiamo tramite dei timer e dei bit di riferimento, in questo se allo scadere del timer il `bit` di riferimento è 1, allora la pagina è parte del *working set*, altrimenti non lo è e viene scartata.
La richiesta totale dei frame di tutti i processi è data da:

$$\mathtt{D} = \sum_i \mathtt{WSS}_i$$

se `D` è maggiore del numero di frame disponibili in memoria, si verifica un errore noto come *thrashing*, in questo caso il sistema operativo deve decidere quali processi terminare o sospendere per liberare frame in memoria.

Trashing Il *thrashing* è una situazione che si verifica quando il numero di processi associati a un processo è minore rispetto ad una certa soglia di frame, in questo caso il sistema operativo deve eseguire un numero eccessivo di *page fault* per soddisfare le richieste del processo. Questo porta a un aumento del numero di accessi alla memoria e a una diminuzione delle prestazioni del sistema

Gestione della frequenza dei *page fault*

La gestione della frequenza dei *page fault* è un altro metodo per determinare il numero di frame da allocare a ciascun processo. In questo metodo, il sistema operativo stabilisce una soglia di frequenza dei *page fault* per ciascun processo e monitora il numero di *page fault* per ciascun processo. Se il numero di

page fault supera la soglia, il sistema operativo aumenta il numero di frame allocati al processo. Se il numero di *page fault* è inferiore alla soglia, il sistema operativo diminuisce il numero di frame allocati al processo.

Altre considerazione

Bisogna tenere le pagine di piccola dimensione in quanto così ridurremo la frammentazione interna. Se abbassiamo la dimensione delle pagine aumentiamo il numero di pagine e quindi deve essere aumentato il numero di frame e la dimensione della tabella delle pagine. Infine bisogna tenere conto dell'`I/O` *overhead* che porterebbe ad una pagina più grande per ammortizzare il costo di accesso alla memoria, và quindi trovato un compromesso tra le dimensioni delle pagine e il numero di frame disponibili in memoria.

Blocco di *frame* (*frame blocking*) Il blocco di *frame* è una tecnica che permette che una pagina non sia rimpiazzata, utile per processi *kernel* o per processi che richiedono un numero elevato di frame. In questo caso il sistema operativo deve tenere traccia dei frame bloccati e non rimpiazzarli durante il rimpiazzo delle pagine.

Capitolo 11

Gestione della memoria secondaria

La memoria secondaria è un'area di memoria non volatile, che viene utilizzata per memorizzare i dati in modo permanente. A differenza della memoria principale (`RAM`), la memoria secondaria non perde i dati quando il computer viene spento e non necessita di un '*refresh*' dei condensatori. Esistono diversi tipi di memoria secondaria, tra cui nastri magnetici, dischi rigidi, unità a stato solido (SSD) e supporti ottici, si rende quindi necessario un sistema di gestione della memoria secondaria per garantire che i dati siano memorizzati e recuperati indipendentemente dalla loro posizione fisica. La gestione della memoria secondaria è un aspetto fondamentale dei sistemi operativi moderni, poiché consente di ottimizzare l'uso delle risorse e migliorare le prestazioni del sistema. In questa sezione, esploreremo i principali aspetti della gestione della memoria secondaria, tra cui la memorizzazione dei dati, l'organizzazione dei file e la gestione dello spazio libero.

11.1 Tipologia della memoria secondaria

Nastri magnetici I nastri magnetici sono un tipo di memoria secondaria che utilizza un nastro magnetico per memorizzare i dati. Tuttora utilizzati per il backup e l'archiviazione a lungo termine, i nastri magnetici offrono una capacità di memorizzazione elevata a un costo relativamente basso. Tuttavia, la velocità di accesso ai dati è molto più lenta rispetto ad altre forme di memoria secondaria, come i dischi rigidi o gli SSD. Questo non tanto a causa della velocità di lettura e scrittura, quanto per il fatto che i nastri magnetici sono dispositivi sequenziali, il che significa che i dati devono essere letti in un ordine specifico. Di conseguenza, l'accesso casuale ai dati su un nastro magnetico può richiedere molto tempo. A differenza di altre forme di memoria secondaria, i nastri magnetici non seguono uno standard di archiviazione dei file, ma piuttosto un dei formati di archiviazione proprietari all'azienda produttrice. I nastri magnetici sono spesso utilizzati in applicazioni di archiviazione a lungo termine, come il backup dei dati e l'archiviazione di grandi volumi di dati.

Dischi rigidi I dischi rigidi (HDD) sono un tipo di memoria secondaria che utilizza piatti magnetici rotanti per memorizzare i dati. I dischi rigidi offrono una capacità di memorizzazione elevata e una velocità di accesso ai dati relativamente rapida, rendendoli una scelta popolare per l'archiviazione dei dati. Tuttavia, i dischi rigidi sono più lenti rispetto agli SSD e possono essere soggetti a guasti meccanici a causa delle parti mobili. Le operazioni di lettura e scrittura sui dischi rigidi avvengono tramite testine magnetizzanti che si muovono sopra i piatti rotanti, se si vuole scrivere un '0' o un '1' su un piatto, la testina magnetizza o de-magnetizza una piccola area del piatto. Quando si legge un dato, la testina rileva il campo magnetico presente nell'area del piatto e lo converte in un segnale elettrico. I dischi rigidi sono ampiamente utilizzati nei computer desktop e nei laptop, nonché nei server e nei data center per l'archiviazione di grandi volumi di dati.

SettoreUn settore è l'unità di memorizzazione più piccola su un disco rigido. Ogni settore ha una dimensione fissa, solitamente di 512 byte o 4 KB, e contiene un blocco di dati. I settori sono raggruppati in *cluster*, che sono la minima unità di allocazione dei file del sistema operativo. Quando un file viene memorizzato su un disco rigido, viene suddiviso in blocchi di dati che vengono memorizzati in settori e cluster. La dimensione del settore e del cluster può influenzare le prestazioni del disco rigido e l'efficienza dell'archiviazione dei dati.

Tempi di accessoI tempi di accesso ai dati su un disco rigido sono influenzati da diversi fattori, tra cui la velocità di rotazione dei piatti, la posizione dei dati sul disco e il tempo necessario per spostare la

testina di lettura/scrittura. La velocità di rotazione dei piatti è misurata in giri al minuto (RPM) e può variare da 5400 a 15000 RPM. Viene quindi calcolato il tempo di latenza come segue: *Seek time* + *Latency time* + *Transfer time*. Il *Seek time* è il tempo necessario per spostare la testina di lettura/scrittura sulla posizione corretta del disco, il *Latency time* è il tempo necessario per far ruotare il piatto fino a quando il settore desiderato si trova sotto la testina e il *Transfer time* è il tempo necessario per trasferire i dati dal disco alla memoria principale. La somma di questi tre tempi determina il tempo totale di accesso ai dati su un disco rigido. Parlando di numeri si può dire che un disco rigido da 7200 RPM ha un tempo di *seek* di circa 9 ms, un tempo di latenza di circa 4.16 ms ed il tempo di trasferimento come la somma tra '*disk-to-buffer*' `1030Mb/s` e '*buffer-to-CPU*' `300MB/s`. Quindi il tempo totale di accesso ai dati su un disco rigido da 7200 RPM è di circa 13.16 ms. La maggior parte di questo è dato dal tempo di *seek* e dal tempo di latenza, mentre il tempo di trasferimento è relativamente breve. Tuttavia, i dischi rigidi più veloci, come quelli da 10000 o 15000 RPM, possono ridurre significativamente il tempo totale di accesso ai dati.

Minimizzazione *seek time*Per minimizzare il tempo di *seek* e migliorare le prestazioni del disco rigido, i sistemi operativi utilizzano diverse tecniche di ottimizzazione. Una di queste tecniche è la *deframmentazione*, che riorganizza i dati sul disco in modo che siano memorizzati in settori contigui. Questo riduce il numero di movimenti della testina di lettura/scrittura e migliora le prestazioni complessive del disco. Un'altra tecnica è la *cache*, che memorizza temporaneamente i dati più frequentemente utilizzati nella memoria principale per ridurre il numero di accessi al disco rigido.

Unità a stato solido (SSD) Le unità a stato solido (SSD) sono un tipo di memoria secondaria che utilizza memoria flash per memorizzare i dati. Gli SSD offrono una velocità di accesso ai dati molto più rapida rispetto ai dischi rigidi, poiché non hanno parti mobili e possono accedere ai dati in modo casuale. Questi usano delle celle di memoria flash `NAND`, che sono costituite da transistor a effetto di campo (FET) e memorizzano i dati in forma elettrica. Il costo delle SSD è diminuito negli ultimi anni, rendendole sempre più popolari per l'archiviazione dei dati nei computer e nei dispositivi mobili, ma dischi di grandi dimensioni sono ancora più costosi rispetto agli HDD. Gli SSD sono disponibili in diverse forme e fattori di forma, tra cui unità SATA, unità NVMe e unità M.2.

11.2 *Scheduling* degli accessi al disco

Dal punto di vista logico il disco rigido è un dispositivo unidimensionale composto da un insieme di blocchi di memoria detti *cluster*, mentre dal punto di vista fisico è un dispositivo bidimensionale composto da un insieme di cilindri e tracce. Si rende dunque necessario un sistema di *scheduling* per ottimizzare l'accesso ai dati memorizzati su disco. Lo *scheduling* degli accessi al disco è un processo che determina l'ordine in cui le richieste di accesso ai dati vengono elaborate dal disco rigido. Esistono diversi algoritmi di *scheduling* degli accessi al disco, ognuno con i propri vantaggi e svantaggi. Di seguito sono riportati alcuni dei più comuni algoritmi di *scheduling* degli accessi al disco:

- **FCFS (First-Come, First-Served)**: Questo algoritmo elabora le richieste di accesso ai dati nell'ordine in cui arrivano. Sebbene sia semplice da implementare, può portare a tempi di attesa elevati se ci sono richieste di accesso lontane tra loro.
- **SSTF (Shortest Seek Time First)**: Questo algoritmo elabora le richieste di accesso ai dati in base alla distanza dalla posizione corrente della testina di lettura/scrittura. In questo modo si riduce il tempo di *seek*, ma può portare a un fenomeno noto come minimo locale *starvation*, in cui alcune richieste di accesso ai dati possono rimanere in attesa per un lungo periodo di tempo se ci sono sempre richieste più vicine alla testina.
- **SCAN** (ascensore): Questo algoritmo muove la testina di lettura/scrittura in una direzione (ad esempio, verso l'esterno del disco) e soddisfa tutte le richieste di accesso ai dati lungo il percorso. Una volta raggiunta la fine del disco, la testina torna indietro e soddisfa le richieste rimanenti. Questo algoritmo è più equo rispetto al SSTF, ma può comunque portare a tempi di attesa elevati per alcune richieste.
- **C-SCAN (Circular SCAN)** (spazzaneve): Questo algoritmo è simile al SCAN, ma quando la testina raggiunge la fine del disco, torna all'inizio senza elaborare le richieste lungo il percorso. Questo garantisce che tutte le richieste siano elaborate in modo equo e riduce i tempi di attesa per le richieste lontane dalla posizione corrente della testina.

- **C-LOOK** (spazzaneve circolare): Questo algoritmo è simile al C-SCAN, ma la testina non raggiunge la fine del disco prima di tornare all'inizio. Invece, si sposta solo fino alla richiesta più lontana nella direzione corrente e poi torna all'inizio. Questo riduce ulteriormente i tempi di attesa per le richieste lontane dalla posizione corrente della testina.

- **N-step SCAN**: Questo algoritmo è una variante del SCAN che consente di elaborare delle code diverse di richieste di accesso ai dati. In questo modo, le richieste vengono suddivise in gruppi di dimensioni fisse e ogni gruppo viene elaborato in modo indipendente. Quando un gruppo è completo questo viene elaborato in modo simile al SCAN, ma le richieste all'interno del gruppo possono essere elaborate in qualsiasi ordine. Questo algoritmo può migliorare le prestazioni del disco rigido e ridurre i tempi di attesa per le richieste di accesso ai dati.

In generale nessun algoritmo di *scheduling* degli accessi al disco è perfetto e la scelta dell'algoritmo dipende dalle esigenze specifiche del sistema e dalle caratteristiche delle richieste di accesso ai dati. Tuttavia, gli algoritmi di *scheduling* degli accessi al disco possono migliorare significativamente le prestazioni del disco rigido e ridurre i tempi di attesa per le richieste di accesso ai dati.

11.3 Gestione del disco

Formattazione La formattazione di basso livello (o fisica) è il processo di preparazione di un disco rigido o di un altro dispositivo di memorizzazione per l'uso. Durante la formattazione, il disco viene suddiviso in settori e tracce, creando una struttura logica per l'archiviazione dei dati. La formattazione fisica aggiunge inoltre lo spazio di correzione degli errori (`ECC`). Viene anche caricato sulk disco il *boot sector*, che contiene le informazioni necessarie per avviare il sistema operativo.
La formattazione di alto livello (o logica) è il processo di creazione di un file system sul disco. Durante la formattazione logica, il disco viene suddiviso in partizioni e viene creato un file system che consente al sistema operativo di gestire i file e le directory. La formattazione logica crea anche una tabella di allocazione dei file (FAT) o un inode, che tiene traccia dei file memorizzati sul disco e delle loro posizioni fisiche. La formattazione logica può essere eseguita su un disco già formattato fisicamente, ma non è necessario eseguire la formattazione fisica ogni volta che si formatta un disco.

Gestione blocchi difettosi Come anticipato in precedenza, i dischi rigidi e gli SSD oltre a memorizzare i dati, memorizzano anche informazioni di correzione degli errori. Solitamente viene calcolato un `ECC` il quale viene memorizzato alla fine del settore e viene utilizzato per verificare l'integrità dei dati memorizzati. Se quando agiamo in lettura su un settore, dopo il calcolo del `ECC` il risultato non è corretto, il settore viene considerato difettoso. Per gestire i *bad block* vengono usate o tecniche *off-line* quali individuazione al momento di formattazione con rimozione e/o marcatura nella `FAT`, utilità come `chkdsk` o `fsck` o tecniche *on-line* quali la mappatura dei settori difettosi in settori di riserva. In questo caso i settori di riserva devono essere riservati in fase di formattazione e non devono essere utilizzati per memorizzare i dati, inoltre questi settori devono essere presenti su ogni cilindro del disco in modo da ridurre il tempo di *seek*.

Interfacce di connessione Le interfacce di connessione sono i protocolli e le tecnologie utilizzate per collegare i dispositivi di memorizzazione al computer. Esistono diverse interfacce di connessione, tra cui:

- **SATA (Serial Advanced Technology Attachment)**: Questa è una delle interfacce di connessione più comuni per i dischi rigidi e gli SSD. SATA offre velocità di trasferimento dati elevate e supporta la connessione di più dispositivi a un singolo controller. Le versioni più recenti di SATA, come SATA III, offrono velocità di trasferimento dati fino a 6 Gbps.

- **IDE (Integrated Drive Electronics)**: Questa è un'interfaccia di connessione più vecchia rispetto a SATA usa trasmissione parallela e supporta solo un numero limitato di dispositivi. IDE è stato ampiamente utilizzato nei computer desktop e nei laptop, ma è stato gradualmente sostituito da SATA.

Gestione spazio di *swap* Il sistema operativo potrebbe decidere di adottare una sezione della memoria secondaria per estendere la memoria principale (vedi capitolo precedente). Per farlo viene o presa una normale sezione dal *file system*, il che può risultare inefficiente, oppure viene creata una partizione

dedicata. In questo caso viene risparmiato spazio in quanto non devono essere memorizzati i dati di *file system*, questa è la soluzione più comune.

Capitolo 12

File System

Il *file system* fornisce il meccanismo fondamentale per la memorizzazione e l'accesso ai dati e ai programmi su un computer. Esso consiste nella collezione di file e nella struttura che li organizza.

12.1 L'interfaccia del *file system*

File Il file è l'astrazione logica per accedere ai supporti di memorizzazione. Un file è costituito da uno spazio di indirizzamento logico e contiguo e raggruppa un insieme di informazioni identificate da un nome.

Attributi di un file Ogni file è caratterizzato da un insieme di attributi, che non sono propriamente parte del file, ma sono informazioni memorizzate nel *file system* e associate al file. Gli attributi di un file sono:

- Il nome del file
- Il tipo di file
- La posizione del file (nella memoria secondaria)
- La dimensione del file
- L'eventuale protezione del file (permessi di accesso)
- La data di creazione, modifica e accesso

Operazioni sui file Le operazioni fondamentali che possono essere eseguite su un file sono:

- Creazione - Si crea uno spazio sul disco per il file e si memorizzano gli attributi del file.
- Scrittura - Si usano delle *system call* per scrivere i dati nel file, conoscendo il puntatore alla posizione del file e della prossima scrittura.
- Lettura - Si usano delle *system call* per leggere i dati dal file, conoscendo il puntatore alla posizione del file e della prossima lettura.
- Riposizionamento - Si usano delle *system call* per modificare il puntatore alla posizione del file.
- Eliminazione - Si eliminano gli attributi del file e si libera lo spazio occupato dal file.
- Troncamento - Si accorcia il file, eliminando i dati oltre una certa posizione.
- Apertura - Si cerca nella struttura del *file system* il file e si copia il file nella memoria principale e si inserisce un riferimento al file aperto nella tabella dei file aperti.
- Chiusura - Si elimina il riferimento, si salvano eventuali modifiche e si libera lo spazio occupato dal file aperto.

Struttura di un file Il tipo di un file ne può definire la struttura, nei sistemi linux i file sono semplicemente sequenze di byte, mentre in altri sistemi operativi i file possono essere strutturati come "record" semplici oppure tramite strutture più complesse, gli ultimi due casi possono essere in alcuni casi emulati tramite il primo caso.

Metodi di accesso I metodi di accesso ai file sono:

Accesso sequenzialeI dati sono letti in modo sequenziale, il puntatore alla posizione del file viene spostato automaticamente alla fine del file dopo la lettura. In questo caso le operazioni consentite sono: *read next*, *write next*, *reset*. Non è permessa l'operazione di *rewrite* dato che c'è il rischio di sovrascrivere i dati.

Accesso direttoI dati sono letti in modo diretto, come in un database, il puntatore alla posizione del file viene spostato automaticamente alla fine del file dopo la lettura. In questo caso le operazioni consentite sono: *read*, *write*, *position to* , *read next*, *write next*, *rewrite*.

12.2 Struttura delle *directory*

Ogni partizione di disco è organizzata in due zone di memoria: la prima contiene le informazioni sulle directory e la seconda contiene i file. Andando ora ad analizzare la struttura delle directory, possiamo astrarre questa come una collezione di nodi che contengono le informazioni sui file (viste sopra) e che sono organizzati in una struttura ad albero.

Operazioni sulle directory Le operazioni fondamentali che possono essere eseguite su una directory sono:

- Aggiunta di un file - Si crea un nuovo nodo nella directory e si memorizzano gli attributi del file.
- Cancellazione di un file - Si elimina il nodo dalla directory e si libera lo spazio occupato dal file.
- Visualizzazione - Si visualizzano i nodi della directory.
- Rinominare un file - Si modificano gli attributi del file.
- Ricercare un file - Si cerca il nodo nella directory e si restituisce il puntatore al file, eventualmente si cerca anche nelle sottodirectory.
- Attraversare il *file system* - Si attraversa il *file system* a partire dalla directory corrente, si possono usare le operazioni di *cd*, *ls*, *pwd* per spostarsi tra le directory e visualizzare i file.

Organizzazione logica delle directory

L'organizzazione logica delle directory è stata sviluppata tenendo conto di tre obbiettivi: l'efficienza, l'accesso ad un file deve essere veloce, la nomenclatura, i nomi dei file devono essere univoci per lo stesso utente ma possono essere duplicati per utenti diversi oltre a più nomi per lo stesso file, e il raggruppamento, i file devono essere raggruppati in modo logico.

Directory ad un livello In questo caso esiste una sola directory per tutti gli utenti, i file sono identificati da un nome univoco. Questo metodo è poco usato in quanto sussistono problemi di nomenclatura e di raggruppamento.

Directory a più livelli In questo caso ogni utente ha una sua directory, i file sono identificati da un nome univoco per ogni utente. Viene introdotti il concetto di "*path*" che è una sequenza di nomi di directory che portano al file. Rispetto al caso precedente, questo metodo permette di usare lo stesso nome per file diversi ed migliora la ricerca dei file, ma non risolve il problema di raggruppamento. Inoltre i programmi di sistema per essere condivisi tra più utenti devono essere memorizzati in una directory comune e per tenerne traccia si usano i puntatori simbolici, in linux si usa la *directory* `PATH` per memorizzare i puntatori simbolici.

Directory ad albero Questo metodo è una evoluzione del metodo a più livelli, in questo caso ogni directory può contenere altre directory, formando una struttura ad albero. In questo modo si risolve il problema di raggruppamento e si migliora la ricerca dei file. Viene introdotto il concetto di "*directory corrente*" che è la directory in cui ci si trova al momento e che viene usata come punto di partenza per le operazioni di ricerca. Inoltre viene introdotto il concetto di percorso assoluto e relativo, il primo è un percorso che parte dalla radice dell'albero, mentre il secondo è un percorso che parte dalla directory corrente.

Directory a grafo aciclico Questo metodo è una evoluzione del metodo ad albero, in questo caso un file può essere "referenziato" da più directory, formando una struttura a grafo aciclico. In questo modo si implementa una prima condivisione dei file tra più utenti, ma si introduce il problema di eliminazione dei file, in quanto se un file è referenziato da più directory, la sua eliminazione deve essere gestita in modo da non lasciare puntatori a file non esistenti. Per risolvere questo problema si usano i puntatori simbolici, che sono dei puntatori a file che possono essere eliminati senza eliminare il file stesso. Esistono due tipi di puntatori, i link simbolici e gli *hard link*, i primi sono dei puntatori a file che possono essere eliminati senza eliminare il file stesso, questi contengono il nome vero del file e se il questo viene eliminato, il puntatore simbolico diventa un "link rotto", mentre gli *hard link* sono dei contatori che contengono il numero di puntatori a file, quando si elimina uno di questi, il contatore viene decrementato e il file viene eliminato solo quando il contatore arriva a zero.

Directory a grafo Questo metodo è una evoluzione del metodo a grafo aciclico, in questo caso anche le *directory* possono essere referenziate da più *directory*, formando una struttura a grafo, che eventualmente contengono dei cicli, in questo caso bisogna gestire queste situazioni nel caso della ricerca di un file, in quanto si potrebbe entrare in un ciclo infinito. Per risolvere questo problema alcuni `fs` permettono i soli collegamenti tra file mentre altri effettuano dei controlli per evitare i cicli.

Mount* di un *file system Il *mount* di un *file system* è l'operazione che permette di rendere questo modulare permettendo di attaccare e staccare i *file system* in altri *file system* già montati. Bisogna definire un punto di attacco, che è una directory in cui viene montato il *file system* (*mount point*). Uno stesso *file system* può essere montato in più punti di attacco, da *file system* diversi, può essere quindi condiviso tra più utenti anche su dispositivi diversi.

Condivisione di file La condivisione di file è un'operazione che permette di condividere file tra più utenti, in questo modo si possono usare gli stessi file senza doverli copiare e tenerne manualmente traccia. Per implementare ciò, però, è necessario implementare prima dei meccanismi di protezione. Il *file system* più usato per la condivisione di file è il *Network File System* che permette la condivisione tramite la rete.

Protezione Il proprietario di un file deve essere in grado di gestire le operazioni eseguibili su questo. In alcuni sistemi obsoleti veniva usata una matrice dei permessi, in sistemi moderni viene usata o la più compatta lista di accesso (Windows) o la divisione degli utenti in tre classi (*user, group, others*)

12.3 Implementazione del File system

La memoria secondaria è organizzata a livelli, infatti si parte del controllore di `I/O` che gestisce i dispositivi di memorizzazione per poi passare al *basic file system* che gestisce le primitive di `R/W` sui singoli blocchi di memoria, a seguire il *file organization module* che gestisce i file e le directory, infine il *logical file system* che gestisce le operazioni sui file e le directory. Ogni livello ha una sua interfaccia e comunica con il livello sottostante tramite delle *system call*.

Strutture dati Per gestire questo complesso sistema vengono usate diverse strutture dati, alcune memorizzate in memoria principale e altre memorizzate su disco, queste dipendono strettamente dal tipo di *file system* e dal tipo di `SO` ma esistono delle strutture generali che vengono usate in tutti i *file system*: su disco possiamo trovare infatti il blocco di boot, che contiene le informazioni necessarie per l'avvio del `SO`, il blocco di controllo delle partizioni, che contiene le informazioni sulle partizioni del disco, la struttura delle directory, che contiene le informazioni sui file e le directory, i descrittori di file, che contengono le informazioni sui file e puntatori a blocchi di dati. Sulla memoria principale troviamo invece la tabella dei file partizioni, struttura della directory, replicate entrambe dalla memoria secondaria, la

tabella globale dei file aperti, che contiene i puntatori ai file descriptor e la tabella dei file aperti per processo che contiene i puntatori ai file aperti per ogni processo.

Allocazione dello spazio su disco

Quando si crea un file, il *file system* deve allocare lo spazio necessario per memorizzare i dati del file. Esistono diversi metodi di allocazione dello spazio su disco, ognuno con i suoi vantaggi e svantaggi.

Allocazione contigua In questo metodo, lo spazio per il file viene allocato in modo contiguo, ovvero i blocchi di dati del file sono memorizzati in posizioni contigue sul disco. Questo metodo permette di memorizzare i file conoscendone solo la posizione iniziale e la dimensione, ma si verifica il problema della frammentazione, ovvero la divisione dello spazio su disco in piccole porzioni non utilizzabili. Inoltre, questo metodo non permette di estendere i file in certe situazioni, in quanto non è possibile allocare spazio contiguo per il file.

Allocazione a lista In questo metodo, si tene traccia del punto di inizio del file e del punto di fine del file, andando a memorizzare su ogni singolo blocco di dati del file il puntatore al blocco successivo, se presente. Questo metodo permette di allocare lo spazio in modo non contiguo, ma richiede più spazio per memorizzare i puntatori e non garantisce un accesso casuale ai dati. Inoltre si perde completamente la località dei dati, in quanto i blocchi di dati possono essere memorizzati in posizioni non contigue sul disco.

Allocazione a lista - variante extent In questo metodo, viene usata la stessa idea del metodo a lista, ma al posto di memorizzare in blocchi di dati in posizione casuale si memorizzano in blocchi di dati contigui finché possibile. In questo modo si riesce a mantenere una quasi contiguità e località dei dati, e si continua a memorizzare i puntatori ai blocchi di dati.

Esempio `FAT`Un esempio di allocazione a lista è il `FAT` (File Allocation Table), che è una tabella che tiene traccia dei blocchi di dati allocati per ogni file. Ogni file ha un puntatore al primo blocco di dati e ogni blocco di dati ha un puntatore al blocco successivo. Il limite di questo metodo è dato dalla dimensione di ogni *entry* della tabella, che può essere di 12, 16 o 32 bit, a seconda della dimensione del disco. Se questa è ad esempio di 32 bit allora si potranno allocare fino a 2^{32} blocchi di dati, che corrispondono a 4 GB di spazio su disco. Questo metodo è stato usato in passato per i dischi rigidi e le memorie flash, ma è stato superato da metodi più moderni. (`NTFS` e `exFAT` sono i successori del `FAT`).

Allocazione indicizzata Ogni file è caratterizzato da un blocco indice (*index block*) che contiene la tabella degli indirizzi dei blocchi di dati del file. Viene dunque memorizzato solo l'indirizzo del blocco indice nella *directory*. In questo modo possiamo garantire l'accesso casuale ai dati, ma non la località, ciò in quanto i blocchi di dati possono essere memorizzati in posizioni non contigue sul disco. Inoltre, questo metodo richiede comunque un ulteriore blocco di dati per memorizzare la tabella degli indirizzi, che può essere di dimensioni variabili a seconda della dimensione del file. Inoltre bisogna considerare che la dimensione del blocco indice limita la dimensione del file, in quanto se il file richiede più blocchi di dati di quelli che possono essere memorizzati nel blocco indice. Si potrebbe aumentare la dimensione del blocco indice, ma ciò porterebbe ad un aumento dello "spreco" di memoria per tutti i file, in quanto non tutti i file richiedono l'intero blocco indice. Per la traduzione degli indirizzi una volta che si conosce la dimensione del blocco (N) e l'indirizzo logico (X) allora nella posizione X/N si trova il blocco indice e $X\%N$ è l'*offset* all'interno del blocco indice. Questa traduzione a differenza delle altre viene effettuata tramite *software* e non tramite *hardware*, in quanto questo è un calcolo relativamente semplice e veloce da effettuare.
Nei sistemi reali si usano o Indici multi-livello, o lo schema concatenato oppure con lo schema combinato.

Indici multi-livello Nell'indicizzazione multi-livello, il blocco indice è memorizzato in più livelli, in modo da poter memorizzare file più grandi rispetto alla dimensione del blocco indice. In questo modo l'indirizzo del blocco indice di primo livello è dato da X/N^2, l'indirizzo del blocco indice di secondo livello è dato da $(X\%N^2)/N$ e l'*offset* all'interno del blocco indice è dato da $(X\%N^2)\%N$. Dato che non è detto che vadano memorizzati tutti i blocchi di dati, si usano dei puntatori nulli per indicare i blocchi di dati non allocati, comunque dovremmo allocare almeno un blocco di primo livello ed un blocco di secondo livello.

Schema concatenato In questo schema la tabella dei file memorizza il puntatore al primo blocco indice, il quale se non è sufficiente a memorizzare gli indirizzi di tutti i blocchi di dati, memorizza nell'ultima posizione un puntatore al blocco indice successivo. Possiamo quindi, in teoria, memorizzare un numero infinito di blocchi di dati, ma in pratica si ha un limite dato dalla dimensione del disco. L'indirizzamento con X l'indirizzo logico ed N la dimensione del blocco, è dato da $X/(N(N-1))$ per il numero di blocco dell'indice e $X\%(N(N-1)) = R$ dove $R/N-1$ è l'offset del blocco indice e $R\%N$ sono gli offset del blocco di dati.

Schema combinato - i-node In questo schema si usano diversi livelli di indici, in modo da poter memorizzare file più grandi rispetto alla dimensione del blocco indice mantenendo comunque un ridotto uso della memoria per i file più piccoli. Nel caso particolare `UNIX` usa 10 puntatori diretti, 1 puntatore indiretto singolo (1 livello), 1 puntatore indiretto doppio (2 livelli) e 1 puntatore indiretto triplo (3 livelli). I file più grandi vengono memorizzati con livelli più alti di indici, mentre i file più piccoli vengono memorizzati con livelli più bassi di indici. In questo modo si riesce a mantenere un buon compromesso tra spazio e velocità di accesso ai dati.

Gestione delle *directory*

Lo stesso meccanismo per la memorizzazione dei file viene usato per la memorizzazione delle directory anche se queste non contengono dati, ma solo puntatori a file. In questo modo possono sorgere problematiche sul come questo contenuto viene memorizzato e come viene gestito l'accesso al contenuto della directory.

Implementazione con lista di nomi In questo metodo, la directory è memorizzata come una lista di nomi di file, con i puntatori ai file. Questo metodo è semplice da implementare, ma è poco efficiente sia in lettura che in scrittura e rimozione.

Implementazione con tabella hash In questo metodo, la directory è memorizzata come una tabella hash, con i puntatori ai file, il nome del fine viene passato come chiave per la tabella hash. Questo metodo è più efficiente rispetto al metodo precedente, ma richiede più spazio per memorizzare la tabella hash e vanno gestite le collisioni. La tabella hash potrebbe "sprecare" spazio per ogni *directory* in quanto non è detto che tutti i puntatori siano usati, ma in questo caso si potrebbe usare una tabella hash con un numero di *entry* maggiore rispetto al numero di file memorizzati nella directory, si devono però gestire le collisioni.

12.4 Gestione dello spazio libero

La gestione dello spazio libero è un'operazione fondamentale per il *file system*, in quanto permette di tenere traccia dei blocchi di dati liberi e di allocare lo spazio necessario per i file. Esistono diversi metodi per gestire lo spazio libero, ognuno con i suoi vantaggi e svantaggi. In linea generale quando si crea un file si cercano dei blocchi di dati liberi nella struttura dati usata per tenere traccia dello spazio libero, per la rimozione di un file si liberano i blocchi di dati occupati dal file e si aggiornano le strutture dati usate per tenere traccia dello spazio libero.

Vettore di `bit` In questo metodo, si usa un vettore di bit per tenere traccia dei blocchi di dati liberi. Ogni bit del vettore rappresenta un blocco di dati, se il bit è 0 il blocco è libero, se il bit è 1 il blocco è occupato. Questo metodo è semplice da implementare ed è efficiente se riesce ad essere mantenuto in memoria principale, inoltre è semplice capire dove si trovano dei blocchi contigui.

Lista concatenata In questo metodo, si usa una lista concatenata per tenere traccia dei blocchi di dati liberi. Ogni blocco di dati libero contiene un puntatore al blocco di dati libero successivo. Questo metodo è più complesso da implementare rispetto al metodo precedente, ma permette di ridurre al minimo lo spazio occupato per tenere traccia dello spazio libero, è però impossibile sapere se ci sono blocchi contigui liberi.

Raggruppamento In questo metodo, si usano delle liste concatenate per tenere traccia dei blocchi di dati liberi, ma si raggruppano al momento della rimozione di un file. In questo metodo si memorizza la dimensione della serie di blocchi liberi nel primo blocco libero e il puntatore al blocco libero successivo

nell'ultimo blocco libero. In questo modo si riesce a tenere traccia dello spazio libero in modo più efficiente rispetto al metodo precedente, ed si riesce a sapere se ci sono blocchi contigui liberi.

12.5 Efficienza e Prestazioni

Abbiamo appurato che il disco è il collo di bottiglia del sistema, dunque è necessario ottimizzare le prestazioni del *file system* per migliorare le prestazioni del sistema. L'efficienza di un *file system* dipende dall'algoritmo di allocazione dello spazio sul disco e dal tipo di dati contenuto nelle *directory*. Le prestazioni del disco dipendono dal *controller* fisico del disco, questo ha a disposizione una *cache* per memorizzare i dati letti e scritti, ma ciò non è sufficiente per migliorare le prestazioni, vengono quindi usati dei meccanismi di dischi virtuali e cache del disco per migliorare le prestazioni.

Dischi virtuali I dischi virtuali sono dei dischi che vengono emulati tramite software e vengono memorizzati in memoria principale. Questi dischi sono gestiti interamente dall'utente e possono essere usati per memorizzare i dati in modo più efficiente rispetto ai dischi fisici, ma quando si spegne il computer i dati memorizzati nei dischi virtuali vengono persi.

Cache del disco In questa tecnica viene memorizzata una copia dei blocchi di dati più usati in memoria principale, in modo da ridurre il numero di accessi al disco fisico. La cache del disco è gestita dal `SO` e sfrutta i principi di località spaziale e temporale per memorizzare i blocchi di dati più usati, i dati una volta modificati nella cache vengono scritti in *background* sul disco fisico. sussistono problematiche di dimensione della cache, della politica di sostituzione e della scrittura dei dati.

Recupero dei dati Inoltre ci sono dei problemi di consistenza dei dati e di controllo di questa consistenza.

File system journaling Il *file system journaling* è una tecnica usata per garantire la consistenza dei dati in caso di crash del sistema. In questa tecnica, ogni operazione eseguita sul *file system* viene registrata in un *journal*, che è un file speciale memorizzato su disco. In caso di crash del sistema, il file system può ripristinare lo stato precedente usando il *journal*.

Capitolo 13

RAID

Il `RAID` (*Redundant Array of Independent Disks*) è una tecnologia di archiviazione dei dati che combina più dischi rigidi in un'unica unità logica per migliorare le prestazioni, la capacità e la tolleranza ai guasti. Il `RAID` è stato creato un quanto l'evoluzione tecnologica ha permesso di avere dischi rigidi sempre più piccoli (fisicamente) e sempre meno costosi, inoltre è semplice equipaggiare un sistema con più dischi rigidi, quindi è possibile sfruttare questa tecnologia per migliorare le prestazioni e la sicurezza dei dati. Gli obbiettivi principali del `RAID` sono il miglioramento dell'affidabilità e delle prestazioni.

Struttura dei dispositivi `RAID`

I sistemi `RAID` possono essere implementati in modi differenti: usando più dischi indipendenti collegati ad un bus ed il `SO` gestisce i dischi come un'unica unità logica, oppure usando un controller del disco *hardware* che gestisce i dischi e presenta al `SO` un'unica unità logica, infine con una batteria `RAID` che è una scheda *hardware* indipendente che gestisce i dischi e presenta al `SO` un'unica unità logica.

Concetti base

Le strutture `RAID` si basano sulla copia speculare dei dati e sul sezionamento dei dati. La copia speculare dei dati è una tecnica che prevede la duplicazione dei dati su più dischi per garantire la tolleranza ai guasti. Il sezionamento dei dati è una tecnica che prevede la suddivisione dei dati in blocchi e la distribuzione di questi blocchi su più dischi per migliorare le prestazioni. Il `RAID` combinando questi permette di garantire maggior affidabilità e prestazioni.

Affidabilità Se vengono memorizzati i dati su più dischi, se uno di questi si guasta, i dati possono essere recuperati tramite la ridondanza creata. Quindi l'affidabilità cresce con il numero di dischi. Questo viene garantito tramite la copia speculare dei dati (*mirroring*) dove un disco logico corrisponde a più dischi fisici, in questo caso ogni scrittura viene eseguita su entrambi i dischi ma i dati vengono persi solo se entrambi i dischi si guastano.

Prestazioni Le prestazioni possono essere migliorate tramite il sezionamento dei dati (*striping*) dove i dati vengono suddivisi in blocchi e distribuiti su più dischi. In questo modo le operazioni di lettura e scrittura possono essere eseguite in parallelo su più dischi, migliorando le prestazioni complessive del sistema. Ma non si può garantire la ridondanza dei dati, quindi se un solo disco si guasta, l'intero array di dischi non è più accessibile.

Sezionamento dei dati Il sezionamento dei dati è una tecnica che prevede la suddivisione o a livello di bit o a livello di blocco. Nel `bit-`*by*`-bit` ad esempio con 8 dischi l'i-esimo bit viene memorizzato sul i-esimo disco. Mentre nel `block-`*by*`-block` i dati vengono suddivisi in blocchi e distribuiti su più dischi. Ad esempio con n dischi, il blocco i viene memorizzato sul disco $i \mod n$.

Livelli di RAID

I livelli di RAID sono delle configurazioni standardizzate che definiscono come i dati vengono distribuiti e protetti su più dischi. Ogni livello ha le proprie caratteristiche in termini di prestazioni, capacità e tolleranza ai guasti.

RAID 0

Il RAID 0 è una configurazione di RAID che utilizza il sezionamento dei dati per migliorare le prestazioni. I dati vengono suddivisi in blocchi e distribuiti su più dischi, senza alcuna ridondanza. Questo significa che se un disco si guasta, tutti i dati memorizzati nell'array sono persi. Il RAID 0 offre prestazioni elevate, ma non garantisce la tolleranza ai guasti. Questo livello è molto economico e permette un aumento delle prestazioni, ma non è adatto per applicazioni critiche dove la perdita di dati non è accettabile.

RAID 1

Il RAID 1 è una configurazione di RAID che utilizza la copia speculare dei dati per garantire la tolleranza ai guasti. I dati vengono duplicati su più dischi, quindi se un disco si guasta, i dati possono essere recuperati dall'altro disco. Il RAID 1 offre prestazioni elevate in lettura, ma le prestazioni in scrittura sono inferiori rispetto al RAID 0 a causa della duplicazione dei dati. Questo livello ha un alto costo in termini di capacità, ed una bassa scalabilità poiché la capacità totale dell'array è pari alla capacità del disco più piccolo. Il RAID 1 è adatto per piccole applicazioni critiche dove la perdita di dati non è accettabile.

RAID 2

Il livello RAID 2 è una configurazione di RAID che introduce il concetto di bit di parità per garantire la tolleranza ai guasti. I dati vengono suddivisi in blocchi e distribuiti su più dischi, e vengono usati tre dischi aggiuntivi per memorizzare le informazioni di parità. In questo livello su 7 dischi 4 memorizzano i dati, 3 memorizzano la parità. Questo livello permette la perdita di un disco e la ricostruzione dei dati, ma non è molto efficiente in termini di costo, anche se permette una migliore prestazione rispetto al RAID 0. Inoltre il RAID 2 permette permette di risparmiare solo un disco per la parità rispetto al RAID 1, quindi non è molto usato.

RAID 3

Il livello RAID 3 è una configurazione di RAID che utilizza il sezionamento dei dati **in** byte[1] ed alloca un intero disco per memorizzare i byte di parità. In questo modo quando il controllore legge i dati può immediatamente calcolare la parità e verificare se i dati sono corretti. In questo modo si può garantire la tolleranza al singolo guasto di un disco e/o la corruzione dei dati di un unico byte. Il livello RAID 3 ha la stessa efficienza del RAID 2 in termini di lettura e scrittura, ma è più efficiente in termini di costo, in quanto non abbiamo bisogno di tre dischi per la parità. Tuttavia questo livello è meno efficiente rispetto al RAID 1 in termini di operazioni di I/O ed richiede più tempo per la scrittura dei dati dato che deve essere calcolata la parità. Infine il RAID 3 ha un problema di usura del disco di parità, in quanto questo non verrà usato uniformemente rispetto agli altri dischi, quindi si usurerà prima degli altri.

RAID 4

Il livello RAID 4 è una configurazione di RAID che utilizza il sezionamento dei dati in blocchi e memorizza le informazioni di parità su un disco dedicato, proprio come il RAID 3, ma a differenza di quest'ultimo il sezionamento dei dati è a livello di blocco. Questo significa che i dati vengono suddivisi in blocchi e distribuiti su più dischi, mentre le informazioni di parità vengono memorizzate su un disco dedicato. Questo livello offre gli stessi vantaggi e svantaggi del RAID 3.

RAID 5

Il livello RAID 5 è la configurazione più comune di RAID e combina il sezionamento dei dati in blocchi con la distribuzione delle informazioni di parità su tutti i dischi. In questo modo si ottiene una maggiore

[1] Tutti gli altri usano il sezionamento in blocchi

tolleranza ai guasti e prestazioni migliori rispetto al RAID 3/4 poiché non c'è un disco dedicato per la parità. Il RAID 5 richiede almeno tre dischi e può tollerare la perdita di un disco senza perdita di dati. Tuttavia, le prestazioni in scrittura sono inferiori rispetto al RAID 1 come per il RAID 3/4, poiché è necessario calcolare la parità e scrivere i dati su più dischi. Inoltre,

RAID 6

Il livello RAID 6 è una configurazione di RAID che estende il RAID 5 introducendo un secondo disco di parità. Questo significa che il RAID 6 può tollerare la perdita di due dischi senza perdita di dati. Il RAID 6 diventa però più costoso in termini di capacità e prestazioni rispetto al RAID 5, poiché è necessario un altro intero disco che non aumenta la capacità dell'array e le prestazioni in scrittura sono inferiori a causa del doppio calcolo della parità. Tuttavia, il RAID 6 è adatto per applicazioni critiche dove la perdita di dati non è accettabile e si desidera una maggiore tolleranza ai guasti.

RAID 0+1

Il livello RAID 0+1 è una configurazione di RAID che combina diverse configurazioni RAID 0 e le combina in un unico RAID 1. Ad esempio supponendo di avere a disposizione 6 dischi allora combineremo i primi 3 dischi in un RAID 0 e gli altri 3 in un RAID 0 ed infine combineremo i due RAID 0 in un RAID 1, in questo modo triplichiamo la capacità e la velocità di lettura e scrittura, ed abbiamo una buona tolleranza ai guasti. Tuttavia, se perdiamo due dischi di due RAID 0 diversi, perdiamo tutti i dati, la tolleranza ai guasti è quindi limitata allo stesso gruppo di dischi. Inoltre, il RAID 0+1 ha un costo elevato in termini di capacità.

RAID 1+0

Il livello RAID 1+0 è una configurazione di RAID che combina diverse configurazioni RAID 1 e le combina in un unico RAID 0. Ad esempio supponendo di avere a disposizione 6 dischi allora combineremo i dischi 2 a 2 in modo da ottenere 3 array RAID 1, e poi combineremo i tre array in un RAID 0. In questo modo come per il RAID 0+1 triplichiamo la capacità e la velocità di lettura e scrittura, ed abbiamo una buona tolleranza ai guasti. Tuttavia, se perdiamo entrambi i dischi di un array RAID 1, perdiamo tutti i dati, la tolleranza ai guasti è quindi limitata ad al massimo un disco per array RAID 1. Inoltre, il RAID 1+0 ha un costo molto elevato in termini di capacità.

Capitolo 14

Il sottosistema di I/O

Il compito principale del sottosistema di I/O è quello di fornire un'interfaccia ai processi (utente e sistema) per l'accesso ai dispositivi di I/O tale che questa sia indipendente dal tipo di dispositivo e dalle sue caratteristiche fisiche. Il sottosistema di I/O deve quindi mettere in comunicazione l'*hardware* ed il *software* di I/O in modo che a prescindere dal tipo/modello di dispositivo, il sistema operativo possa fornire un'interfaccia uniforme e standardizzata per l'accesso ai dispositivi.

14.1 Hardware di I/O

Il mercato libero ha portato alla proliferazione di dispositivi di I/O di ogni tipo e forma. Troviamo infatti dispositivi di memoria, di rete, di interazione con l'utente, di visualizzazione, di stampa, ecc. Possiamo inoltre distinguere i dispositivi dai controllori dei dispositivi, mentre i primi costituiscono la parte non elettronica del dispositivo, i secondi sono i circuiti elettronici che si occupano di gestire il dispositivo stesso.
Definiamo ora alcuni termini comuni a quasi tutti i dispositivi di I/O: Infatti ognuno di questi ha una porta che è il punto di connessione fisico tra il dispositivo e il computer. La porta è composta da un insieme di linee elettriche che possono essere utilizzate per inviare o ricevere dati. Le porte possono essere di tipo seriale o parallelo. Poi alcuni dispositivi funzionano su un BUS di sistema, che è un insieme di linee elettriche condivise da più dispositivi, il BUS può essere del tipo *daisy chain* o condiviso. Infine ogni dispositivo ha un suo controllore che è un circuito elettronico che si occupa di gestire il dispositivo stesso e i segnali elettrici che provengono dalla porta. Il controllore è in grado di generare segnali di I/O e di interpretare i segnali provenienti dal dispositivo e dal sistema o da altri controllori.

Controllore dei dispositivi

Come già accennato, il controllore è un circuito elettronico che si occupa di gestire il dispositivo stesso e i segnali elettrici che provengono dalla porta. Questo è connesso al BUS di sistema, ed gli è associato un indirizzo univoco all'interno del sistema. Ogni controllore necessita di un registro/registri di stato, che contengono informazioni sullo stato del dispositivo e del controllore stesso (se ad esempio è pronto a ricevere dati o se ha terminato un'operazione di I/O). Inoltre ogni controllore ha un registro di controllo che viene usato per inviare comandi al dispositivo. Infine ogni controllore ha un *buffer* per passare dai dati nel formato del BUS a quello del dispositivo e viceversa.

Comunicazione tra controllori e sistema

Per mettere in comunicazioni i registri del controllore con il resto del sistema operativo possono essere usate due principali tecniche ed una combinazione delle due. La prima è la *memory-mapped I/O*, mentre la seconda è la *I/O port-mapped*.

Memory-mapped I/O Nella *memory-mapped I/O* i registri del controllore vengono mappati in un'area di memoria del sistema. In questo modo i processi possono inviare e ricevere comandi tramite le istruzione di accesso alla memoria. Ciò permette di scrivere dei *driver* di I/O ad un livello più altro ed inoltre basta allocare la memoria in un'area che và al di fuori dell'area di memoria del sistema operativo per evitare problemi di sicurezza.

Port-mapped `I/O` In questo caso l'accesso ai registri del controllore avviene tramite delle istruzioni specifiche per l'`I/O` che potenzialmente sono diverse per ogni controllore/dispositivo. In questo modo i registri del controllore non sono mappati in memoria il che risparmi un'area di memoria.

Combinazione delle due Un esempio di combinazione ibrida è l'architettura pentium dove i registri del controllore sono mappati tramite *port-mapped `I/O`* e i *buffer* coi dati sono mappati in memoria. In questo modo si risparmia memoria delle istruzioni di `I/O` e si possono usare le istruzioni di accesso alla memoria per accedere ai *buffer*.

Accesso ai dispositivi

L'accesso ai dispositivi può avvenire in tre modi:

- *Polling*
- *Interrupt*
- `DMA` *Direct Memory Access*

Polling Il *polling* è una tecnica di accesso ai dispositivi in cui il sistema operativo controlla periodicamente se il *busy-bit* del registro di stato del controllore è attivo. Se il *busy-bit* è impostato a `0` un nuovo comando può essere scritto nel registro di controllo del controllore e il *command-ready-bit* viene impostato a `1`, viene quindi eseguita l'operazione di `I/O` e quando questa è terminata il *busy-bit* viene impostato a `0`. Il *polling* è una tecnica semplice da implementare ma causa un ciclo di attesa attiva che consuma risorse di sistema.

Interrupt L'*interrupt* è una tecnica di accesso ai dispositivi in cui il controllore invia un segnale al sistema operativo tramite una connessione fisica alla `CPU`. Quando il sistema operativo riceve il segnale di `I/O` interrompe l'esecuzione del processo corrente e salva il suo stato. Il sistema operativo esegue quindi il codice di gestione dell'*interrupt* che si occupa di gestire l'operazione di `I/O` e ripristina lo stato del processo interrotto. Vanno però gestite delle situazioni dove il segnale di *interrupt* arriva ma non può essere gestito immediatamente, ad esempio se si sta eseguendo un'operazione critica o se il sistema operativo è in uno stato di *deadlock*. In questo caso il segnale di *interrupt* deve essere messo in coda e gestito successivamente. Inoltre i segnali di *interrupt* devono essere numerati in modo che la `CPU` esegua le corrette istruzioni di gestione dell'*interrupt*. Infine *interrupt* multipli devono essere gestiti tramite un sistema di priorità, in modo che i segnali più importanti vengano gestiti prima di quelli meno importanti.

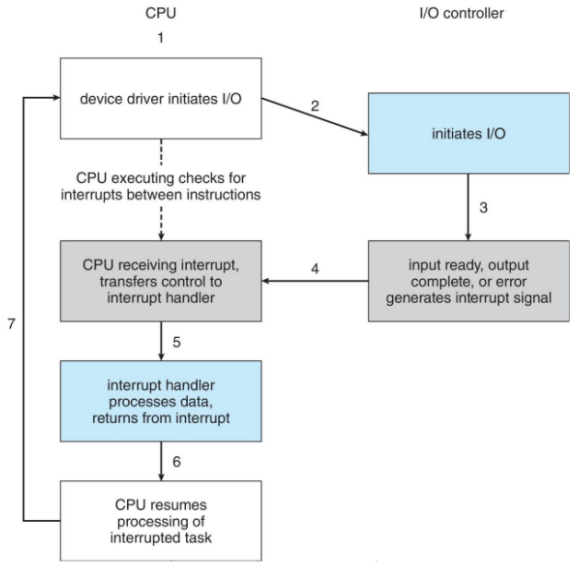


Figura 14.1: Ciclo della `CPU` per la gestione di *interrupt*

DMA - *Direct Memory Access* Il DMA è una tecnica di accesso ai dispositivi ideata per grandi spostamenti di dati. Visto che usare la CPU per controllare solo pochi byte dal registro di stato del controllore è uno spreco di risorse, il DMA permette, tramite un *hardware* dedicato, di trasferire i dati dalla memoria al dispositivo e viceversa senza l'intervento della CPU. Quando la CPU deve trasferire un grande blocco di dati, essa invia un comando al DMA contenete l'indirizzo di partenza dei blocchi di memoria da trasferire, l'indirizzo di destinazione, il numero di byte da trasferire e la direzione del trasferimento. Il DMA si occupa di gestire il trasferimento comunicando direttamente con il controller del dispositivo e con la memoria. Quando il trasferimento è terminato il DMA invia un segnale di *interrupt* alla CPU per informarla che il trasferimento è terminato. Il DMA può essere usato in combinazione con il *polling* o con gli *interrupt*. Infatti il DMA può essere usato per trasferire i dati tra la memoria e il dispositivo, mentre la CPU può essere usata per gestire gli *interrupt* e il *polling*.

14.2 Interfacce di I/O

Come già accennato, i dispositivi di I/O sono molto diversi tra loro e si differenziano per modalità di funzionamento, modalità di accesso, velocità di trasferimento, ecc. Per questo motivo è necessario usare un *layer* aggiuntivo di software che si occupa di nascondere le differenze tra i vari dispositivi al SO e ai processi.
Questo *layer* deve fornire una interfaccia comune ed uniforme per l'accesso ai dispositivi di I/O e deve essere in grado di gestire le differenze tra i vari dispositivi. Ecco quindi che introduciamo i *driver*
Una prima grande distinzione è tra *driver* per dispositivi a blocco e *driver* per dispositivi a caratteri. I primi sono dispositivi nei quali è possibile memorizzare e trasferire i dati in blocchi di dimensione fissa, in questo genere di dispositivi è possibile leggere (*read*), scrivere (*write*) e cercare (*seek*) i dati, questo è il caso di dischi rigidi, ecc... Questi dispositivi solitamente sono *memory-mapped* e usano dei *block-device drivers*.
I *driver* per dispositivi a caratteri sono invece dispositivi nei quali i dati vengono trasferiti in modo sequenziale e non è possibile accedere ai dati in modo casuale. Quindi al posto di avere i comandi *read*, *write* e *seek* abbiamo i comandi *get* e *put*. Questi dispositivi sono solitamente *port-mapped* e usano dei *character-device drivers*. Esempio di dispositivi a caratteri sono le stampanti, le porte seriali, i terminali, ecc...

14.3 Software di I/O

La parte del SO che si occupa di gestire i dispositivi di I/O è chiamata *I/O subsystem*. Questa parte deve essere indipendente dal dispositivo, la notazione deve essere uniforme e standardizzata, deve essere in grado di gestire gli errori e le varie opzioni di trasferimento dei dati. Il software è organizzato in quattro *layer* principali:

1. Gestione degli *interrupt*
2. *Device drivers*
3. SW del SO indipendente dal dispositivo
4. Programmi utente

Gestione degli *interrupt* La gestione degli *interrupt* è la parte del SO che si occupa di gestire gli *interrupt* provenienti dai dispositivi di I/O. Gestisce inoltre le iterazioni coi controller dei dispositivi e la comunicazione con il SW del SO indipendente dal dispositivo.

Device drivers I *device drivers* sono i programmi che si occupano di gestire i dispositivi di I/O. Questi programmi sono scritti dai produttori dei dispositivi e sono specifici per ogni dispositivo, prima di essere distribuiti questi sono firmati digitalmente dai produttori dei sistemi operativi. I *device drivers* contengono tutto ciò che è *device-dependent* e si occupano di gestire le differenze tra i vari dispositivi. I *device drivers* sono spesso condivisi tra più dispositivi simili e sono tipicamente scritti in linguaggio macchina o in linguaggio C.

SW del SO indipendente dal dispositivo Questa parte del SO si occupa di gestire le operazioni di I/O in modo uniforme e standardizzato. Inoltre deve gestire ed individuare i dispositivi collegati e gestirne i nomi. Inoltre deve proteggere tutte le operazioni di I/O dato che tutte le primitive di I/O sono privilegiate, và anche gestito il *buffering* dei dati e la sincronizzazione tra i vari processi che accedono ai dispositivi di I/O, questo per gestire le differenti velocità, dimensioni e dimensioni di blocco dei vari dispositivi. Infine deve gestire gli errori, anche quelli *device-dependent*, e l'allocazione ed il rilascio dei dispositivi.

SpoolingLo *spooling* è una tecnica di I/O che permette di gestire i dispositivi di I/O in modo efficiente. Questa tecnica consiste nel memorizzare i dati in un'area di memoria temporanea (chiamata *spool*) prima di inviarli al dispositivo. In questo modo solo il processo *spooler* deve gestire il dispositivo e gli altri processi possono continuare a lavorare senza dover aspettare il completamento dell'operazione di I/O. Inoltre lo *spooling* permette di gestire i dispositivi in modo asincrono, ovvero i processi possono continuare a lavorare mentre i dati vengono trasferiti al dispositivo. Questa tecnica è molto utile per i dispositivi di I/O lenti come le stampanti o i dischi rigidi.

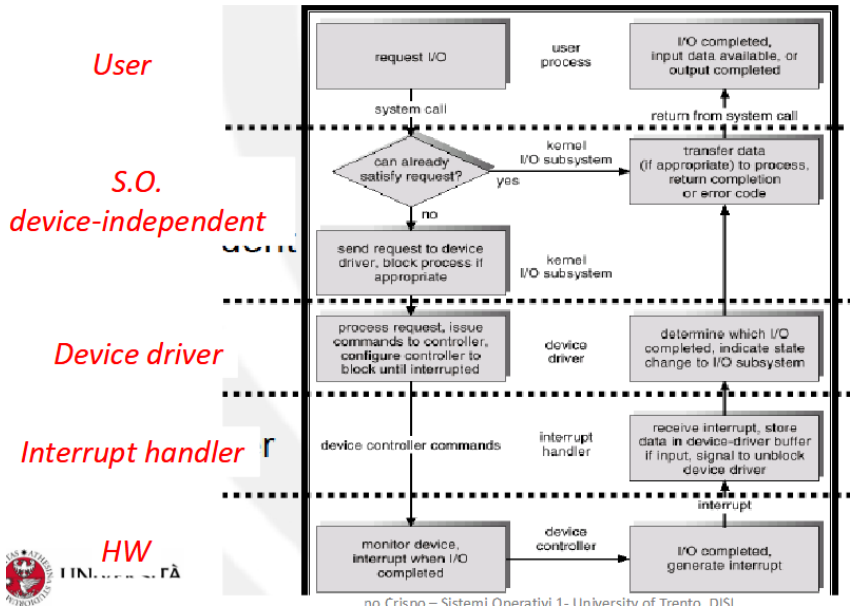


Figura 14.2: Ciclo della CPU per la gestione di I/O