

Sistemi Operativi

Università degli Studi di Padova

ALESSANDRO TRIGOLO

2022—2023

1

*

INDICE

PAGINA 1

CAPITOLO 1

*

CAPITOLO 2

PROCESSI

PAGINA 2

2.1	Allocazione in memoria	2
2.1.1	Process Control Block (PCB)	3
2.2	Stati di un processo	4
2.2.1	Context switch.....	5
2.2.2	Creazione di un processo.....	6
2.2.3	L'albero dei processi in Linux	7
2.2.4	Terminazione di un processo	7
2.3	Comunicazione tra processi (IPC).....	8
2.3.1	Memoria condivisa	8
2.3.2	Passaggio di messaggi	9

CAPITOLO 3 **THREADS** **PAGINA 11**

3.1	Concorrenza e parallelismo	11
3.1.1	Tipi di parallelismo	12
3.1.2	Legge di <i>Amdahl</i>	13
3.2	Modelli multithreading	14
3.2.1	Many-to-One	14
3.2.2	One-to-One	15
3.2.3	Many-to-Many	15
3.3	Librerie di thread	16
3.4	Threading implicito	17
3.4.1	Modello fork-join	17
3.4.2	Thread pools e OpenMP	18
3.5	Problematiche	18
3.5.1	Semantica <code>exec</code> e <code>fork</code>	18
3.5.2	Segnalazione ed eliminazione	19

CAPITOLO 4 **CPU SCHEDULING** **PAGINA 20**

4.1	Nozioni fondamentali	20
4.2	Algoritmi non preemptive	22
4.2.1	First-Come First-Served (FCFS)	22
4.2.2	Shortest-Job-First (SJF)	23
4.2.3	Stima del <i>CPU burst time</i>	24
4.3	Algoritmi preemptive	25
4.3.1	Shortest-Remaining-Time-First (SRTF)	25
4.3.2	Round Robin (RR)	26
4.3.3	Reattività	27
4.4	Scheduling con priorità	28
4.4.1	Scheduling con priorità e RR	29

4.4.2	Coda multilivello	30
4.5	Scheduling multiprocessore	30
4.5.1	Symmetric multiprocessing (SMP)	31
4.5.2	31
4.6	Scheduling real-time	31
4.7	Valutazione di un algoritmo	31

CAPITOLO 5	SINCRONIZZAZIONE	PAGINA 32
5.1	Sezione critica	32
5.1.1	Requisiti.....	33
5.1.2	Soluzioni inefficienti	33
5.2	Soluzione di Peterson	34
5.2.1	Architetture moderne	35
5.3	Sincronizzazione via hardware	36
5.3.1	Test and set.....	36
5.3.2	Compare and swap	37
5.3.3	Variabili atomiche	38
5.4	Mutex lock e Semafori.....	38
5.4.1	Waiting queue	40
5.5	Monitor	41
5.5.1	Struttura e implementazione	42
5.5.2	Variabile di condizione	43
5.6	Problemi comuni della sincronizzazione	44
5.6.1	Buffer limitato	44
5.6.2	Problema dei lettori e degli scrittori	45
5.6.3	Problema dei 5 filosofi	46

CAPITOLO 6 DEADLOCKS **PAGINA 50**

6.1	Nozioni fondamentali	50
6.1.1	Caratterizzazione	51
6.1.2	Grafo risorsa-allocazione	52
6.2	Avoidance	55
6.2.1	Safe state	55
6.2.2	Algoritmo sul grafo risorsa-allocazione	56
6.2.3	Algoritmo del banchiere	57
6.3	Detection	60
6.3.1	Istanza singola	60
6.3.2	Istanze multiple	61
6.3.3	Deadlock recovery	62

CAPITOLO 7 MEMORIA PRINCIPALE **PAGINA 64**

7.1	Introduzione	64
7.1.1	Protezione	65
7.1.2	Binding	66
7.1.3	Memory-Management Unit (MMU)	67
7.1.4	Caricamento e collegamento dinamico	68
7.2	Primi modelli di allocazione	69
7.2.1	Allocazione contigua	69
7.2.2	Allocazione a partizione fissa	69
7.2.3	Allocazione a partizione variabile	70
7.2.4	Problema della frammentazione	72
7.3	Paginazione (<i>paging</i>)	73
7.3.1	Frammentazione interna	74
7.3.2	Traduzione degli indirizzi	75
7.3.3	Allocazione nei frames liberi	77

7.4	Page table	78
7.4.1	Translation Look-aside Buffer (TLB)	79
7.4.2	Bit di validità	81
7.4.3	Page table gerarchica	82
7.4.4	Page table con tabella <i>hash</i>	83
7.4.5	Page table invertita	84
7.5	Swapping	86
7.5.1	Swapping con paginazione	87
7.5.2	Swapping nei dispositivi mobili	87
7.6	Segmentazione (<i>segmentation</i>)	88
7.6.1	Segment table	89
7.6.2	Modello ibrido	90

CAPITOLO 8 MEMORIA VIRTUALE PAGINA 91

8.1	Introduzione	91
8.1.1	Spazio degli indirizzi virtuali	92
8.1.2	Memoria condivisa	92
8.2	Demand Paging	93
8.2.1	Page fault	94
8.2.2	Pure demand paging	95
8.2.3	Performance e ottimizzazione	96
8.2.4	Prepaging	97
8.3	Page replacement	98
8.3.1	FIFO	99
8.3.2	Algoritmo ottimale	100
8.3.3	LRU	101
8.3.4	Second-chance (clock)	102
8.3.5	Algoritmo counting	104

8.3.6	Ottimizzazione	104
8.4	Allocazione dei frames	105
8.4.1	Allocazione globale e locale.....	105
8.4.2	Richiesta delle pagine	105
8.5	Thrashing	106
8.5.1	Modello Working-set.....	107
8.5.2	Frequenza dei page fault (PFF).....	108
8.6	Allocare la memoria del kernel.....	109
8.6.1	Buddy system.....	109
8.6.2	Slab allocation	110

CAPITOLO 9 MEMORIA DI MASSA **PAGINA 112**

9.1	Tipi di memoria secondaria.....	112
9.1.1	HDD	112
9.1.2	NVM	114
9.1.3	Memoria volatile.....	115
9.1.4	Nastro magnetico	115
9.1.5	Dispositivi di memorizzazione esterna	116
9.2	Indirizzamento	116
9.3	HDD scheduling	116
9.3.1	FCFS	117
9.3.2	SSTF	117
9.3.3	SCAN e C-SCAN	118
9.3.4	Scelta dell'algoritmo	119

CAPITOLO 10 SISTEMA INPUT/OUTPUT **PAGINA 120**

10.1	Componenti hardware	120
------	---------------------------	-----

10.2	Tecniche di comunicazione.....	121
10.2.1	Polling	122
10.2.2	Interrupt	122
10.2.3	DMA.....	123
10.3	Gestione software.....	125
10.3.1	Caratteristiche delle periferiche I/O.....	125
10.3.2	Tipi di device-drivers.....	125
10.3.3	Device sincroni e asincroni	126
10.4	Task del kernel.....	127
10.4.1	Gestione degli errori	128
10.4.2	Strutture dati	128

CAPITOLO 11 | INTERFACCIA DEL FILE SYSTEM **PAGINA 130**

11.1	Concetto di file	130
11.1.1	Attributi e operazioni	131
11.1.2	Files aperti e file <i>locking</i>	131
11.1.3	Struttura del file	132
11.2	Metodi di accesso	132
11.2.1	Accesso sequenziale	132
11.2.2	Accesso diretto	133
11.3	Struttura della directory	133
11.3.1	Directory a uno e due livelli.....	134
11.3.2	Directory strutturata ad albero.....	134
11.3.3	Problema dei cicli.....	136
11.3.4	Disco.....	137
11.4	Protezione	137
11.4.1	Access list in Unix/Linux	137

12.1	Struttura e operazioni del file system	138
12.1.1	Livelli	139
12.1.2	Strutture per le operazioni	140
12.1.3	File Control Block (FCB).....	140
12.2	Metodi di allocazione	141
12.2.1	Allocazione contigua.....	141
12.2.2	Allocazione concatenata	142
12.2.3	Allocazione indicizzata.....	144
12.2.4	Performance.....	147
12.3	Gestione dello spazio libero	147
12.3.1	Free list	147
12.3.2	Counting e grouping	148
12.3.3	Altri metodi	148

2 PROCESSI

Iniziamo dalle basi. Un **processo** è un programma in esecuzione, è un’istanza del programma che viene eseguita sulla CPU. Possiamo infatti avere diverse istanze dello stesso programma, ognuna che viene eseguita indipendentemente dall’altra. Possiamo quindi dire che il programma, ovvero il file eseguibile (`.exe`) è qualcosa di passivo mentre il processo è qualcosa di **attivo**.

2.1 Allocazione in memoria

Andando un po’ più in dettaglio, quando il programma è in esecuzione, questo viene eseguito in maniera sequenziale. Al processo, una volta che è eseguito, viene dedicato dello spazio in memoria dal sistema operativo. Come è possibile osservare nella figura 2.1, la memoria messa a disposizione dal sistema operativo è suddivisa in diverse zone, ciascuna con un particolare compito. Prima di tutto, il codice sorgente del programma viene caricato nella zona **text**. Dopo di che, nella parte dedicata ai dati (**data**) vengono salvate generalmente le variabili globali, che permangono per tutta la vita del processo. Sono infine presenti due parti: lo **stack** e l'**heap** che crescono in direzione opposta. Lo stack contiene dati temporanei come variabili locali mentre l’heap è utilizzato al fine di allocare la memoria dinamicamente durante la vita del programma¹.

¹Come abbiamo visto con C++, heap e *freestore* sono quasi dei sinonimi.

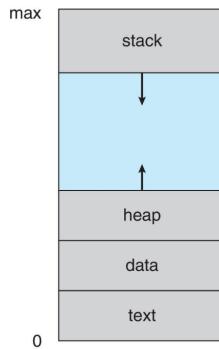


Figura 2.1: Spazio in memoria allocato per il processo dal sistema operativo.

2.1.1 Process Control Block (PCB)

Ad ogni processo che è mandato in esecuzione è assegnata una particolare struttura dati dal sistema operativo, ovvero il *Process Control Block* (figura 2.2). Il PCB contiene diverse informazioni riguardanti il processo, in particolare:

1. Lo **stato** del processo;
2. Informazioni sul **program counter**, in particolare è importante sapere se il processo è fermato temporaneamente e poi fatto ripartire più tardi;
3. Valori dei registri, utili nel caso in cui un processo venga messo in pausa;
4. Altre informazione riguardanti lo scheduling della CPU (vedi capitolo 4), come la priorità del processo;
5. Informazioni per la gestione della memoria e dell'I/O.

In particolare, in Linux, nel PCB di un processo (che in Linux è chiamato `task`) sono presenti le seguenti informazioni: `pid` (numero assegnato al particolare processo), puntatori al processo genitore (che vedremo saranno utili nella fase di creazione di un processo), puntatori ai processi figli e altre informazioni come la lista dei file aperti. Quando un nuovo processo è creato in Linux, le sue informazioni sono detenute in una lista concatenata (*doubly-linked list*) dove ogni nodo della lista è il PCB di un processo specifico (figura 2.3). Al fine di andare a modificare delle informazioni del processo (come lo stato corrente) il sistema operativo scorre la lista e, dopo aver selezionato il PCB del processo desiderato, andrà a modificare il campo.



Figura 2.2: Rappresentazione del contenuto di un generico PCB.

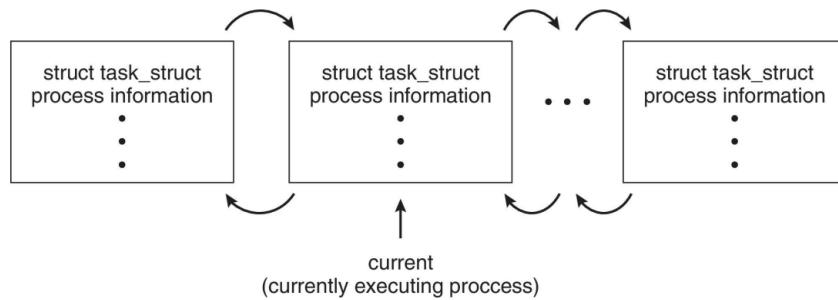


Figura 2.3: Lista concatenata che mantiene tutti i PCB dei processi (task) in Linux.

2.2 Stati di un processo

Durante l'intera vita del processo, questo passa in diversi stati (figura 2.4). I principali sono:

1. **New**: il processo è appena stato creato;
2. **Ready**: il processo è pronto per essere eseguito, quindi non è ancora stato assegnato al processore e sta aspettando l'assegnazione;
3. **Running**: dopo essere stato associato alla CPU il processo inizia ad essere eseguito. Come vedremo nel capitolo 4 e successivi, il processo può essere interrotto e quindi ritorna allo stato ready;
4. **Waiting**: se il processo deve aspettare qualche input da esterno si mette in attesa e una volta che riceve l'input ritorna nello stato ready;

5. **Terminated**: una volta che il processo finisce la sua esecuzione, questo ovviamente termina e viene rimosso dalla CPU.

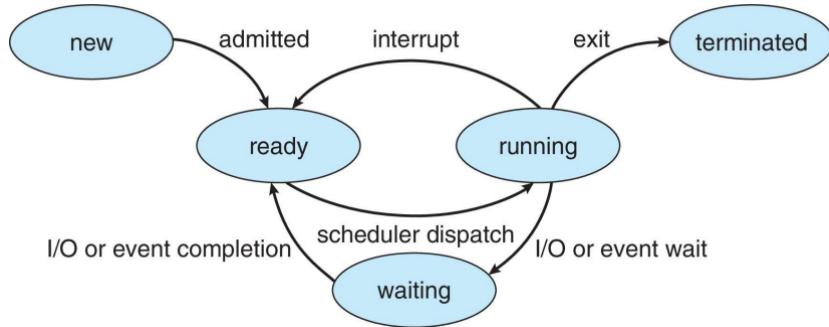


Figura 2.4: Gli stati della vita di un processo.

Lo stato *ready* e lo stato *wait* contengono delle code dove i processi attendono di essere eseguiti ovvero la **ready queue** e la **wait queue** che non sono altro che delle liste concatenate. Le liste contengono i PCB dei processi: il sistema operativo tiene traccia del primo e dell'ultimo processo nella lista al fine di riuscire a implementare le due code. Possiamo inoltre suddividere i processi in due macro categorie:

- ◊ **CPU bound** che sono i processi che hanno un uso massiccio della CPU;
- ◊ **I/O bound**, ovvero processi che spendono la maggior parte del loro tempo in una situazione di wait per leggere o scrivere sulle periferiche.

2.2.1 Context switch

Quando un processo A viene rimosso dalla CPU, nel caso in cui sia stato interrotto per far spazio ad un altro processo B, è necessario salvare l'informazione del processo A in modo tale da poterlo sostituire con il processo B per poi, in un secondo momento, riuscire a ricaricare il processo A. Questa operazione è detta **context switch** (figura 2.5) e viene effettuata in pochi microsecondi. Ciò nonostante se è effettuata in maniera molto frequente durante l'esecuzione di diversi processi può causare molto spreco di tempo: il context switch può quindi generare un **overhead** che va sicuramente preso in considerazione negli algoritmi di scheduling (capitolo 4). È importante notare che il context switch tipicamente richiede anche un aggiuntivo utilizzo della memoria, che andremo a discutere nel capitolo 7 quando discuteremo di *paginazione* e *swapping*.

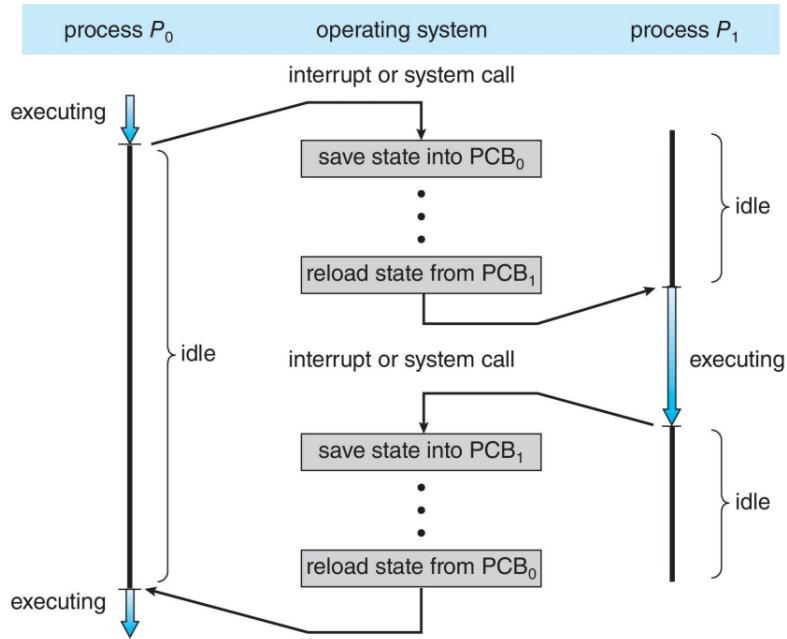


Figura 2.5: Il context switch.

2.2.2 Creazione di un processo

Al fine di creare un processo ce ne deve sempre essere uno iniziale (*parent*) che genera il quello nuovo (*child*). Ogni nuovo processo ha un identificativo, il **pid**, che distingue univocamente il processo creato. Al momento della creazione è possibile specificare alcune opzioni al fine di creare il processo child in un determinato modo. Prima tra tutte è l'opzione di **condivisione di risorse**, dove si può specificare se il figlio condivide le stesse risorse del genitore, un sottoinsieme oppure si può specificare che il figlio non condivide alcuna risorsa con il *parent*. Inoltre si possono specificare le opzioni di **esecuzione**: si specifica se il figlio e il genitore possano essere eseguiti in maniera concorrente oppure se il *parent* deve aspettare il termine dell'esecuzione del *child*. Infine si può anche specificare lo **spazio degli indirizzi**, in particolare si sceglie se il figlio crea una copia identica della memoria utilizzata dal genitore oppure se carica un programma completamente nuovo.

Vediamo ora un esempio di creazione di un processo in **UNIX** (figura 2.6). Questo sistema operativo fornisce tre particolari *system calls*:

- ◊ `fork()`: questa system call non fa altro che creare un processo. Il processo parent, dopo aver

chiamato la funzione `fork()` viene duplicato. La funzione inoltre ritorna un valore intero, se questo valore è maggiore di zero vuol dire che ci troviamo all'interno del processo genitore; se invece il valore è zero vuol dire che il codice eseguito è all'interno del child. L'unico modo per distinguere se il processo è parent o child è attraverso il valore di ritorno di `fork`.

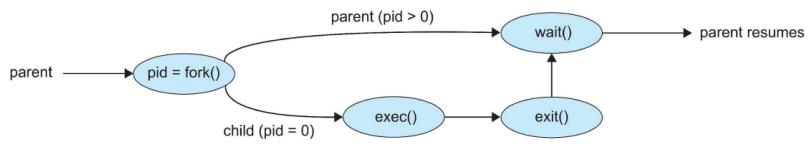


Figura 2.6: Rappresentazione delle 3 *system calls* fondamentali.

- ◊ `exec()`: è una funzione utilizzata dal processo figlio nel caso in cui è necessario far partire un processo completamente diverso dal parent;
- ◊ `wait()`: è una system call utilizzata dal genitore al fine di aspettare il termine dell'esecuzione del figlio.

2.2.3 L'albero dei processi in Linux

Come fa il sistema operativo a generare tutti i processi di cui ha bisogno? Ci deve sempre essere un processo iniziale, un programma all'inizio da cui tutti si genera. In particolare in Linux il primo processo da cui tutto è generato è chiamato `systemd` ed è il processo padre di tutti gli altri processi, quello il quale `pid` vale 1. Da `systemd`, *forkando* processo dopo processo vengono generati tutti i processi necessari all'avvio del sistema, come il terminale, generando quindi un albero (figura 2.7).

2.2.4 Terminazione di un processo

Naturalmente, un processo può anche terminare. La terminazione del processo può essere spontanea (attraverso la *system call* `exit`) e quindi il processo viene deallocated dal sistema operativo, oppure il processo può essere terminato dal genitore attraverso la *system call* `abort`. Questo di solito avviene quando il processo figlio supera il limite delle risorse allocate, quando la task che sta completando non è più richiesta oppure nel momento in cui il genitore termina e di conseguenza il sistema operativo termina anche i figli.

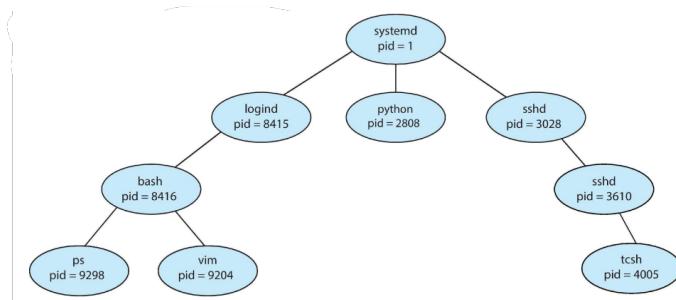


Figura 2.7: L'albero dei processi generato da `systemd`.

Come abbiamo visto in precedenza, esiste una funzione (`wait()`), che serve per evitare che il processo parent termini prima del processo child: la funzione infatti obbliga il parent ad aspettare che il child termini. Inoltre, se al termine di un processo child non c'è nessun processo parent che stava aspettando il termine del child, ci troviamo davanti ad un **processo zombie**. Infine, se il processo parent termina senza aspettare la terminazione del child, quest'ultimo è chiamato **orfano** e verrà terminato dal sistema operativo.

2.3 Comunicazione tra processi (IPC)

Passiamo ora a discutere i diversi modi per comunicare tra diversi processi. In alcuni casi avremo a che fare con processi indipendenti, altre volte invece necessiteremo di processi **cooperanti**. Per questi ultimi è necessario un modello di *Inter-Process Communication*, chiamata anche **IPC**. In questo paragrafo ci occuperemo di due modelli: comunicazione tramite memoria condivisa e tramite il passaggio di messaggi.

2.3.1 Memoria condivisa

Il primo modello di cui ci occuperemo è la tecnica di memoria condivisa. In questo modello, come possiamo notare anche dalla figura 2.8, l'unica cosa di cui si fa carico il sistema operativo è l'assegnazione di una memoria che è condivisa tra i processi A e B. Ciò significa che la comunicazione è molto veloce tra i due processi in quanto non c'è nessun intermediario tra i due. Allo stesso tempo però è più facile che si generino errori come la sovrascrittura di valori e in genere problemi di

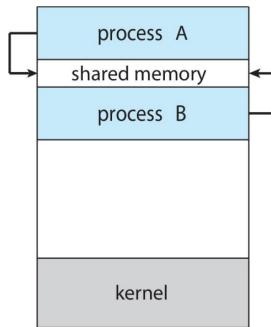


Figura 2.8: Il modello di memoria condivisa per IPC.

sincronizzazione con i dati (come la *race condition*, vedi capitolo 5) in quanto il sistema operativo non ha alcun tipo di controllo sulla memoria secondaria.

Partiamo ora da un esempio di questo modello al fine di ottenere informazioni più dettagliate: stiamo infatti parlando del problema Producer – Consumer. Ipotizziamo quindi che due processi abbiano dello spazio in memoria condiviso; la comunicazione attraverso questa memoria può avvenire in due modi:

- ◊ **unbounded**, ovvero che il produttore continua a generare dati da mettere nell'area condivisa, e che il consumatore continua ad utilizzare quei dati fino a che non finiscono (al più attende la creazione di altri dati).
- ◊ **bounded**, dove si ha un **buffer** che entrambi devono aspettare: il consumer attende che ci siano dati nel buffer e il producer aspetta nel momento in cui il buffer è pieno.

2.3.2 Passaggio di messaggi

In questo secondo caso invece il sistema operativo si prende carico di gestire la coda dei messaggi (**message queue**) che vengono scambiati tra i due processi (figura 2.9). In questo caso è il kernel che fa da intermediario tra i due e di conseguenza la velocità di comunicazione sarà ridotta dall'**overhead**. Allo stesso tempo però il kernel garantisce la sincronizzazione e la correttezza tra i messaggi scambiati e di conseguenza è un modello più sicuro.

Discutiamo ora più dettagliatamente questo modello. In particolare, sono fornite due operazioni fondamentali: `send(message)` e `recv(message)`. Ciò nonostante la comunicazione di tali messaggi fa sorgere diversi dubbi e domande legate alla progettazione: come sono stabilite le

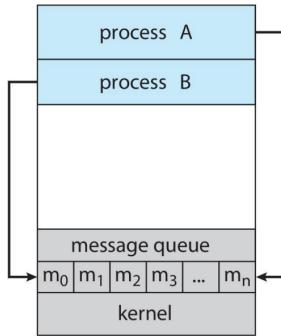


Figura 2.9: Il modello di memoria condivisa per IPC.

connessioni? qual è la sua capacità? è unidirezionale o bidirezionale? la dimensione del messaggio è fissa o variabile? un collegamento è associato solo a 2 processi o a più? Per rispondere a queste domande andiamo a vedere l'implementazione di questo modello che può essere di due tipi.

Quando si parla di comunicazione **diretta**, si intende che i processi specifichino esplicitamente il destinatario del messaggio: `send(Q, message)` e `receive(P, message)`. Nella comunicazione **indiretta** invece si ha a che fare con un **buffer**, chiamato anche porta, il quale ha un ID unico tra gli altri. In questa implementazione due processi si possono parlare solo se condividono questo buffer (**mailbox**). Se più processi condividono la stessa mailbox si ha quindi un link che collega diversi processi, inoltre attraverso questa porta il collegamento è bidirezionale in quanto un processo può sia spedire un messaggio che riceverlo. Con la comunicazione indiretta le primitive però sono diverse: una volta creata la porta (chiamiamola A, per comodità), al fine di scambiare i messaggi è necessario utilizzare le operazioni `send(A, message)` e `receive(A, message)`.

3 THREADS

Partiamo con il distinguere un thread da un processo. Il thread è un **filo di esecuzione**: in un processo ci possono essere diversi thread i quali possono avere dei diversi *pattern* per ciascuna esecuzione.

Perché?

I thread sono entità più semplici rispetto ai processi e sono quindi più facili da gestire. Grazie ai thread si ottengono diversi vantaggi:

- ◊ Il sistema è più **recettivo**;
- ◊ Dato che le risorse sono condivise è più facile gestirle;
- ◊ Richiede meno risorse rispetto alla creazione di un processo;
- ◊ Può utilizzare tutti i *core* messi a disposizione dal sistema (*multicore programming*).

Per esempio, al posto di far eseguire 4 processi differenti è molto meglio eseguire un processo con 4 thread differenti: in questo modo si evita di allocare in memoria 4 volte le stesse risorse che, grazie all'utilizzo dei thread, sono allocate solo una volta.

3.1 Concorrenza e parallelismo

Quando parliamo di **multicore programming** è necessario fare una netta distinzione tra il significato di concorrenza e parallelismo. Con **parallelismo** si intende che un sistema è in grado di preformare

più di un compito in maniera simultanea (tipico dei sistemi multicore). Con **concorrenza** si intende la possibilità di far progredire più di un compito (non in maniera simultanea).

Come possiamo vedere della figura 3.1, quando parliamo di concorrenza ci riferiamo ad un singolo core che esegue a frammenti più thread diversi.

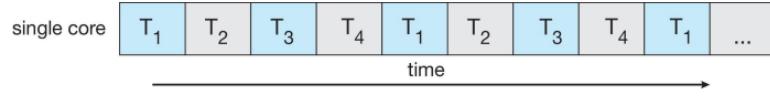


Figura 3.1: Esempio di concorrenza tra 4 processi.

Quando invece facciamo riferimento al parallelismo (figura 3.2) indichiamo la capacità del sistema di effettivamente riuscire ad eseguire parallelamente diversi thread. Si osserva che le due pratiche non

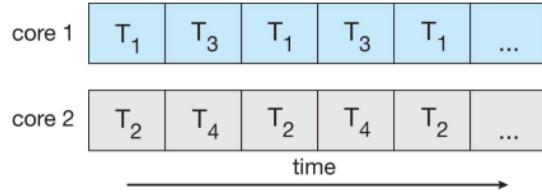


Figura 3.2: Esempio di parallelismo in un sistema dual core.

sono esclusive: notiamo che il core 1 esegue in maniera concorrente T₁ e T₃, mentre il core 2 esegue in maniera concorrente T₂ e T₄.

3.1.1 Tipi di parallelismo

Possiamo dividere i parallelismi in due tipi. Il primo, chiamato **data parallelism** (figura 3.3) implica un sottoinsieme preciso di dati sia distribuito per ogni core. In altre parole, avendo a che fare un un largo database, lo si suddivide in parti e ciascuna viene assegnata ad un core. Tale core potrà operare solo in quella porzione di dati.

Il secondo tipo di parallelismo è detto **task parallelism**, rappresentato in figura 3.4. Questo tipo di parallelismo concede la memoria condivisa a ciascun core solo che ogni core ha un compito ben preciso: un core sarà ottimizzato per la scrittura, un altro sarà più veloce in lettura e così via.

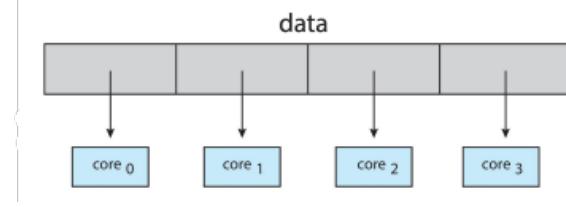


Figura 3.3: Esempio di parallelismo di dati.

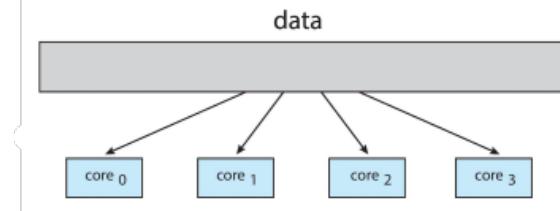


Figura 3.4: Esempio di parallelismo di compiti.

3.1.2 Legge di *Amdahl*

La legge di *Amdahl* è una funzione che mette in relazione due variabili importanti:

1. **S**, ovvero la percentuale di codice che non può essere parallelizzato, ovvero codice **seriale** (di conseguenza il numero di codice che può essere parallelizzato è $1 - S$);
2. **N**, che rappresenta il numero di core disponibili.

Con questi due dati abbiamo la possibilità di calcolare lo **speedup** attraverso la seguente formula:

$$\text{speedup} \leq \frac{1}{S + \frac{1-S}{N}}$$

Osservando la formula osserviamo che se se la percentuale di codice seriale tende a zero e il numero di core tende a infinito, lo *speedup* sarebbe infinito. Questa però è una situazione utopica: non esistono casi in cui si è privi di codice seriale.

Osserviamo quindi il seguente grafico (figura 3.5) che mostra lo *speedup* all'aumentare del numero di core, essendo a conoscenza della percentuale di codice seriale. Notiamo che quando la percentuale di codice seriale è circa la metà non ha importanza il numero di core del sistema: lo speedup rimarrà pressoché invariato. Anche solo con il 5% di codice seriale notiamo un forte abbassamento rispetto allo speedup ideale. È evidente quindi che l'aumento di core non causa l'aumento di speedup, soprattutto in una situazione dove il codice seriale è molto.

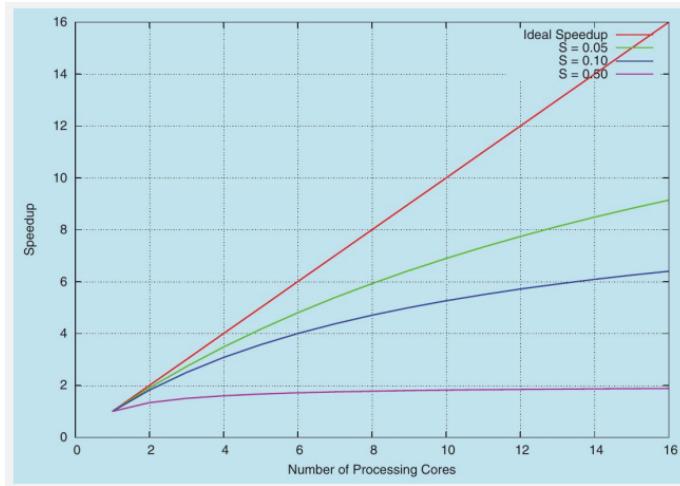


Figura 3.5: Il grafico che descrive lo *speedup* a seconda della percentuale di codice seriale.

3.2 Modelli multithreading

È importante fare una distinzione tra due classi principali di thread: **user threads** e **kernel threads**. La principale differenza tra i due è che i kernel threads hanno molti più privilegi rispetto a quelli utente. Esistono quindi dei modelli che cercano di associare agli user threads i kernel threads al fine di sfruttare al meglio il principio di *multithreading*.

3.2.1 Many-to-One

In questa prima architettura, il kernel mette a disposizione solo un thread che è collegato e deve soddisfare tutte le richieste di tutti gli user threads (figura 3.6). È evidente che l'efficienza di questo modello non è il suo forte: sono presenti diversi thread utente ai quali deve rispondere solamente un thread del kernel. Basta pesare al fatto che se il kernel thread è in attesa di un input esterno, tutti gli user thread sono bloccati (**collo di bottiglia**). Questo modello è infatti poco utilizzato dato che non sfrutta le potenzialità del *multicore*.

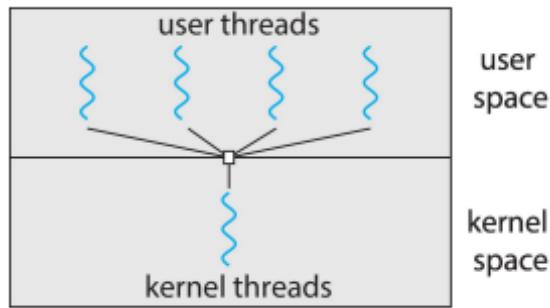


Figura 3.6: Il modello di *multithreading* molti a uno.

3.2.2 One-to-One

In questa architettura (figura 3.7), quando viene creato uno user thread, il suo rispettivo kernel thread viene creato; così facendo esiste un kernel thread associato ad ogni user thread. Ad ogni modo ci

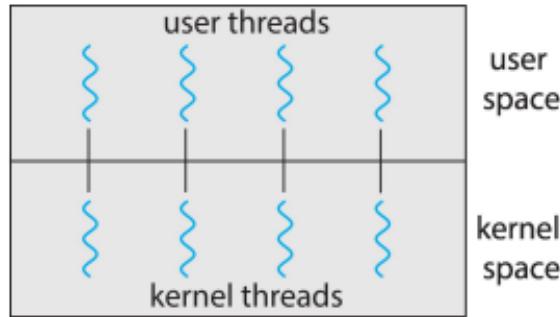


Figura 3.7: Il modello di *multithreading* uno a uno.

possono essere alcune restrizioni in modo da evitare la creazioni di troppi kernel threads (e quindi limitare la creazione di user threads). È comunque evidente che questo modello è sicuramente più **efficace** del modello precedente in quanto fornisce la possibilità di sfruttare a pieno un sistema multicore.

3.2.3 Many-to-Many

L'ultimo modello è un buon compromesso tra il modello *Many-to-One* e il modello *One-to-One*. Il modello molti a molti (figura 3.8) fornisce diversi vantaggi ed è più flessibile rispetto ai primi due. Non

esiste infatti la corrispondenza univoca, generalmente si hanno più user threads che fanno riferimento ad alcuni kernel threads (è come se una sala da 20 clienti fosse gestita da 5 camerieri). È possibile infatti

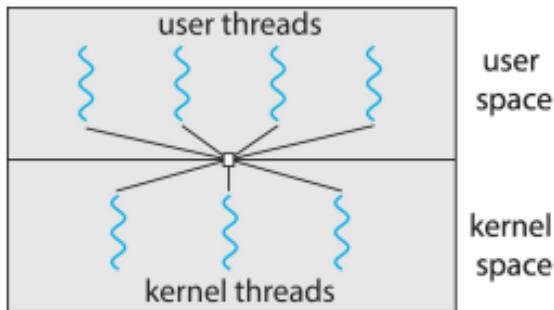


Figura 3.8: Il modello di *multithreading* molti a molti.

controllare in maniera più efficace i kernel threads e questo rende il modello *Many-to-Many* molto **robusto**. Ad ogni modo la sua implementazione è molto più complessa rispetto ai modelli precedenti. Per questo, a volte, si fa riferimento ad un *ibrido* tra il modello *Many-to-Many* e il modello *One-to-One*: stiamo parlando del **Two-level model** che consente sia la corrispondenza *1 a 1* che la corrispondenza *N a M*.

3.3 Librerie di thread

In questa piccola sezione ci interessiamo alle librerie disponibili per la creazione e la gestione di threads. Queste possono essere di due tipi:

- ◊ Librerie in **user space**, dove i thread sono gestiti completamente a livello utente (come nel caso di Pthreads);
- ◊ Librerie di tipo **kernel-level** che sono supportate dal sistema operativo e si appoggiano al kernel attraverso le *system calls*; questo comporta un livello maggiore di complessità nel kernel ma il programmatore ha meno da implementare.

Spendiamo due parole su **POSIX Threads**. Questa, non è propriamente una libreria ma è un insieme di specifiche (non implementazioni!) che aiuta con la creazione e la gestione di thread. **POSIX Threads** fornisce specifiche sia a livello di utente che a livello di kernel.

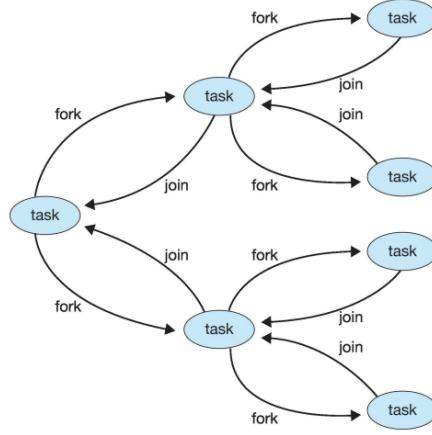


Figura 3.9: Rappresentazione grafica del modello fork-join per la risoluzione di un task.

3.4 Threading implicito

Cerchiamo ora di cambiare approccio: attraverso le librerie l'approccio era esplicito, stava al programmatore l'implementazione e/o la gestione dei thread. Nel momento in cui si fa riferimento a più threads, diventa sempre più difficile controllare la correttezza del codice. Ecco che si è pensato ad un altro approccio: il **threading implicito**. Con questo approccio i thread sono gestiti maggiormente dal compilatore oppure da librerie *run-time* le quali si occupano di creare e gestire i thread. In questa sezione vedremo alcuni dei metodi utilizzati per ottenere il threading implicito.

3.4.1 Modello fork-join

Questo tipo di modello si ispira alla creazione di processi (paragrafo 2.2.2). In questo modello, diversi threads sono divisi e, una volta che sono terminati, vengono uniti (figura 3.9). La divisione dei threads è una scelta presa completamente dalla libreria. Come possiamo notare dalla figura 3.9, in questo caso, la libreria sceglie un task da far risolvere ad un thread dove, eventualmente verrà spartita in altri due thread (che vengono appositamente creati) e così via fino a che il task non sia completamente risolto.

3.4.2 Thread pools e OpenMP

Altri due modelli molto importanti sono *thread pools* e OpenMP. Nel primo caso la libreria mette a disposizione un numero di thread (*pool* di thread) che attendono un task da risolvere. In questo modo, dato che i thread sono già pronti, non si spreca tempo per l'effettivo processo di creazione. Inoltre il problema della limitazione dei thread è risolto in quanto una volta creati quelli per la *pool*, non vengono creati altri thread.

Quando parliamo di OpenMP invece parliamo di una serie di **direttive** per il compilatore e un insieme di librerie per C, C++ e FORTRAN. Questo modello fornisce tutte le risorse per la condivisione della memoria tra i thread e i parallelismi. Essendo delle direttive, queste devono essere proprio scritte nel codice; per esempio: `#pragm omp parallel`.

3.5 Problematiche

Come vedremo in questo paragrafo, anche i thread portano a delle problematiche che devono essere gestite come, per esempio, l'interruzione di tali e il modo di implementare l'eliminazione di un thread.

3.5.1 Semantica `exec` e `fork`

Il primo dubbio che è necessario chiarire è il comportamento durante la chiamata della funzione `exec()`: se si fa una chiamata alla funzione `exec()`, vengono rimpiazzati tutti i thread del processo oppure solo il thread su cui è stata chiamata `exec`? Generalmente la funzione `exec` cancella il processo in esecuzione e, di conseguenza, tutti i suoi thread.

Il secondo dubbio, riguarda la funzione `fork()`: se si fa una chiamata a `fork()` ad un processo multithread, si effettua una copia a tutti i thread del processo oppure solo al thread su cui è stata chiamata la funzione? La risposta a questa domanda è che dipende dall'obiettivo della funzione `fork()`:

- ◊ Se la funzione deve cambiare subito il codice attraverso `exec()`, non vale la piena copiare tutti i thread se tanto si si che verranno eliminati.
- ◊ Se invece il nuovo processo deve supportare anch'esso il multithreading, allora ha senso effettuare una copia di tutti i thread e non solo del thread su cui è stata chiamata la funzione

3.5.2 Segnalazione ed eliminazione

I segnali sono usati per notificare un processo che un determinato evento è accaduto. Tali segnali però debbono essere gestiti: ecco che emerge la figura del **signal handler** che può essere di **default** oppure **user-defined**, ovvero definito dall'utente. Generalmente ogni segnale ha il suo specifico *default handler* che è utilizzato anche dal kernel. Cosa succede però nel caso in cui il sistema è multi-threading? Ci sono diverse opzioni:

1. Spedire il segnale al thread ad esso compatibile;
2. Propagare il segnale ad ogni thread del processo;
3. Mandare il segnale ad dei thread specifici del processo;
4. Assegnare ad uno specifico thread il compito di ricevere i segnali degli altri thread.

È inoltre importante riuscire a gestire la **cancellazione** dei thread. Questi però devono attivare la possibilità di essere cancellati: in altre parole, se un thread ha la cancellazione disabilitata non potrà essere cancellato fino a che non viene riattivata. Nel momento in cui la cancellazione è attivata, il thread può essere cancellato attraverso due tecniche:

- ◊ **Asincrona**, ovvero il thread termina immediatamente;
- ◊ **Deferred** che significa *posticipata*. Può ritornare utile nel momento in cui il thread sta compiendo un'operazione delicata (chiusura di un file) e non ha la possibilità di terminare immediatamente.

4 CPU SCHEDULING

In questa sezione ci occupiamo di tutti gli aspetti che concernono lo scheduling del processore, ovvero la pratica secondo la quale, attraverso dei precisi criteri, si sceglie quale processo all'interno della *ready queue* verrà eseguito.

4.1 Nozioni fondamentali

Prima di iniziare a discutere di scheduling vero e proprio è necessario avere chiari alcuni concetti importanti.

Burst. La prima è la nozione di burst. Chiamiamo **CPU burst** il periodo di tempo nel quale un processo esegue operazioni all'interno della CPU; diversamente, l'**I/O burst** è il tempo che il processo spende interfacciandosi con le periferiche di input e output. In particolare, un processo che occupa per molto tempo la CPU si dice *CPU bounded*, mentre nel caso di I/O si parla di un processo *I/O bounded*. In questo capitolo ci occupiamo solo di CPU burst.

CPU scheduler. Il CPU scheduler è il responsabile dell'organizzazione e dell'ordinamento della **coda dei processi** (figura 4.1). Le decisioni su quale processo deve essere eseguito prima stanno a lui. Questo tipo di decisioni avvengono generalmente quando un processo:

1. Passa dallo stato *running* a *waiting*;

2. Passa da *running* allo stato *ready*;
3. Passa dallo stato *waiting* allo stato *ready*;
4. Termina.

Osserviamo che per i punti 1 e 4 non è presente una vera e propria decisione: un nuovo processo deve essere messo in esecuzione. Questo non vale per i punti 2 e 3 dove si fa una decisione.

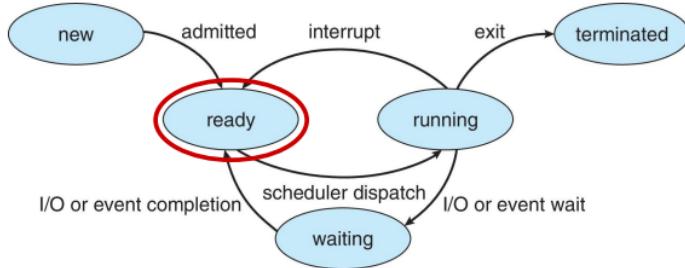


Figura 4.1: Lo scheduler entra in gioco nel passaggio da *ready* a *running*.

Preemption. Ci sono due macro gruppi di scheduling: preemptive e non. Uno scheduling è detto **non preemptive** nel momento in cui un processo non può essere fermato. In altre parole, quando la CPU è assegnata ad un processo, questo la utilizza fino a che non è terminato. Differentemente, uno scheduling è detto **preemptive** quando un processo può essere fermato per dare precedenza ad un altro per poi essere fatto ripartire: questo tipo di scheduling è sicuramente più performante e moderno; è infatti utilizzato nei sistemi operativi più diffusi come Windows, MacOS, Linux e altri. Ciò porta comunque a delle situazioni indesiderate come le **race conditions**: poniamo il caso di avere due processi che condividono dei dati; immaginiamo che il primo processo stia aggiornando questi dati ma allo stesso tempo il secondo processo li stia utilizzando. Questo è un problema dato che il processo due sta utilizzando dei dati che non sono consistenti dato che sono in fase di aggiornamento dal processo uno. Come vedremo nel capitolo 5, questo è un problema di sincronizzazione che va risolto.

Dispatcher. Nel momento in cui lo scheduler ha scelto quale processo verrà eseguito, il dispatcher si occupa di cambiare il processo nella CPU. In particolare viene effettuato il *context switch* (2.2.1), passa in *user mode* e va alla giusta locazione del programma per iniziare la sua esecuzione. Durante tutto ciò

la CPU però non lavora: è importante quindi minimizzare questa latenza e fare in modo che non se ne effettuino un numero troppo elevato al fine di mantenere un alta percentuale di utilizzo della CPU.

4.2 Algoritmi non preemptive

In questo paragrafo ci occupiamo dei primi algoritmi non preemptive, ovvero gli algoritmi che non fermano i processi che sono in esecuzione.

4.2.1 First-Come First-Served (FCFS)

Questo è l'algoritmo più banale da implementare: il primo processo che entra nella coda sarà anche il primo ad essere eseguito. Questo algoritmo ha un approccio praticamente identico al **FIFO** (*First In First Out*). Attraverso un semplice esempio, possiamo osservare che questo algoritmo non è efficiente. Poniamo che entrino nella coda 3 processi: P_1 , di durata 24 unità di tempo (in genere millisecondi) e P_2 e P_3 con durata di esecuzione 3. Osservando la figura 4.2 è banale notare che se P_1 fosse arrivato in

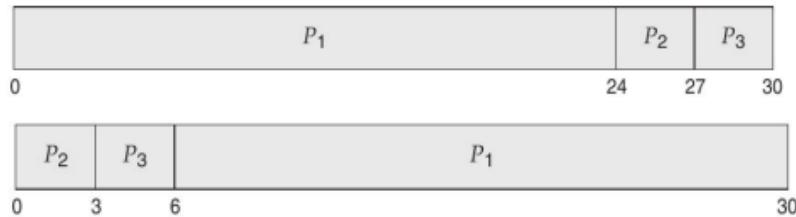


Figura 4.2: Diagramma di Gantt dell'algoritmo FCFS.

coda per ultimo, i processi P_2 e P_3 avrebbero aspettato meno. Possiamo dimostrarlo anche in maniera più matematica, attraverso dei brevi calcoli. Nel primo caso P_1 ha aspettato 0 prima di essere eseguito, P_2 ha aspettato l'esecuzione di $P_1 = 24$ mentre P_3 ha aspettato l'esecuzione di $P_1 + P_2 = 24 + 3 = 27$. Se provassimo a fare una media del tempo di attesa otteniamo:

$$\langle T \rangle = \frac{0 + 24 + 27}{3} = 17$$

Nel secondo caso invece la situazione migliora notevolmente in quanto P_2 aspetta 0, P_3 aspetta solamente l'esecuzione di $P_2 = 3$ e infine P_1 attende l'esecuzione di $P_2 + P_3 = 3 + 3 = 6$. Attraverso la stessa espressione matematica otteniamo che l'attesa media in questo caso diventa:

$$\langle T \rangle = \frac{6 + 0 + 3}{3} = 3$$

Diversi sono i problemi di questo algoritmo. Prima di tutto può generare il **convoy effect** (effetto convoglio): se viene eseguito un processo *I/O bounded*, tutti gli altri processi in coda devono aspettare che questo si "sblocchi" generando quindi un rallentamento generale. In secondo luogo, questo algoritmo di scheduling dipende dall'ordine di entrata dei processi e di conseguenza **non** è nemmeno possibili analizzare in modo **deterministico** le prestazioni dell'algoritmo.

4.2.2 Shortest-Job-First (SJF)

Come abbiamo notato dalla figura 4.2, se i processi più bervi sono eseguiti prima, il tempo medio di attesa $\langle T \rangle$ si abbassa. Implementiamo quindi un algoritmo che dia la precedenza ai processi con il tempo di esecuzione più breve tra quelli che sono presenti in coda. In questo algoritmo stiamo assumendo che siamo a conoscenza del CPU burst time di ciascun processo: si osserva che stiamo facendo un'ipotesi, spesso questo dato non è a disposizione. Se tutti i CPU burst sono conosciuti, l'algoritmo fornisce il minor tempo medio di attesa di un insieme finito di processi.

Il funzionamento di questo algoritmo è molto semplice: in base ai processi arrivati in coda, questi, prima di essere eseguiti, vengono riordinati in base al loro tempo di burst. Per esempio, poniamo di avere in coda un processo P_1 con un burst time di 6, P_2 da 8, P_3 da 7 e P_4 da 3. Osservando la figura

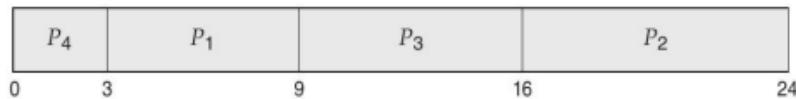


Figura 4.3: Diagramma di *Gantt* dell'algoritmo SJF

4.3 notiamo che i processi sono stati riordinati in modo tale da minimizzare il tempo medio d'attesa. In questo esempio il processo P_4 attende 0, il processo P_1 attende l'esecuzione di $P_4 = 3$, il processo P_3

aspetta la conclusione di $P_4 + P_1 = 3 + 6 = 9$ e infine il processo P_2 aspetta $P_4 + P_1 + P_3 = 3 + 6 + 7 = 16$. Ecco che il tempo di attesa medio diventa:

$$\langle T \rangle = \frac{3 + 16 + 9 + 0}{4} = 7$$

Anche in questo caso però se durante l'esecuzione è in coda un processo con un burst molto alto e entrano solo processi con un burst basso, è probabile che si verifichi una situazione di attesa perenne (chiamata **starvation**); vedremo, nel corso di questo capitolo, come ciò può essere evitato (4.4).

4.2.3 Stima del *CPU burst time*

Come abbiamo affermato poco fa, quasi mai il burst time è a disposizione. Si è quindi trovato un modo per stimare al meglio il burst time di un processo, in base ai processi che sono stati eseguiti in precedenza. Il metodo utilizzato è chiamato **exponential averaging** il quale, in essenza, dà peso maggiore ai processi eseguiti da poco tempo e, pian piano, più i processi sono remoti, meno influenza hanno sulla stima. Le variabili in gioco nella formula sono:

- ◊ t_i indica la durata del CPU burst del processo i -esimo, dove $i \in [0, n]$, dove n indica il numero di processi;
- ◊ τ_{n+1} rappresenta la stima, la predizione (*guess*), che si calcola sul processo che si sta per eseguire;
- ◊ α che è un coefficiente che indica quanto pesa la storia dei processi. Quando questo coefficiente è basso, la storia recente non conta, mentre quando è alto, la storia recente ha più peso (con $\alpha = 1$ si ha che solo l'ultimo processo influisce sulla stima). Generalmente si utilizza il valore $\alpha = \frac{1}{2}$.

Nel caso generale, con n processi si ha che per stimare il bursts dell' $n + 1$ -esimo:

$$\begin{aligned} \tau_{n+1} = & \alpha t_n + (1 - \alpha) \alpha t_{n-1} + \dots \\ & + (1 - \alpha)^i \alpha t_{n-i} + \dots \\ & + (1 - \alpha)^{n+1} \tau_0 \end{aligned}$$

Osserviamo ora la figura 4.4 che rappresenta come viene effettuata la *guess* in base alla storia dei processi ($\alpha = .5$). La prima colonna composta dalla coppia $\binom{6}{10}$ è la colonna "base", la partenza del nostro grafico. Con questi due dati, si calcola la seconda colonna $\binom{4}{8}$, dove $8 = \frac{10 + 6}{2}$; a questo punto si può calcolare la terza colonna $\binom{6}{6}$, dove il secondo $6 = \frac{4 + 8}{2}$. Così facendo si è in grado di calcolare una buona stima per il CPU burst time del processo corrente.

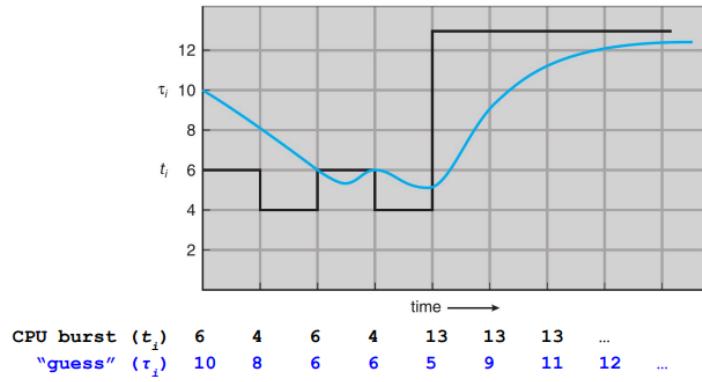


Figura 4.4: Grafico che indica come viene stimato il burst di un processo.

4.3 Algoritmi preemptive

In questo secondo paragrafo discutiamo invece di due algoritmi preemptive, ovvero algoritmi che possono fermare l'esecuzione di un processo per favorirne un altro.

4.3.1 Shortest-Remaining-Time-First (SRTF)

Il primo algoritmo che andremo a discutere è la versione preemptive del SJF: in questo caso viene servito per primo il processo al quale manca minor tempo per essere completato. Quindi se si sta eseguendo un processo P e arriva un processo Q il quale burst time è minore rispetto al burst time che manca a P per terminare, quest'ultimo viene fermato per dare la precedenza a Q . Una volta terminata l'esecuzione di Q , il processo P riprende da dove era stato fermato in precedenza. Prendiamo in considerazione i seguenti 4 processi:

Processo	Tempo di arrivo	Stima burst time
P_1	0	8
P_2	1	4
P_3	2	9
P_4	3	5

Osservando la figura 4.5, cerchiamo di capire come si comporta questo algoritmo nel momento in cui

i 4 processi sono inseriti all'interno della coda. Al tempo zero arriva in coda il processo P_1 , di durata 8.

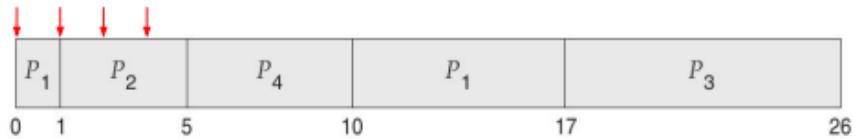


Figura 4.5: Diagramma di *Gantt* dell'algoritmo SRTF.

Al tempo 1 arriva in coda il processo P_2 che ha una durata di 4. A P_1 rimangono ancora $8 - 1 = 7$ unità di tempo prima di terminare, mentre a P_2 ne servono solo 4. P_1 viene quindi fermato e P_2 inizia la sua esecuzione (si effettua un *depatching*). Al tempo 2 e al tempo 3 sono aggiunti alla coda P_3 e P_4 i quali però hanno un burst time maggiore rispetto a P_2 che quindi termina l'esecuzione al tempo $1 + 4 = 5$. A questo punto rimangono P_1 , P_3 e P_4 . Viene eseguito P_4 in quanto il suo burst (5) è minore rispetto a quello di P_1 (7) e P_3 (9). Terminata l'esecuzione di P_4 inizia quella di P_1 e poi quella di P_3 .

Osserviamo che con questo algoritmo può capitare che sia in esecuzione un processo con un tempo di burst molto elevato e che poi continuino ad arrivare dei processi con un tempo di burst ridotto. In questo caso, il processo con il tempo maggiore verrebbe sempre interrotto dagli altri processi e non riuscirebbe mai a terminare andando quindi in una situazione di **starvation**. Come vedremo, una soluzione a questo problema è fornita dallo scheduling con priorità (4.4).

4.3.2 Round Robin (RR)

Passiamo ora ad un algoritmo un po' più sofisticato ed elegante: stiamo parlando del Round Robin. Alla base di questo algoritmo c'è il **quanto** di tempo (generalmente tra i 10 e i 100 millisecondi): ogni processo all'interno della coda, ha diritto ad essere eseguito per 1 quanto di tempo alla volta. Di conseguenza, ogni quanto di tempo viene effettuato un *context switch* e si prosegue ad un altro processo nella coda: si continua così in maniera ciclica finché ciascun processo viene eseguito completamente lasciando spazio ai nuovi.

Osserviamo che se $q \rightarrow \infty$ si ha un comportamento FIFO, molto simile all'algoritmo First-Come First-Served. Allo stesso tempo però q deve comunque essere maggiore del tempo che ci si impiega per effettuare un context switch (ordine dei μs), altrimenti la CPU viene sprecata solo per fare i context switch al posto di effettivamente eseguire i processi.

Consideriamo il caso in cui $q = 4$ e in coda sono presenti i tre processi utilizzati nell'algoritmo FCFS:

Processo	Stima burst time
P_1	24
P_2	3
P_3	3

Il comportamento del Round Robin, in questo caso, è illustrato nella figura 4.6. Osserviamo che il RR,

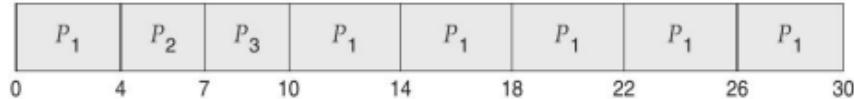


Figura 4.6: Diagramma di *Gantt* dell'algoritmo RR.

per il processo P_1 , ogni $q = 4$, si ferma per fare spazio agli altri processi mentre, per i processi P_2 e P_3 , che durano meno di un quanto, quando terminano il Round Robin, non aspetta la scadenza del quanto per eseguire un altro processo ma comincia subito, che è un comportamento ragionevolmente ovvio.

4.3.3 Reattività

Cerchiamo ora di capire il vantaggio che forniscono gli algoritmi di scheduling preemptive (in particolare il RR) rispetto agli algoritmi non preemptive. Prendiamo come esempio i seguenti processi:

Processo	Burst time
P_1	6
P_2	3
P_3	1
P_4	7

Poniamo ora che il quanto di tempo q sia 4. Calcoliamo ora il turnaround time medio tra questi processi: P_1 viene eseguito per 4 unità, dopo di chè viene fermato (gliene rimangono 2) e viene eseguito P_2 che termina ($T_{P_2} = 4 + 3 = 7$); viene quindi eseguito P_3 che termina anche lui ($T_{P_3} = 7 + 1 = 8$). A questo punto inizia l'esecuzione di P_4 che viene fermato dopo 4 unità (ne rimangono ancora 3) e viene fatto

ripartire P_1 che termina ($T_{P_1} = 8 + 4 + 2 = 14$) e infine viene fatto terminare anche P_4 ($T_{P_4} = 14 + 3 = 17$).

Per trovare il turnaround time medio si effettua la media dei 4 turnaround trovati.

$$\langle T \rangle = \frac{T_{P_1} + T_{P_2} + T_{P_3} + T_{P_4}}{4} = \frac{14 + 7 + 8 + 17}{4} = \frac{46}{4} = 11.5$$

Osserviamo però che se avessimo utilizzato l'algoritmo SJF (vedi 4.2.2) il turnaround time medio è 8 che è minore rispetto a quello fornito da RR. Ciò significa che utilizzare un algoritmo pre-emptive, in termini di tempistiche, non è necessariamente la scelta migliore, è semplicemente un altro modo per schedulare l'esecuzione dei processi, ma non ne garantisce il miglioramento della prestazione. Allora perchè utilizzare questi algoritmo? Perchè migliora la reattività (**responsiveness**) del sistema. Ipotizziamo di avere un coda moltissimi processi che stanno in esecuzione. Un algoritmo non preemptive esegue uno ad uno (secondo determinati criteri, più o meno efficienti) ma tutti i processi devono sempre stare in attesa che uno termini. Il RR invece garantisce che tutti i processi vengano eseguiti per almeno un certo quanto q di tempo: è quindi un algoritmo più **equo** rispetto agli algoritmi non preemptive.

4.4 Scheduling con priorità

Fino ad ora abbiamo trattato i processi in modo equo se non per la stima del *burst time*. Introduciamo ora una seconda informazione, la **priorità** che non è altro che un numero che indica quanto sia importante (urgente) l'esecuzione di un determinato algoritmo. La priorità può essere sia legata al CPU burst time ma può anche essere legata ad altri aspetti.

Arriva però un problema: la **starvation**. Anche in questo caso, come nel SRTF, può capitare che i processi che hanno una priorità di grado molto basso non vengano mai eseguiti in quanto sono sempre presenti processi con una priorità più alta. Con l'introduzione della priorità però, se un processo è da troppo tempo in coda, si aumenta di un grado la priorità al fine da mandarlo in esecuzione. Questa tecnica rappresenta allegoricamente l'invecchiamento (**aging**) del processo nella coda di attesa.

Vediamo ora un primo esempio di scheduling con priorità puro. Abbiamo a che fare con i 5 processi seguenti:

Osserviamo che noi consideriamo come numero più basso la priorità più alta (di conseguenza P_2 è il processo con priorità più alta e P_4 quello con priorità più bassa). Detto ciò procediamo con il diagramma di Gantt (figura 4.7). Notiamo infatti che i processi sono eseguiti in ordine in base alla

Processo	Stima burst time	Priorità
P_1	10	3
P_2	1	1
P_3	2	4
P_4	1	5
P_5	5	2



Figura 4.7: Diagramma di Gantt dell'algoritmo basato puramente sulla priorità.

loro priorità: P_2, P_5, P_1, P_3 e P_4 . Ovviamente, se durante l'esecuzione di P_1 (che ha priorità 3) fosse arrivato in coda un algoritmo con priorità 2, P_1 sarebbe stato interrotto per favorire l'esecuzione del nuovo processo. Inoltre, nel caso in cui due processi abbiano la stessa priorità si segue la dinamica FIFO, ovvero il primo processo che entra nella coda viene eseguito per primo rispetto ai processi con medesima priorità

4.4.1 Scheduling con priorità e RR

Cerchiamo ora di raffinare un po' di più l'algoritmo fondendo la priorità con l'algoritmo di Round Robin. In particolare, nel momento in cui si hanno più processi che hanno lo stesso livello di priorità, al posto di seguire un approccio FIFO, si utilizza il Round Robin, garantendo quindi una maggiore reattività ai processi. Partiamo dal seguente set di processi:

Processo	Stima burst time	Priorità
P_1	4	3
P_2	5	2
P_3	8	2
P_4	7	1
P_5	3	3

Come è possibile osservare dalla figura 4.8, notiamo che il processo 4, essendo che è l'unico processo con priorità 1, viene eseguito per primo e non viene interrotto da nessun'altro processo. Dopo di che

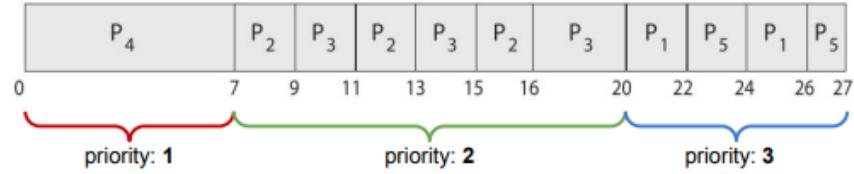
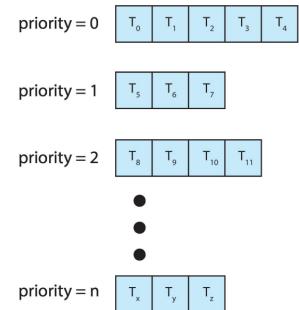


Figura 4.8: Diagramma di *Gantt* dello scheduling con priorità unito all'algoritmo RR per i processi con lo stesso grado di urgenza.

si effettua l'algoritmo RR con $q = 2$ sui processi 2 e 3 in quanto hanno la stessa priorità. Infine, si fa lo stesso procedimento che con i processi 1 e 5 che hanno priorità 3.

4.4.2 Coda multilivello

Il fatto che per ogni grado di priorità venga eseguito il Round Robin ci porta a creare uno scheduling multilivello dove ogni livello corrisponde ad un grado di priorità. Di conseguenza nell'esecuzione viene prima eseguito il primo livello attraverso il RR; dopo di che si passa al secondo livello e così via fino all'ultimo grado di priorità. In particolare, le priorità più alte sono assegnati a processi che hanno un bisogno **real-time** (come per il controllo di un braccio robotico) che poi sono seguiti dai processi di sistema etc. Questo ci porta al caso più complesso: immaginiamo che i processi non solo debbano essere inseriti nella coda giusta ma che questi debbano anche essere in grado di sposarsi da una coda all'altra e quindi cambiando il loro grado di priorità. Stiamo infatti parlando delle **Multilevel Feedback Queue** (MFQ).



4.5 Scheduling multiprocessore

Fino ad ora abbiamo discusso di algoritmi tenendo conto del fatto che il sistema disponesse di un singolo processore; sappiamo bene però che nei sistemi moderni ormai una situazione del genere non si verifica più, con sistemi multicore.

Come possono essere gestiti diversi threads all'interno di queste architetture multiprocessore? Nel caso del **Symmetric multiprocessing (SMP)**

4.5.1 Symmetric multiprocessing (SMP)

Partiamo dalla situazione più semplice, nel casi in cui abbiamo a che fare con un'architettura SMP. In questo caso infatti abbiamo *cores* che vengono gestiti in modo simmetrico (vedi paragrafo 3.1) e, disponendo di n corse, l'unico problema da risolvere è come questi possano andare a gestire n thread. Come è infatti possibile notare dalla figura 4.9, un modo può essere quello di utilizzare una *ready-queue* comune a tutti i cores oppure quello di creare una coda apposita ad ogni core per gestire i processi. Entrambe le soluzioni sono più che lecite anche se ad oggi i sistemi operativi moderni tendono ad

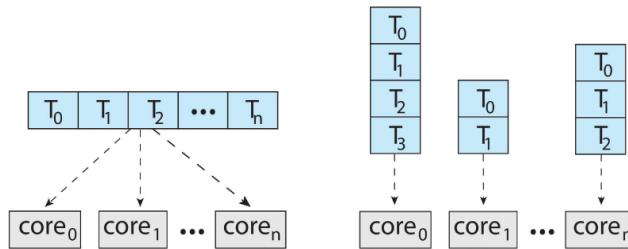


Figura 4.9: Ci sono diverse modi per gestire n threads con n cores.

usare la seconda in quanto la prima soluzione, essendo si che si ha a che fare con una coda condivisa è complicato gestire la condivisione di tale risorsa tra i cores (problemi di *race conditions*).

4.5.2

4.6 Scheduling real-time

4.7 Valutazione di un algoritmo

5 SINCRONIZZAZIONE

Come abbiamo visto nel capitolo 2, i processi possono essere eseguiti sia in parallelo che in concorrenza, la quale non è altro che un modo per far apparentemente girare due processi in maniera parallela quando in realtà stanno condividendo lo stesso *core* del sistema. Un processo può quindi essere interrotto in qualunque momento da un algoritmo di scheduling (come il *Round Robin*) per fare spazio ad un altro processo. Cosa succede però se viene interrotto in un momento in cui sta accedendo nella memoria condivisa con un altro processo e, lasciando spazio all'altro, vengono modificati dei dati? Capiremo, in questa sezione, che l'obiettivo da raggiungere è quindi la **cooperazione** tra processi in modo tale che non si verifichino situazioni spiacevoli.

5.1 Sezione critica

Poniamo ora di avere a che fare con due processi i quali effettuano due `fork` simultanei (figura 5.1). Supponiamo di avere una variabile, chiamata `next_available_pid` che contiene il primo `PID` disponibile. Se i due processi, in questo caso P_0 e P_1 , fanno accesso alla variabile in maniera simultanea, verranno creati due figli con lo stesso `PID`. I due processi hanno fatto un accesso in maniera non esclusiva alla stessa zona di memoria condivisa generando quindi un problema.

Quello che abbiamo appena visto non è altro che un esempio di **critical section problem**. Generalizzando, possiamo dire che ogni processo è composto di un segmento di codice che deve avere un **accesso esclusivo** in quanto ha un accesso in scrittura dove aggiorna tabelle, modifica variabili o

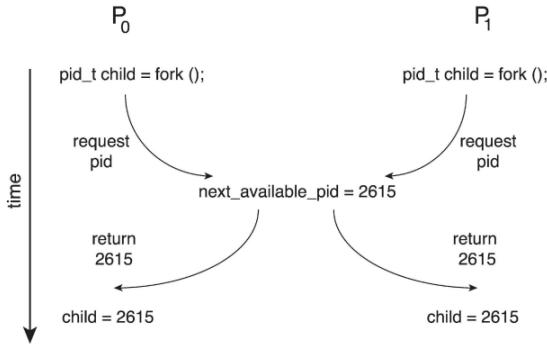


Figura 5.1: Creazione di due processi figli con lo stesso P ID.

scrive su files. Di conseguenza si cerca di non interrompere un processo nel momento in cui questo è nella sua **sezione critica**.

5.1.1 Requisiti

È stato stilato un elenco di 3 punti di requisiti che devono essere rispettati affinché non vengano generate situazioni dove un processo viene interrotto durante la sua sezione critica.

1. **Mutua esclusione:** se il processo P sta eseguendo la sua parte critica nessun altro processo deve interromperlo;
2. **Progresso:** bisogna garantire che nessun processo rimanga in un'attesa perenne mentre aspetta la conclusione della parte critica di un altro processo. In altre parole, l'esecuzione della sezione critica di un processo non può essere posticipata in maniere indeterminata;
3. **Attesa limitata:** deve esistere un limite massimo per cui un processo deve concludere la sua parte critica.

5.1.2 Soluzioni inefficienti

Una prima idea banale è quella di **disabilitare gli interrupt** nel momento in cui un processo (un *thread*) sta eseguendo la sua parte critica. Questa però è una soluzione poco elegante e anche poco funzionale. Si rischia infatti di far aspettare molto altri processi i quali necessitano di eseguire la loro sezione critica (starvation).

Si è quindi pensato ad una soluzione un po' più elegante, ma che comunque è una forma embrionale rispetto a ciò che vedremo. Stiamo parlando di una semplice **soluzione software** tra due processi dove entrambi condividono una variabile *booleana* `turn` che indica di chi è il turno per eseguire la sezione critica.

```

1  while(true) {
2      while(turn == j); /* attendo che sia il mio turno */
3      /* SEZIONE CRITICA */
4      turn = j; /* rimando il turno all'altro processo */
5      /* sezione rimasta */
6  }

```

Con questa soluzione, è rispettata la richiesta di mutua esclusione ma i punti (2) e (3) non sono rispettati: il processo `j` potrebbe anche durare un'ora e ci sarebbe un'ora di attesa. Non sono quindi i rispettati i limiti di attesa e nemmeno il progresso dei processi.

5.2 Soluzione di Peterson

Una prima soluzione un po' più raffinata è quella di Peterson. In questa soluzione i due processi condividono due variabili:

- ◊ `turn` che è la variabile che indica di quale processo è il turno;
- ◊ `flag[2]` che è un array di due valori booleani che indicano se il processo *i*-esimo è pronto o meno per entrare nella sezione critica.

Come possiamo notare dal seguente segmento di codice, sono stati aggiunti alcuni controlli.

Codice 5.1: Soluzione di Peterson

```

1  while (true) {
2      flag[i] = true;
3      turn = j;
4      while (flag[j] && turn == j); /* aspetto fino a che non sono pronto e
   fino a che j non ha finito */
5      /* SEZIONE CRITICA */

```

```

6     flag[i] = false;
7     /* sezione rimasta */
8 }
```

Osserviamo che possiamo intendere l'esecuzione della parte critica di un processo come una galleria a senso unico alternato: se un processo deve entrarci ma al suo interno ce n'è già un altro attende, altrimenti ci entra. L'esecuzione in "parallelo" è quindi mantenuta dato che un processo può essere molto lontano dalla galleria mentre il secondo la sta attraversando. Con questo modello, quindi, tutti e 3 i requisiti sono rispettati.

5.2.1 Architetture moderne

Ciò nonostante, questa soluzione diventa obsoleta per i sistemi moderni *multicore* e *multithread*. Questo perché nelle architetture moderne il compilatore si prende la libertà di riordinare e riorganizzare il codice. Vediamo un esempio di due thread (assumiamo che le variabili condivise siano *x* e *flag*):

```

1  /* THREAD 1 */
2  while(!flag);
3  print x
4
5  /* THREAD 2 */
6  x = 100;
7  flag = true;
```

Una volta eseguito il codice, ci aspettiamo che l'output sia 100. Però, se le istruzioni nel thread 2 vengono invertite, il risultato è completamente diverso, perché il thread 1 viene eseguito prima che la variabile *x* venga cambiata a 100.

```

1  /* THREAD 2 */
2  flag = true;
3  x = 100;
```

Al fine di fare in modo che la soluzione di Peterson funzioni anche nelle architetture moderne si introduce la **memory barrier**: questa è un'istruzione che rende le modifiche effettuate in memoria

visibili a tutti i processori (*cores*). Questo tipo di modello di memoria si dice *strongly ordered*, ovvero una modifica in memoria è propagata immediatamente a tutti i core; si contrappone al *weakly ordered* dove una modifica in memoria non è propagata istantaneamente.

Con l'introduzione delle memory barrier ora la soluzione di Peterson rimane valida. Questo perché quando viene invocato un `memory_barrier()` il sistema si assicura che tutti i load e store in memoria vengano completati prima di ogni altro load e store. Quindi anche se il codice venisse riordinato, la memory barrier si assicura che le modifiche in memoria siano completate.

5.3 Sincronizzazione via hardware

Tra le diverse soluzioni che abbiamo visto, la più gettonata rimane comunque l'implementazione di particolari **istruzioni hardware** che permettono di modificare il contenuto di una *word* in memoria oppure di effettuare uno *swap* di due word.

5.3.1 Test and set

La prima delle due istruzioni HW che discuteremo è la `test_and_set` dove prende un input booleano e lo memorizza in una variabile temporanea `tmp`; in secondo luogo il valore in ingresso viene settato a `true` e infine la variabile temporanea viene restituita. Lo pseudocodice di quest'istruzione è il seguente (ricordiamo che sono istruzioni a livello HW, quindi il codice è solo a scopo concettuale):

```
1 | boolean test_and_set(boolean *target) {  
2 |     boolean tmp = *target; /* salvo il valore di target */  
3 |     *target = true; /* setto a TRUE */  
4 |     return tmp;  
5 | }
```

Si osserva che dopo l'esecuzione dell'istruzione, il valore di `target` è sempre `true` e viene ritornato il vecchio valore della variabile in ingresso. Ricordiamo infine che è un'istruzione **atomica**, ciò significa che non può essere interrotta.

Vediamo ora come questa istruzione possa esserci utile per l'esecuzione della sezione critica di un processo:

Codice 5.2: Utilizzo test_and_set

```
1 while(true) {  
2     while(test_and_set(&lock)); /* lock = true, attendo per la risorsa,  
3         ora e' occupato */  
4     /* lock = false, posso cominciare a eseguire la parte critica perch'e'  
5         si e' liberato; metto lock = true */  
6     /* SEZIONE CRITICA */  
7     lock = false; /* lock e' libero */  
8     /* sezione rimanente */  
9 }
```

5.3.2 Compare and swap

La seconda istruzione HW che andiamo a discutere si chiama compare_and_swap e, come per la precedente, anch'essa è un'istruzione atomica. Come vedremo, questa istruzione, ha ben 3 variabili in ingresso:

- ◊ value, che indica il valore da modificare;
- ◊ expected, che indica il valore che ci si aspetta contenga value;
- ◊ new_value, ovvero il nuovo valore con cui vogliamo cambiare value.

Osserviamo ora lo pseudocodice dell'istruzione:

```
1 int compare_and_swap(int *value, int expected, int new_value) {  
2     int temp = *value;  
3     if (*value == expected)  
4         *value = new_value;  
5     return temp; /* ritorna vecchio valore di value */  
6 }
```

Notiamo che, a differenza di test_and_set, in questo caso il valore di value viene modificato solo nel momento in cui coincide con il valore che ci aspettiamo questo viene modificato con il valore inserito (new_value). Capiamo ora come questa istruzione possa essere utilizzata per la gestione della sezione critica e come questa sia più **flessibile** rispetto alla precedente.

Codice 5.3: Utilizzo di compare_and_swap

```
1 while(true) {  
2     while(compare_and_swap(&lock, 0, 1) != 0);  
3         /* lock = 1 = true, continuo ad aspettare */  
4         /* lock = 0 = false, viene ritornato 0, quindi esco dal while, e  
5            occupo lo spazio, lock = 1 */  
6         /* SEZIONE CRITICA */  
7         lock = 0;  
8         /* sezione rimasta*/  
9     } }
```

5.3.3 Variabili atomiche

Istruzioni come compare_and_swap sono utilizzate per comporre dei blocchi per comporre altri oggetti di sincronizzazione. Uno tra questi oggetti è la variabile atomica che fornisce degli aggiornamenti elementare a dei dati primitivi. Segue infatti l'esempio di increment() dove il valore intero v viene incrementato senza interruzioni

Codice 5.4: Esempio di una variabile atomica increment()

```
1 void increment(atomic_int *v) {  
2     int temp;  
3     do{  
4         temp = *v;  
5     } while (temp != compare_and_swap(v, temp, temp+1));  
6 }
```

5.4 Mutex lock e Semafori

Le soluzioni che abbiamo visto in precedenza erano intricate e spesso non erano accessibili da applicazioni esterni; inoltre creano ulteriori complicazioni in sistemi multithreading. A questo

proposito i progettisti di sistemi operativi hanno costruito diversi *tool* al fine di risolvere il problema della sezione critica.

Il primo che andremo a vedere è il **mutex lock** (MUTual EXclusion lock) che è una variabile booleana che indica se il **lock**, ovvero la sezione critica è disponibile o meno. Questa è protetta da due istruzioni atomiche di acquisizione e rilascio:

- ◊ `acquire()` che blocca la CS¹;
- ◊ `release()` che rilascia il lucchetto e quindi la CS è liberata.

Spesso queste sono implementate via HW attraverso istruzioni come `compare_and_swap`.

Come possiamo notare dallo pseudocodice sottostante, ci troviamo in una situazione di **busy waiting** in quanto il thread continua ad entrare nel ciclo fino a che non acquisisce il lucchetto. È evidente che stiamo sprecando CPU, che potrebbe essere usata per altri thread.

```
1 | while(true) {  
2 |     acquire();  
3 |     /* SEZIONE CRITICA */  
4 |     release();  
5 |     /* sezione rimanente */  
6 | }
```

A questo proposito, questa pratica è raffinata tramite i **semafori**, che non sono altro che una **variabile intera** (`S`) accessibile, anche in questo caso attraverso due operazioni atomiche: `wait()` e `signal()`.

```
1 | wait(S) {  
2 |     while (S<=0); /* busy wait */  
3 |     S--;  
4 | }  
5 |  
6 | signal(S) {  
7 |     S++;  
8 | }
```

¹CS sta per *Critic Session*, ovvero sezione critica

Se la variabile S può assumere solo il valore di 0 o 1, si parla di semafori **binari** (ovvero dei mutex lock), altrimenti si fa riferimento ai **counting semaphores**.

Attraverso i semafori abbiamo la possibilità di fare eseguire una parte di processo prima che un altro processo inizi. Siano P_1 e P_2 due processi che contengono il segmento S_1 ed S_2 ; per fare in modo che S_1 sia eseguito prima di S_2 possiamo utilizzare i semafori.

```
1 P1:
2     S1;      /* eseguo S1 */
3     signal(synch); /* dico a P2 che ho finito */
4
5 P2:
6     /* busy wait */
7     wait(synch); /* aspetto che P1 abbia finito */
8     S2;      /* eseguo S2 */
```

Possiamo notare però che anche in questo caso P_2 è in una fase di busy waiting.

5.4.1 Waiting queue

Per evitare che accada è necessario implementare una **waiting queue**:

Codice 5.5: Struttura del semaforo con waiting queue

```
1 typedef struct {
2     int value;
3     struct process *list; /* indica la prossima entry nella lista */
4 } semaphore
```

Qui abbiamo a disposizione altre due operazioni per implementare la sincronizzazione tra processi:

- ◊ `sleep()` che mette a dormire il processo o il thread e lo inserisce nella waiting queue;
- ◊ `wakeup()` che rimuove il primo processo della coda e lo inserisce nella *ready queue*, coda che contiene i processi pronti per essere eseguiti.

A questo punto possiamo ridefinire le due operazioni basi del semaforo: `wait` e `signal`.

Codice 5.6: Utilizzo di semaforo con *waiting queue*

```

1  wait(semaphore *s) {
2      S->value--; /* decremento il valore del semaforo */
3      if (S->value < 0) {
4          aggiungi processo in S->list;
5          sleep(); /* evito il busy waiting, mi tolgo dalla CPU e aspetto che
6              signal() mi svegli dalla coda */
7      }
8
9  signal(semaphore *S) {
10     S->value++; /* incremento il valore del semaforo */
11     if (S->value <= 0) {
12         rimuovii processo in S->list;
13         wakeup();
14     }
15 }
```

Bisogna però fare attenzione a utilizzare `wait` e `signal` nel modo corretto: non si può prima invocare `signal` e poi `wait` oppure invocare `wait` due volte di seguito. Questo utilizzo incorretto delle due operazioni può generare errori.

5.5 Monitor

Cerchiamo ora di risolvere il problema dei semafori: le operazioni di `wait` e `signal` erano lasciate al programmatore. Questa scelta ha come unica conseguenza la generazione di errori da parte dell'utilizzatore (che, come vedremo più avanti prendono il nome di **deadlock**, capitolo 6). Cerchiamo quindi di implementare un'astrazione dei semafori al fine di limitarne i danni. Queste astrazioni di alto livello sono proprio i **monitor** i quali garantiscono l'esecuzione di un processo alla volta.

5.5.1 Struttura e implementazione

La struttura di astrazione del monitor è formata da una variabile condivisa da tutti i processi, da del codice di inizializzazione per il monitor e da tante funzioni o procedure (ogni processo ha le stesse procedure).

Codice 5.7: Struttura del monitor

```
1 monitor name{  
2     /* dichiarazione della variabile condivisa */  
3     codice di inizializzazione (...) { ... }  
4  
5     procedura F1 (...) { ... }  
6     procedura F2 (...) { ... }  
7     /* ... */  
8     procedura Fn (...) { ... }  
9 }
```

Come possiamo osservare dalla figura 5.2, concettualmente, il monitor raggruppa tutti i processi attraverso dei dati condivisi in una coda.

Per implementare un monitor ci appoggiamo sui semafori, in particolare su un mutex:

```
1 semaphore mutex = 1; /* semaforo binario = mutex */  
2  
3 wait(mutex) /* aspetto che il mutex si liberi per occuparlo */  
4     /* corpo della funzione Fi */  
5 signal(mutex) /* libero il mutex **/
```

È quindi necessario modificare ogni procedura del monitor bloccando e poi rilasciando il mutex. In questo modo, a differenza dei semafori, le operazioni di `wait` e `signal` sono già implementate all'interno delle procedure che eseguono quelle operazioni, al posto del programmatore. Un esempio di utilizzo completo è presente nella sezione 5.6.3.

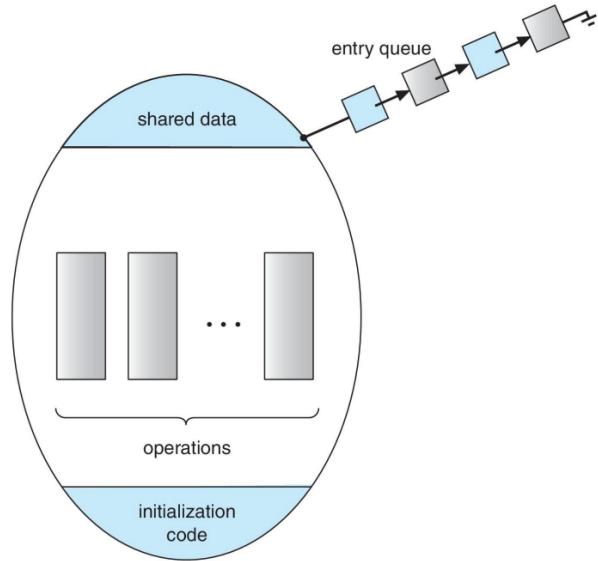


Figura 5.2: La struttura di astrazione del monitor.

5.5.2 Variabile di condizione

Oltre alla struttura base, illustrata nel paragrafo precedente, possiamo anche inserire le *condition variables* (rappresentazione in figura 5.3) che possono eseguire le stesse due operazioni dei semafori: `wait` e `signal`. In questo caso però le due operazioni sono modificate e non portano ad errori se utilizzate male. Ricordiamo che, nel caso dei semafori, se `signal` fosse stato invocato senza aver prima invocato il corrispettivo `wait`, avrebbe erroneamente incrementato il contatore. In questo caso invece, se non è presente alcun processo in `wait`, allora `signal` non fa nulla.

Emerge ora un piccolo dubbio. Se ci sono diversi processi che sono in `wait`, quando viene effettuato un `x.signal()`, quale di questi processi viene eseguito per prima? Notiamo che ci troviamo in una situazione molto simile allo scheduling (capitolo 4) solo che, in questo caso non è più a livello di sistema operativo ma è a livello di monitor e di accesso alla sezione critica. Per risolvere questo problema ci possono essere algoritmi come FCFS oppure a livello di priorità: chi ha la priorità più alta viene fatto eseguire prima. Spesso la priorità è definita in base al tempo di utilizzo della sezione critica, ma si fa presente che se entrano nella coda processi con un tempo di utilizzo molto breve, i processi con un utilizzo lungo possono essere in attesa perenne ed andare in **stravation**. Ecco che non viene più

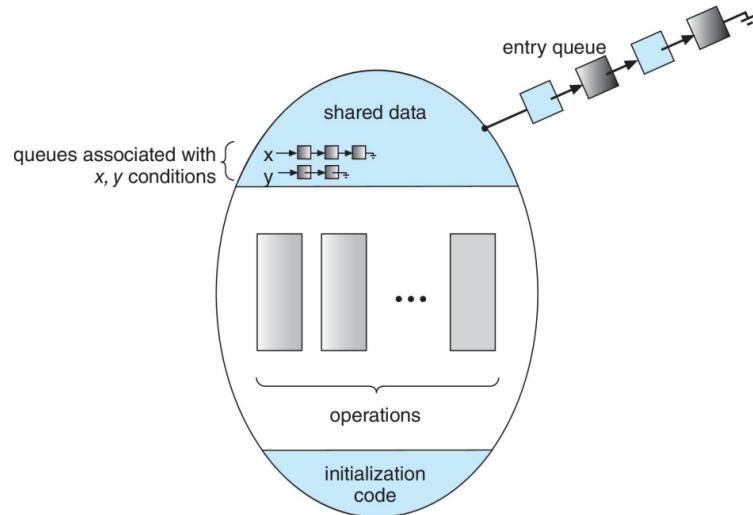


Figura 5.3: Caption

rispettato il terzo requisito, ovvero che l'attesa di un processo per la sezione critica deve essere limitata (vedi paragrafo 5.1.1).

5.6 Problemi comuni della sincronizzazione

5.6.1 Buffer limitato

Il primo problema che andiamo a discutere è il problema del buffer limitato. Abbiamo a disposizione un buffer che contiene n elementi, ovvero il numero di processi in coda per accedere alla loro sezione critica. Disponiamo inoltre di tre semafori:

- ◊ `mutex` che consente l'accesso al buffer in maniera esclusiva (inizializzato ad 1);
- ◊ `full` che segnala in numero di elementi contenuti all'interno del buffer (inizializzato a 0);
- ◊ `empty` che indica la quantità di spazi disponibili all'interno del buffer, sarebbe $n - full$ (inizializzato ad n).

Analizziamo la soluzione del problema, che è molto simile al problema del produttore e consumatore.

Codice 5.8: Problema del buffer limitato

```
1 | /* Prodottore */
```

```

2  while (true) {
3      /* produco un elemento */
4      wait(empty); /* aspetto che ci sia spazio all'interno del buffer,
   ovvero che ci sia almeno una locazione libera (empty > 0) */
5      wait(mutex); /* ora che c'e' un posto libero, richiedo l'accesso alla
   sezione critica */
6      /* SEZIONE CRITICA: aggiungo il codice nel buffer */
7      signal(mutex); /* libero la sezione critica */
8      signal(full); /* incremento di 1 il numero di elementi nel buffer
9  }
10
11 /* Consumatore */
12 while(true) {
13     wait(full); /* aspetto che ci siano >0 elementi nel buffer */
14     wait(mutex); /* aspetto l'accesso esclusivo nel buffer */
15     /* SEZIONE CRITICA: rimuovo l'elemento dal buffer */
16     signal(mutex); /* libero la sezione critica */
17     signal(empty); /* incremento il numero di locazioni libere */
18 }

```

Osserviamo che i comando `wait(empty)` e `wait(mutex)` all'interno di produttore **non** possono essere invertiti. Se prima infatti viene bloccato il mutex, il consumatore non sarà più in grado di liberare il buffer che, se è pieno, porta ad un'attesa perenne il produttore in `wait(empty)`: siamo ricaduti in una situazione di *deadlock*.

5.6.2 Problema dei lettori e degli scrittori

Il secondo problema è detto dei lettori e degli scrittori. In questo caso: (1) più lettori possono leggere lo stesso dataset (che può essere anche un file su disco) contemporaneamente (per definizione, non possono modificarlo) e (2) lo scrittore deve poter scrivere sul dataset senza che quest'ultimo venga letto. Solo nel momento in cui ha terminato (mutua esclusione), allora i lettori possono effettuare nuovamente l'accesso simultaneo.



Figura 5.4: Disposizione dei 5 filosofi.

Non ci soffermiamo sul codice, ci limitiamo a dire che è sufficiente utilizzare due semafori:

- ◊ `rw_mutex`, semaforo binario inizializzato ad 1, che indica che la risorsa è occupata dallo scrittore;
- ◊ `mutex`, anch'esso un semaforo binario che segnala che la risorsa è in lettura da almeno un lettore;
- ◊ `read_count`, un intero che conta il numero di lettore che stanno leggendo la risorsa.

In questo caso i problemi generati sono due. Il primo è che c'è la possibilità che lo scrittore non riesce mai ad effettuare l'accesso, entra quindi in una fase di attesa perenne. Il secondo problema che emerge è che quando lo scrittore ha effettuato l'accesso, nessun lettore è in grado di accedere: anche in questo caso si può ricadere in una fase di *starvation*.

5.6.3 Problema dei 5 filosofi

Il problema dei 5 filosofi, rimane il più conosciuto (e il più importante). Un filosofo, può pensare o mangiare. Facendo riferimento alla figura 5.4, osserviamo che i filosofi sono disposti in un tavolo rotondo che ha al centro una scodella di riso (che allegoricamente indica il dataset, ovvero la risorsa condivisa). Alla destra e alla sinistra di ciascun filosofo è presente una bacchetta: questa rappresenta il

semaforo in quanto se un filosofo smette di pensare e inizia a mangiare, non può farlo se il filosofo alla sua sinistra o alla sua destra sta mangiando, in quanto almeno una delle due bacchette sono occupate. Proviamo a dare una prima soluzione a questo problema di sincronizzazione con i semafori.

Codice 5.9: Risoluzione del problema dei filosofi con i semafori

```

1  while(true) {
2      wait(chopstick[i]); /* attende la bacchetta a sinistra */
3      wait(chopstick[ (i + 1) % 5]); /* attende la bacchetta a destra */
4      /* SEZIONE CRITICA: mangia dalla ciotola di riso */
5      signal(chopstick[i]); /* rilascia la bacchetta a sx */
6      signal(chopstick[ (i + 1) % 5]); /* rilascia la bacchetta a dx */
7      /* pensa */
8  }

```

Sembrerebbe tutto ottimo, ma cosa succede se tutti e 5 i filosofi prendono la loro bacchetto sinistra allo stesso tempo? Entrano in una situazione di attesa infinita (deadlock) un quanto tutte le loro bacchette destre sono occupate dal filosofo alla loro destra. Proviamo a risolvere questo nuovo problema attraverso i **monitor**.

Codice 5.10: Risoluzione del problema dei filosofi con i monitor

```

1  monitor DiningPhilosophers{ /* creo la struttura astratta */
2      enum { THINKING, HUNGRY, EATING } state[5]; /* per ogni filosofo e' definito lo stato in cui si trova */
3      condition self[5]; /* condition variable */
4
5      void test(int i){
6          if ((state[(i + 4) % 5] != EATING) /* se il vicino di sinistra non sta mangiando */ &&
7              (state[i] == HUNGRY) /* se ho fame */ &&
8              (state[(i + 1) % 5] != EATING)) /* se il vicino di destra non sta mangiando */ {
9                  state[i] = EATING; /* mi metto a mangiare */

```

```

10         self[i].signal();
11     }
12 }
13
14 void pickup(int i){ /* acquire */
15     state[i] = HUNGRY;
16     test(i); /* se posso, mi metto a mangiare */
17     if (state[i] != EATING) /* se non sto mangiando */
18         self[i].wait(); /* mi metto in attesa*/
19 }
20
21 void putdown(int i){ /* release */
22     state[i] = THINKING; /* mi rimetto a pensare */
23     /* controllo che il mio vicino sinistro e destro non siano in attesa
24     */
25     test((i + 4) % 5);
26     test((i + 1) % 5);
27 }
28
29 initialization_code(){ /* all'inizio tutti i filosofi stanno pensando
30 */
31     for (int i = 0; i < 5; i++)
32         state[i] = THINKING;
33 }
34 }
```

Dopo aver implementato questa struttura astratta è semplicemente necessario utilizzare le due operazione `pickup()` e `putdown()` al fine di fare in modo che tutti i filosofi mangino sincronizzati. Ad ogni modo, anche in una situazione così raffinata la **starvation** è possibile, nel caso in cui un filosofo si metta a mangiare e non smetta più.

1 | DiningPhilosophers.pickup(i);

```
2 | /* SEZIONE CRITICA (mangio) */  
3 | DiningPhilosophers.putdown(i);
```

6 DEADLOCKS

In questo capitolo cercheremo di approfondire i deadlocks: capiremo sotto quali condizioni si giungerà ad una soluzione di deadlock e i diversi algoritmi per la gestione dei deadlocks.

Modello di sistema

Prima di discutere effettivamente dei deadlock è importante fare un piccola premessa. In questo capitolo avremo a che fare con sistemi che possono essere composti da m tipi di risorse generiche: queste potrebbero essere dei dispositivi I/O, dati su disco, delle allocazioni in memoria, la CPU oppure dei semafori generici. Queste risorse vengono definite con la notazione R_m . Inoltre, ogni tipo di risorsa può essere **singolo**, come un slot del DVD sul computer, oppure **multiplo** come la presenza di diverse porte USB oppure molti buffer in memoria.

6.1 Nozioni fondamentali

Ripassiamo innanzitutto il concetto di deadlock. Prendiamo il caso in cui due semafori S_1 ed S_2 , entrambi inizializzati a 1, siano condivisi da due thread T_1 e T_2 . Osservando lo pseudo-codice seguente, si ha una classica situazione di deadlock. Se T_1 e T_2 sono eseguiti contemporaneamente, T_1 occupa subito S_1 e T_2 occupa subito S_2 . Dopo di che T_1 aspetta S_2 che però è occupato da T_1 il quale è in

attesa di S_1 , occupato da T_1 . Siamo quindi in una situazione di stallo dove entrambi i thread stanno aspettando l'altro e nessuno dei due termina l'esecuzione

Codice 6.1: Classica situazione di deadlock

```
1  /* T1 */
2  wait(S1); /* occupa S1 */
3  wait(S2); /* aspetta per S2 */
4
5  /* T2 */
6  wait(S2); /* occupa S2 */
7  wait(S1); /* attende S1 */
```

6.1.1 Caratterizzazione

Possiamo osservare che si devono verificare 4 condizioni affinché si arrivi ad una situazione di deadlock.

1. **Mutua esclusione:** banalmente, ci deve essere almeno una risorsa che sia utilizzata da un thread alla volta;
2. **Hold and wait:** deve esistere almeno un thread che sta cercando di accedere alla risorsa che è già occupata e, di conseguenza, deve attendere;
3. **No preemption:** una risorsa può essere rilasciata volontariamente dei thread. Le risorse non possono quindi essere "rubate" da altri thread e, di conseguenza, un thread non può essere spezzato durante la sua fase critica;
4. **Attesa circolare:** è il requisito più importante, deve verificarsi una situazione come quella dei cinque filosofi (paragrafo 5.6.3), ovvero che un thread attende la conclusione di un altro thread che ... che aspetta la conclusione del thread iniziale.

Se almeno una di queste quattro condizioni **non è verificata** allora non si è in una situazione di deadlock.

6.1.2 Grafo risorsa-allocazione

Uno strumento molto importante, utile per visionare i diversi casi di deadlock è il grafo risorsa-allocazione. Questo grafo è composto, come ogni grafo, da vertici e archi. I vertici si suddividono in due tipi: il primo tipo rappresenta la risorsa R_m mentre il secondo tipo indica il thread T_n . Il grafo è **diretto**, ciò significa che gli archi hanno una direzione. In particolare $R \rightarrow T$ indica che il thread T detiene/occupa la risorsa R , mentre $T \rightarrow R$ segnala che il thread T è in attesa della risorsa R . Inoltre, si osserva che generalmente i vertici che rappresentano una risorsa multipla contengono tanti pallini (\bullet) quante sono le istanze di quella risorsa disponibili (come le 5 porte USB).

Prendiamo come esempio la figura 6.1. Possiamo notare che rispetta tutte "regole" del grafo risorsa-allocazione. In particolare possiamo affermare che R_1 ed R_3 sono risorse singole mentre R_2 ed R_4 sono risorse multiple (con, rispettivamente, 2 e 4 istanze). Inoltre possiamo dire che ci troviamo in

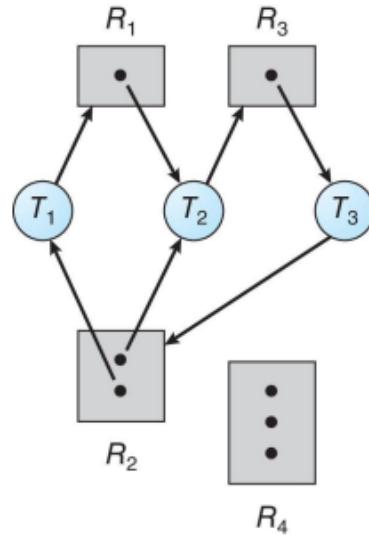


Figura 6.1: Rappresentazione grafica di un grafo che presenta una situazione di deadlock.

una situazione di *deadlock*. Perché? Partiamo da R_2 , questa è occupata da T_1 e T_2 ; T_1 è in attesa di R_1 la quale è occupata da R_2 che, a sua volta, sta aspettando che R_3 venga liberata da T_3 il quale sta attendendo la liberazione due R_2 , occupata per l'appunto da T_1 e T_2 . È proprio una situazione di deadlock: a primo impatto possiamo dire ciò perché è presente un **ciclo**. Osservando però la figura 6.2, capiamo che il fatto che ci sia un ciclo non implica che ci sia per forza una situazione di deadlock. Se

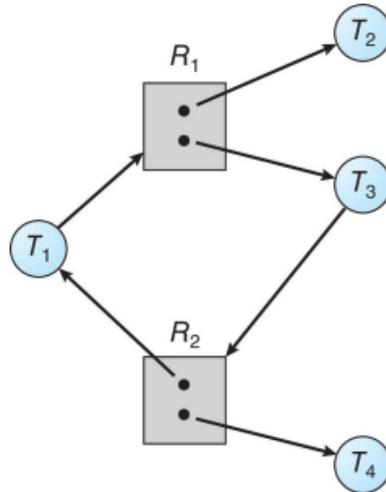


Figura 6.2: Rappresentazione grafica di un grafo senza deadlock.

osserviamo infatti la figura 6.2 riconosciamo che prima o poi il thread T_2 oppure T_4 rilascerà la risorsa e di conseguenza le richiesta degli altri thread saranno soddisfatte.

Riassumiamo ora queste informazioni nel seguente diagramma:

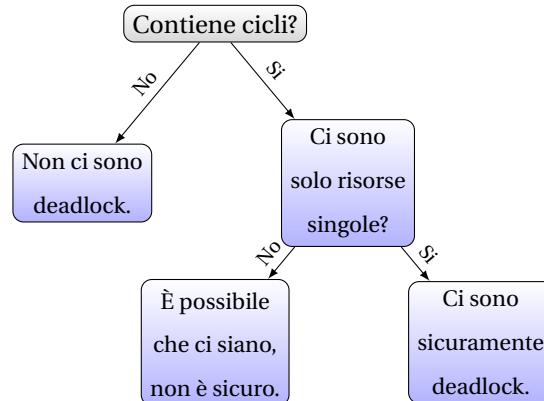


Figura 6.3: Piccolo *decision tree* utile per capire se è presente un deadlock oppure no.

Livelock

Oltre al deadlock esiste anche una seconda situazione di stallo chiamata *livelock*. In questo caso i due processi non sono completamente in una situazione di stallo ma non riescono a terminare la loro esecuzione. Possiamo allegoricamente spiegare questa situazione con un corridoio dove ci sono due persone, A e B, che vanno in direzione opposta. Ad un certo punto, nel tempo, si incontrano, allora, per non scontrarsi, A si sposta, ma contemporaneamente lo fa anche B. A questo punto A si sposta di nuovo, ma lo fa anche B. I due processi quindi si ostacolano l'un l'altro non riuscendo a terminare.

Prevention

Come vedremo in questo capitolo, ci sono diversi metodi finalizzati alla gestione dei deadlock. Il primo metodo che andiamo a discutere è quello di "prevenzione". In essenza, questo metodo si occupa di **invalidare** una delle quattro **condizioni** necessarie che generano una situazione di deadlock (vedi paragrafo 6.1.1). Anticipiamo già che questo approccio non è assolutamente efficiente e non è utilizzato.

1. La prima condizione, ovvero quella di mutua esclusione raramente può essere invalidata. Proprio perché la sincronizzazione tra processi nasce proprio per far sì che solo un processo abbia accesso ad una risorsa, la mutua esclusione è fondamentale. Altrimenti si incombe in situazioni dove più processi sono, per esempio, in scrittura, sullo stesso file.
2. Per quanto riguarda l'*hold and wait*, in questo caso si cerca di fare in modo che un processo che detiene una risorsa non ne può richiedere un'altra (si rimuove l'*hold* dalla definizione). Di conseguenza si fa in modo che un thread, prima di essere eseguito, faccia una richiesta contemporanea a tutte le risorse di cui ha bisogno lasciano tutti gli altri processi ad aspettare. Si entra in situazioni di inefficienza e, talvolta, c'è il rischio che un processo vada in **starvation**.
3. Affine di invalidare il punto tre si aggiunge la *preemption*, ovvero che i processi possono essere interrotti al fine di lasciare la risorsa ad altri processi. Anche in questo caso però si possono verificare situazioni assurde: si pensi all'utilizzo di una stampante, un thread non può fermare la stampa e iniziare a stampare. Si ha anche in questo caso una soluzione poco ottimale.

4. Solo nel quarto punto, dove si cerca di evitare l'attesa circolare si ha effettivamente una soluzione sensata. Nella pratica si cerca di eliminare la circolarità imponendo dei vincoli particolare ai processi, per esempio un **ordine** di esecuzione.

6.2 Avoidance

Il secondo approccio che andremo a discutere più a fondo è una sorta di miglioramento della prevenzione. In questo caso si cerca proprio di **evitare** di ricadere in situazioni che possono portare al deadlock, è una prevenzione molto più *strict*.

Affinché questo approccio funzioni però deve essere fornita un'informazione importante da parte del processo, ovvero il numero massimo di risorse per tipo che il processo necessita durante la sua esecuzione. Per esempio, 6 risorse di tipo A, 5 risorse di tipo B e 2 risorse di tipo C.

6.2.1 Safe state

Il safe state, ovvero lo **stato sicuro**, è una situazione in cui esiste un ordine di esecuzione di thread secondo il quale tutti i thread possono essere completati uno dopo l'altro. Quando un processo fa la richiesta per le risorse, prima di concederle, l'algoritmo (che vedremo in seguito) controlla se dopo la cessione delle risorse si è in una situazione ancora safe. Se sì, allora le risorse vengono occupate dal processo, altrimenti il processo deve aspettare il termine di alcuni thread.

Nella primo grafo della figura 6.4 si è in una condizione non sicura in quanto non esiste una sequenza di esecuzione dei processi dove non si verifichi una situazione di deadlock. Nel secondo grafo invece si è in una situazione safe in quanto esiste una sequenza dove tutti i processi possono essere eseguiti senza entrare in una condizione di deadlock: $T_1 \rightarrow T_4 \rightarrow T_3 \rightarrow T_5 \rightarrow T_2$ infatti è una sequenza che permette il completamento dell'esecuzione dei threads. Anche in questo caso possiamo riassumere tutto in un piccolo *decision tree*:

Ricordiamo infine che gli algoritmi che vedremo in questo paragrafo al fine di prevenire i deadlock **evitano** di entrare in **unsafe state**, rimanendo quindi in safe state.

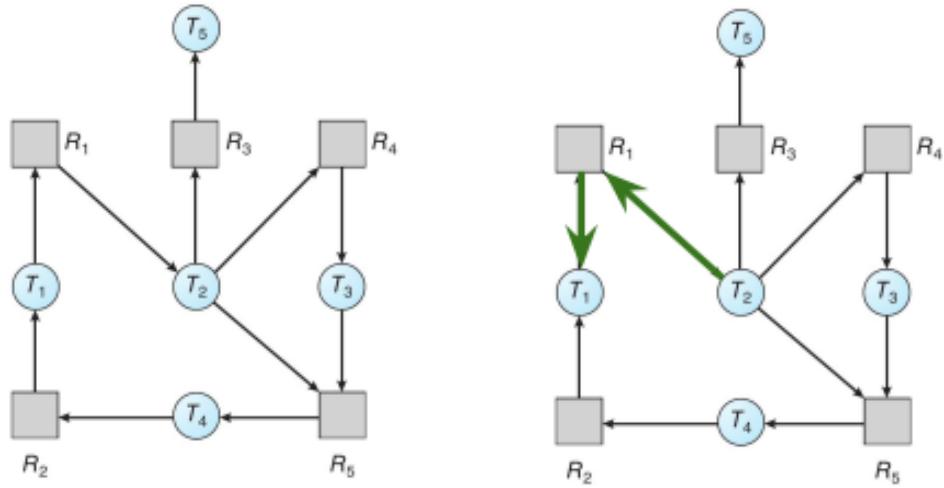
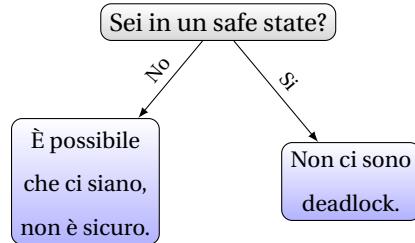


Figura 6.4: Differenza tra situazione unsafe e safe.



6.2.2 Algoritmo sul grafo risorsa-allocazione

Iniziamo con un algoritmo che funziona solo nel caso di **risorse singole**, ovvero nel caso in cui si ha solo un pallino (\bullet) nel grafo, e quindi ogni tipo di risorsa ha una sola istanza. Introduciamo ora la nozione di **claim edge** ovvero un arco tratteggiato che indica che il processo T_i può richiedere la risorsa R_j . Questi archi diventano **request edge** nel momento in cui si effettua la richiesta e poi si trasformano a loro volta in **assignments edge** nel momento in cui stanno utilizzando la risorsa. Osservando il grafo in figura 6.5 notiamo che $T_1 \rightarrow R_2$ è un claim edge, $T_2 \rightarrow R_1$ è un request edge e $R_1 \rightarrow T_1$ è un assignment edge.

Nella situazione presentata in figura 6.5, si rischia di essere in una situazione *unsafe*. Questo perché se R_2 è assegnata al thread T_2 nel caso in cui il thread T_1 effettui la richiesta per R_2 si è in una situazione di deadlock. Possiamo infatti osservare che si ha la presenza di un ciclo e che tutte le risorse abbiano

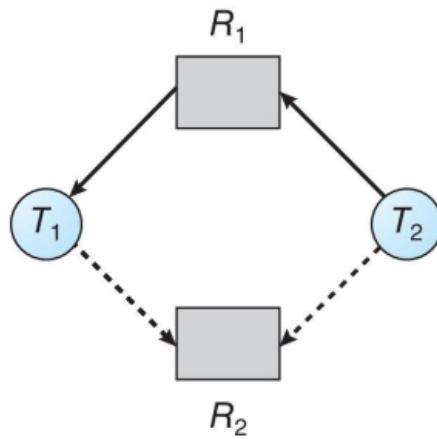


Figura 6.5: Un grafo composto da tutti e tre i tipi di archi.

un'istanza, facendo riferimento allo schema in figura 6.3, si ha che si è sicuramente in una situazione di deadlock. Ponendo infatti che R_2 venga assegnata a T_2 , nel momento in cui T_1 effettui una richiesta, l'algoritmo la nega, in modo tale da non essere in una situazione di deadlock.

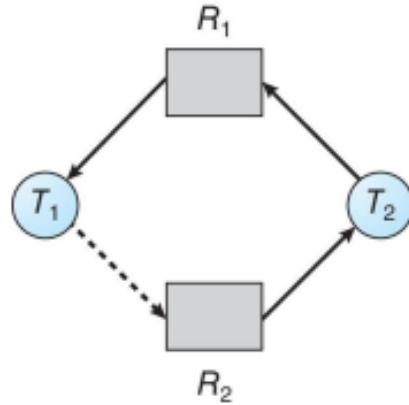


Figura 6.6: L'algoritmo nega la richiesta di T_1 per R_2

6.2.3 Algoritmo del banchiere

Passiamo ora ad una situazione più reale, quella in cui esistono più risorse con più istanze ciascuna. L'algoritmo a cui si fa affidamento è detto *Banker's algorithm*, ovvero l'algoritmo del banchiere.

Innanzitutto, definiamo con n il numero di processi ed m il numero di tipi di risorse; definiamo poi tre matrici e un'array, strutture sul quale l'algoritmo si appoggia:

- L'array **available** (disponibili) che indica quante istanze di risorse sono disponibili per ogni tipo.

Per esempio, la risorsa j ha a disposizione 3 istanze, e questo $\forall j \in [0, m]$.

- La prima matrice, **max** che indica il massimo numero di istanze della risorsa j esima che il processo i esimo ha al più bisogno ($\forall i \in [0, n]$).
- La seconda matrice **allocated** (allocato), che indica quante istanze della risorsa j esima il processo i esimo ha già allocato.
- La terza matrice, **need** (bisogno), che non è altro che la differenza tra *max* e *allocated*; questa matrice indica di quante altre istanze della risorsa j esima il processo i esimo potrebbe, al più, aver bisogno.

Per capire l'algoritmo al meglio si può tranquillamente trascurare lo pseudo-codice e partire subito con un **esempio**. Abbiamo $n = 5$ thread, da T_0 a T_4 , e 3 tipi di risorse: A, con 10 istanze, B, con 5 e C, che ha a disposizione 7 istanze. Al tempo iniziale t_0 , la situazione è la seguente:

<u>Threads</u>	<u>Allocated</u>			<u>Max</u>		
	A	B	C	A	B	C
T_0	0	1	0	7	5	3
T_1	2	0	0	3	2	2
T_2	3	0	2	9	0	2
T_3	2	1	1	2	2	2
T_4	0	0	1	4	3	3

Un esempio di lettura della tabella è il seguente: T_0 detiene un'istanza di tipo B e zero istanze di tipo A e C. Può richiedere un massimo di 7 istanze di tipo A, 5 istanze di tipo B e 3 istanze di tipo C.

Con questi dati possiamo ora calcolare le due strutture che ci mancano: *Available* e *Need*, che rispettivamente indicano le risorse che sono disponibili (per ogni tipo) e il numero di risorse che un determinato thread può richiedere. Osserviamo che per calcolare il vettore di risorse disponibili è necessario sottrarre alle istanze iniziali la somma delle risorse occupate dai threads. Per esempio, nel caso delle istanze disponibili della risorsa A abbiamo che sono $10 - (2 + 3 + 2) = 3$, occupate rispettivamente da: T_1 , T_2 e T_3 . per calcolare invece la matrice *Need*, per ogni cella della matrice si sottrarre il valore di *Max* al valore di *Allocated* nella cella $[i, j]$.

	<u>Need</u>		
	A	B	C
<u>Available</u>	7	4	3
A	1	2	2
3	6	0	0
	0	1	1
	4	3	1

Abbiamo ora tutte le risorse necessarie per iniziare a eseguire l'algoritmo. Iniziamo scorrendo la matrice *Need* e comparandola con il vettore *Available*. Partiamo dal thread T_0 : con le risorse disponibili $[3, 3, 2]$, il thread ha la possibilità di terminare? No, perché per terminare necessita, al più, di 7 istanze di A, 4 istanze di B e 3 istanze di C. Passiamo ora al secondo thread: T_1 può terminare con le risorse disponibili? Sì, perché tutte le risorse di cui ha bisogno sono \leq rispetto alle risorse disponibili. A questo punto allora si spostano le risorse necessarie a T_1 per terminare, ovvero $[1, 2, 2]$:

<u>Threads</u>	<u>Allocated</u>			<u>Max</u>			<u>Need</u>			<u>Available</u>
	A	B	C	A	B	C	A	B	C	
T_0	0	1	0	7	5	3	7	4	3	
T_1	3	2	2	3	2	2	0	0	0	A B C
T_2	3	0	2	9	0	2	6	0	0	2 1 0
T_3	2	1	1	2	2	2	0	1	1	
T_4	0	0	1	4	3	3	4	3	1	

Una volta terminato il thread T_1 , le risorse vengono rilasciate e il vettore *Available* va aggiornato, sommando alle risorse disponibili le risorse che sono appena state rilasciate da T_1 .

Available

A	B	C
5	3	2

A questo punto ci sono abbastanza risorse per terminare un altro processo come, per esempio T_3 oppure T_4 . Così facendo, mano che i thread vengono eseguiti e terminati vengono rilasciate abbastanza risorse per completare quelli più onerosi. Una sequenza esempio dell'ordine con il quale i thread sono completati può essere: $\{T_1, T_3, T_4, T_2, T_0\}$.

Può succedere che durante l'esecuzione dei thread possano arrivare altre richieste che devono essere accettate o meno. Per esempio se prima dell'esecuzione T_4 avesse fatto una richiesta $[3, 3, 0]$, la richiesta sarebbe stata rifiutata dato che le risorse disponibili erano minori rispetto alla richiesta.

6.3 Detection

L'ultimo approccio che andiamo a discutere è il rilevamento del deadlock: si va alla ricerca di tale e, una volta trovato, si passa al salvataggio, alla **recovery**. Anche in questo caso, come per l'approccio di *avoidance* (paragrafo 6.2) sono presenti due algoritmi, uno studiato per avere solo un'istanza per risorsa e il secondo che copre anche il caso in cui ci siano più istanze per risorse.

6.3.1 Istanza singola

In questo primo caso si fa riferimento ad un secondo grafo particolare chiamato **wait-for graph**. È una modifica del *resource-allocation graph* dove si evidenzia il fatto che un thread sta aspettando il termine dell'esecuzione di un altro thread: non sono infatti presenti vertici che rappresentano risorse a ci sono solo nodi che indicano i thread. In figura 6.7 si osserva la traduzione di un grafo risorsa allocazioni in grafo *wait-for*. Si nota che nel secondo grafo si accentua ancora di più la presenza di cicli e, dato che si tratta di risorse con istanza singola, la presenza di un ciclo ha l'immediata conseguenza di generare una situazione di deadlock.

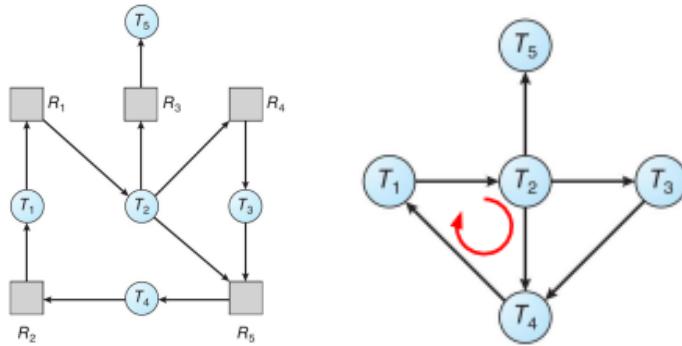


Figura 6.7: Un grafo *resource-allocation* e il suo rispettivo *wait-for graph*

L'algoritmo di cui stiamo parlando infatti, per rilevare un deadlock, si limita a creare un *wait-for graph* e controllare la presenza di cicli¹. Una volta trovato il ciclo si è sicuri di aver trovato il deadlock e si passa alla *deadlock recovery* (vedi paragrafo 6.3.3). Possiamo quindi notare che nella figura 6.7 si è in una condizione di deadlock dato che il thread T_1 sta attendendo il termine di T_2 che è in attesa di T_4 che a sua volta sta aspettando T_1 . Una situazione analoga è presente anche con i thread T_1, T_2, T_3, T_4 .

6.3.2 Istanze multiple

Nel caso invece delle istanze multiple, anche in questo algoritmo si fa affidamento a diverse strutture dati, proprio come nell'algoritmo del banchiere. Avremo quindi a che vedere con n processi che hanno a che fare con m tipi di risorse che interagiranno con due matrici e un vettore.

- ◊ Il vettore *Available*, proprio come nell'algoritmo del banchiere, tiene traccia di quante istanze sono libere per ogni tipo di risorsa.
- ◊ La prima matrice *Allocated* memorizza il numero di risorse che ogni processo detiene.
- ◊ La seconda matrice, **Request**, indica quante risorse il thread sta richiedendo al fine di terminare.

Come nel caso dell'algoritmo del banchiere, non ci soffermiamo sullo pseudo-codice ma partiamo subito con un esempio. Poniamo di avere cinque threads, da T_0 a T_4 e di avere anche in questo caso tre tipi di risorse: A, con 7 istanze, B con 2 istanze e C con 6 istanze. Al tempo iniziale t_0 ci troviamo nella seguente situazione:

¹Da studente che ha seguito il corso "Dati e Algoritmi" penso che si utilizzino algoritmi basati su BFS oppure DFS.

<u>Threads</u>	<u>Allocated</u>			<u>Request</u>			<u>Avaliable</u>
	A	B	C	A	B	C	
T_0	0	1	0	0	0	0	
T_1	2	0	0	2	0	2	A B C
T_2	3	0	3	0	0	0	0 0 0
T_3	2	1	1	1	0	0	
T_4	0	0	2	0	0	2	

Dalla tabella *Request* possiamo notare che sia T_0 che T_2 possono completare senza attendere altri processi dato che non richiedono nulla. Una volta terminati T_0 e T_2 tutti gli altri thread possono man mano terminare. Non siamo quindi in una situazione di deadlock.

Mettiamoci però nel caso in cui T_2 debba richiedere una risorsa di tipo C. In questo caso le risorse che

<u>Threads</u>	<u>Allocated</u>			<u>Request</u>			<u>Avaliable</u>
	A	B	C	A	B	C	
T_0	0	1	0	0	0	0	
T_1	2	0	0	2	0	2	A B C
T_2	3	0	3	0	0	1	0 0 0
T_3	2	1	1	1	0	0	
T_4	0	0	2	0	0	2	

vengono rilasciate al termine di T_0 non sono necessarie per terminare gli altri thread e ci troveremo quindi in una situazione di deadlock.

6.3.3 Deadlock recovery

Come dobbiamo comportarci nei momenti in cui rileviamo un deadlock? La soluzione più brutale è quella di terminare tutti i processi che si trovano all'interno del deadlock (come fare un `CTRL + C` ad ogni processo). Esiste però un altro modo un po' più elegante che fa comunque affidamento alla **terminazione dei processi**: si cerca infatti di terminare ad uno ad uno tutti i thread che compongono il ciclo. L'ordine di terminazione può essere scelto in base alla priorità, in base alle risorse che il thread ha utilizzato o altro.

Una seconda strada invece si basa sul **resource preemption**, ovvero sul rubare delle risorse che sono detenute da un altro processo o thread. In questo caso si sceglie una vittima a cui rubare le risorse e si cerca di far terminare un thread. Si ricorda che anche in questo caso si può arrivare ad una situazione di *starvation* se lo stesso thread è sempre preso come vittima.

7

MEMORIA PRINCIPALE

Fino ad ora abbiamo discusso di come un sistema operativo gestisce lo scheduling (4), la sincronizzazione (5) e i deadlock (6) sorvolando sempre sulla locazione effettiva su cui questi processi vengono mantenuti, ovvero la memoria. Osserviamo sin dall'inizio che con memoria intendiamo la memoria RAM del calcolatore, le memorie secondarie molto capienti sono i dischi e le memorie di massa, che verranno discusse approfonditamente nel capitolo 9. In particolare in questo capitolo ci occuperemo di dare un'introduzione ai concetti generali, passando poi a discutere il metodo più semplice per risolvere l'utilizzo della memoria, ovvero l'allocazione continua per poi discutere un tecnica più avanzata che tutt'ora si trova nei sistemi operativi moderni, ovvero il *paging*. Scopriremo che saranno necessarie delle strutture ausiliarie, come la tabella delle pagine, e ne studieremo i diversi metodi di implementazione. Infine ci dedicheremo al concetto di *swapping* e a modelli di allocazioni alternativi alla paginazione.

7.1 Introduzione

Come sappiamo, il programma, una volta che viene eseguito, deve essere caricato dal disco in memoria. Solo una volta che è stato caricato in memoria può essere eseguito; questo perché la CPU non ha accesso diretto alla memoria di massa. Sappiamo inoltre che l'accesso della CPU è effettuato seguendo 3 tipi di memorie (figura 7.1):

- ◊ Registri, che hanno un accesso rapidissimo;

- ◊ RAM, che ha una velocità sicuramente ridotta rispetto ai registri;
- ◊ Cache, che è una piccola memoria che fa da intermediario tra i registri e la RAM ed ha un tempo di accesso molto più veloce della memoria.

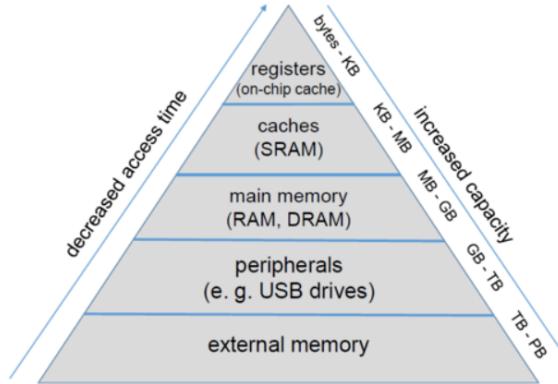


Figura 7.1: La gerarchie delle memorie nel calcolatore.

7.1.1 Protezione

In generale, se vogliamo avere molti processi che vengono caricati contemporaneamente nell'area della memoria e che vengono eseguiti in modo concorrente, è importante che ci sia una protezione dei processi fornita dal sistema operativo al fine di garantire che ogni processo abbia il suo spazio di memoria senza che vada a collidere con un altro processo. La soluzione più banale, rappresentata in figura 7.2, è quella di utilizzare due registri che tengono traccia di due valori: l'indirizzo **base**, ovvero l'indirizzo in memoria iniziale da cui il processo parte, e il valore **limite** (*offset*) che indica la massima espansione del processo in memoria. In altre parole, il processo in memoria non può superare l'indirizzo "base + limite" (più avanti in questo capitolo vedremo soluzioni più complesse e raffinate). In questo caso l'unica preoccupazione del sistema operativo è controllare, ogni volta che un indirizzo viene generato dalla CPU durante l'esecuzione del programma, se questo indirizzo è compreso tra la base e l'offset fornito dai due registri. Nel caso in cui l'indirizzo è compreso allora la locazione di memoria può essere acceduta. Osserviamo che la figura 7.3 rappresenta un piccolo schema dove viene illustrata questa verifica da parte del sistema operativo.

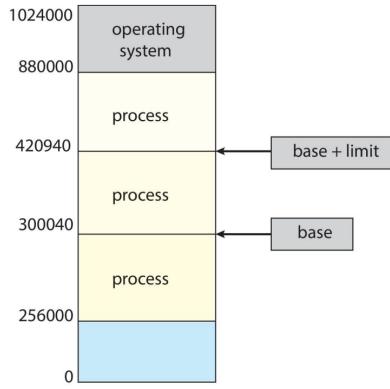


Figura 7.2: Limite inferiore e limite superiore dello spazio del processo in memoria.

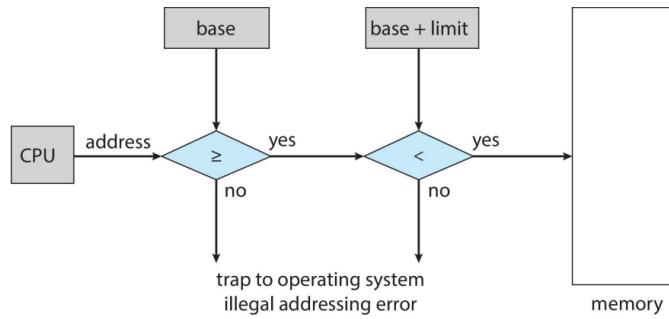


Figura 7.3: Procedura di controllo di un indirizzo mediante il SO.

7.1.2 Binding

Quando creiamo un processo, di default parte dall'allocazione 0000. Naturalmente non è possibile far partire tutti i processi dalla stessa allocazione altrimenti si genererebbero conflitti che causerebbero problemi non indifferenti. Inoltre ogni processo che viene eseguito, anche se il suo indirizzo in memoria è diverso 0000, viene eseguito come se la sua prima cella disponibile fosse la 0000. Come è possibile? Il procedimento è chiamato **binding** e mappa due tipi di indirizzi: gli indirizzi **assoluti**, che sono gli effettivi indirizzi nella memoria, e gli indirizzi **rilocabili** che sono gli indirizzi relativi al programma in esecuzione. Per chiarire le idee, facciamo un esempio: poniamo di avere un processo che parte dall'indirizzo in memoria 31400 (indirizzo assoluto) e che stia eseguendo una linea di codice che secondo il processo è all'indirizzo 15. In realtà il programma non sta eseguendo l'indirizzo 15 ma

sta eseguendo l'indirizzo assoluto 31415, che è la somma tra l'indirizzo fisico di inizio e l'indirizzo logico (31400 + 15). Questa somma, è infatti detta binding e permette di collegare gli indirizzi logici del processo agli indirizzi fisici della memoria.

Il binding però può avvenire in diversi momenti, come è possibile anche osservare dalla figura 7.4. Possiamo distinguere principalmente 3 momenti:

1. **Tempo di compilazione:** in questo caso, il compilatore, compilando il codice automaticamente converte gli indirizzi rilocabili in indirizzi assoluti.
2. **Tempo di caricamento:** nel momento di *linking* del programma e quindi si ha un eseguibile che ha ancora gli indirizzi temporanei ma che poi, una volta caricato, verranno sostituiti con gli indirizzi assoluti.
3. **Tempo di esecuzione:** è infine possibile trasformare gli indirizzi rilocabili in indirizzi assoluti durante l'esecuzione del programma. Questo è il metodo tipicamente utilizzato dai sistemi operativi moderni, attraverso tecniche che approfondiremo in questo capitolo.

È quindi fare un'importante distinzione tra due tipi di indirizzi. Lo spazio di indirizzi che è visto dal programma (e quindi è indipendente dalla macchina in cui risiede) è chiamato spazio degli indirizzi **logici** (o **virtuali**, come vedremo nel capitolo 8). D'altro canto, lo spazio degli indirizzi in memoria, ovvero gli indirizzi assoluti, è chiamato spazio degli indirizzi **fisici**.

7.1.3 Memory-Management Unit (MMU)

Parte dei meccanismi utilizzati per gestire la memoria e il binding non sono lasciati solamente al sistema operativo ma sono un ibrido tra hardware e software. Questo hardware è detto *Memory-Management Unit (MMU)* ed aiuta a mappare gli indirizzi logici in indirizzi fisici. Questo modulo era originariamente un chip a parte, tra la CPU e la memoria, con l'avanzare del tempo l'MMU è una parte integrata nella CPU stessa.

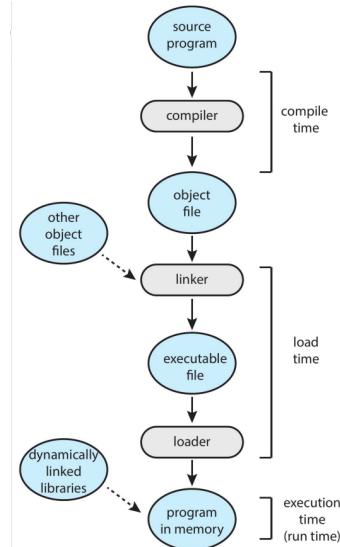


Figura 7.4: I tre tempi di binding.

Una delle funzionalità dell'MMU è quella di fornire il controllo sull'indirizzo del processo nella memoria. È la stessa funzionalità che implementava il sistema operativo (vedi figura 7.3) solo che in questo caso il controllo è effettuato autonomamente dell'MMU attraverso un particolare registro chiamato **relocation register**. Il processo è rappresentato nell'illustrazione 7.5.

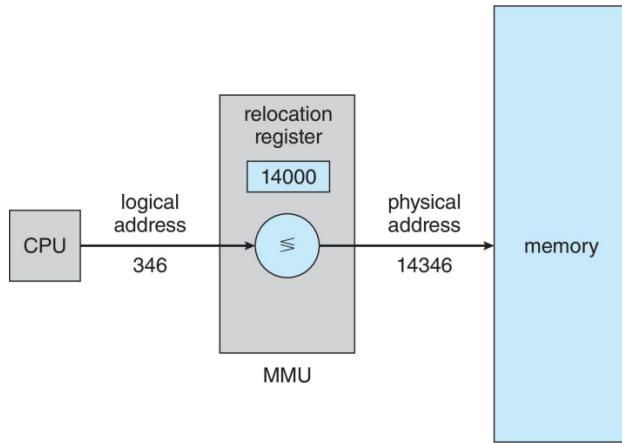


Figura 7.5: Procedura di controllo di un indirizzo mediante la MMU.

7.1.4 Caricamento e collegamento dinamico

Uno dei concetti fondamentali nell'uso della memoria è il **dynamic loading**. Ciò significa che dal punto di vista del programma non è possibile mantenere tutte le funzioni necessarie del programma sempre in memoria, ma mantenere solo un sottoinsieme che sono al momento necessarie. Nel momento in cui una nuova routine è necessaria la si va a caricare dalla memoria.

Un secondo approccio è quello del *linking*. In particolare il collegamento può essere fatto in due modi: statico e dinamico. Lo **static linking** avviene quando colleghiamo le librerie di sistema al programma e una volta compilato il programma le librerie sono immediatamente integrate nel file eseguibile. A questo si contrappone il **dynamic linking** dove il codice delle librerie non viene caricato nel file eseguibile ma le librerie sono collegate solo durante l'esecuzione del programma. Sono infatti presenti delle **shared libraries**, come le `dll` di Windows, che vengono caricate in memoria e condivise a tutti i processi che le necessita e, un volta che nessun processo le utilizza più, queste vengono rimosse dalla memoria. Questo è molto più conveniente rispetto al collegamento statico dove in memoria ci possono essere potenzialmente diversi programmi che hanno una copia della libreria nel codice.

Questo significa che la stessa libreria occupa molto più spazio in memoria perché è copiata da diversi processi. Se invece ci fosse stata una libreria dinamica, tutti i processi avrebbero usufruito il codice della stessa in modo tale da risparmiare dello spazio prezioso in memoria.

7.2 Primi modelli di allocazione

Passiamo ora a vedere la storia e i concetti della paginazione della memoria per poi arrivare a discutere dei metodi moderni e tuttora utilizzati.

7.2.1 Allocazione contigua

La prima soluzione che è stata pensata è l'allocazione contigua (figura 7.6) in memoria: ciò significa



Figura 7.6: Allocazione contigua dei processi in memoria.

avere una parte dedicata al sistema operativo (kernel) e poi, in modo contiguo, inizia il codice per i processi.

Il modo più semplice per poter implementare questo meccanismo è attraverso il *limit register* e il *relocation register* che abbiamo visto poco fa. Questo infatti permette di proteggere i processi tra di loro, evitando quindi che si sovrappongano in memoria. Osservando infatti l'illustrazione 7.7 possiamo notare che dalla CPU viene preso l'indirizzo logico (ovvero l'indirizzo relativo per il processo), viene controllato se "sfiora" il limite massimo in memoria; una volta che è stato appurato che l'*upper bound* non è stato superato, il relocation register procede con il *binding*.

7.2.2 Allocazione a partizione fissa

Un'altra soluzione un po' più raffinata, è l'allocazione a partizione fisso. Questo tipo di allocazione consiste nel dividere la memoria in un insieme di partizioni di dimensione non variabile e andare

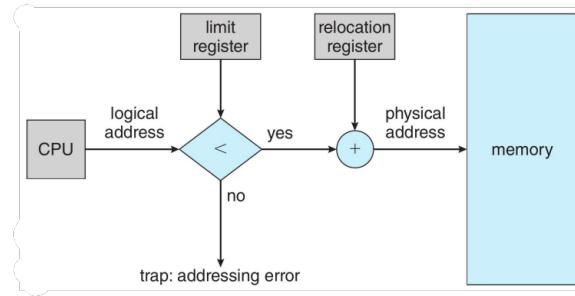


Figura 7.7: I registri utilizzati per il binding tra logico e fisico.

ad allocare i processi, non in modo contiguo, ma in una partizione la cui dimensione è adeguata e consona alla dimensione del processo. Osservando la figura 7.8 notiamo che, innanzitutto le partizioni

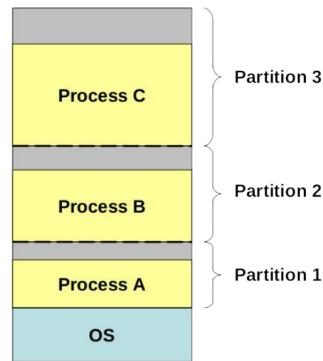


Figura 7.8: La memoria divisa in partizioni fisse per i processi.

possono essere di dimensione diversa (l'importante è che non cambi nel tempo) al fine di gestire processi si dimensione diversa. Notiamo già i possibili problemi di questo tipo di modello, primo tra tutti è lo spazio in memoria inutilizzato grigio alla fine di ogni partizione (vedi il paragrafo 7.2.4: frammentazione interna). L'idea però della partizioni fisse è un concetto che, come vedremo verrà ripreso nella paginazione (paragrafo 7.3), presente nei sistemi operativi moderni.

7.2.3 Allocazione a partizione variabile

In alternativa alla partizione fissa, possiamo avere un tipo di allocazione in cui la dimensione della partizione varia a seconda del processo che fa richiesta. Osservando l'illustrazione 7.9 notiamo che il processo 8, una volta che è terminato lascia un buco ad un secondo processo, il 9 che però occupa

meno spazio, di conseguenza la dimensione della partizione è variata. Anche in questo caso, come

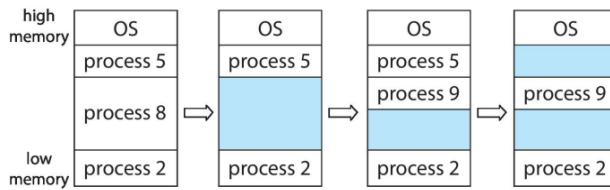


Figura 7.9: Rappresentazione di una partizione variabile in memoria.

nella partizione fissa, si può incombere in una frammentazione interna (paragrafo 7.2.4) dato che una volta che il processo 9 ha occupato lo spazio rilasciato dal processo 8, è rimasto comunque un **bucco** in memoria, che non sempre può essere occupato da un altro processo in quanto, magari, non è presente abbastanza spazio contiguo.

Scelta della cella da allocare

Poniamo di essere in una situazione in cui abbiamo alcuni processi sparsi in memoria ed è necessario inserire un nuovo processo, come in figura 7.10. Quale posto sceglio?

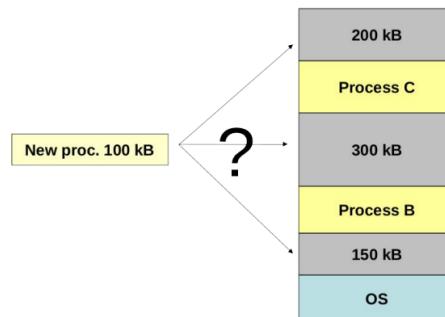


Figura 7.10: In quale allocazione di memoria verrà inserito il nuovo processo?

La scelta può essere effettuata secondo 3 criteri diversi:

- ◊ *First-fit*: il primo posto disponibile viene occupato dal nuovo processo. Per esempio, se abbiamo la memoria ad allocazione variabile e all'interno di essa abbiamo un processo B e un processo C, si va a scorrere la memoria e si inserisce il nuovo processo nel primo *hole* disponibile;

- ◊ *Best-fit*: in questo caso si cerca di allocare il nuovo processo nella zona il più si avvicina alla dimensione del processo. Di conseguenza, al fine di allocare il processo in memoria è prima necessario scorrerla tutta con una ricerca **lineare**;
- ◊ *Worst-fit*: infine, con questa tecnica, si sceglie lo spazio con la dimensione peggiore possibile, ovvero la dimensione più grande possibile; ciò significa che il buco lasciato in memoria sarà il più grande possibile. Anche in questo caso è necessario fare una ricerca **lineare**.

7.2.4 Problema della frammentazione

Abbiamo ormai constatato che quando si allocano nuovi processi in memoria, questi lasciano in memoria dei buchi che non sempre possono essere riempiti. In particolare, nel caso della partizione fissa (7.2.2) questo tipo di problema viene chiamato **frammentazione interna**, questo perché è interna alla partizione in quanto il processo allocato non riesce a riempire pienamente lo spazio della partizione (osservare la figura 7.8). Questo rappresenta ovviamente un problema perché può essere che in memoria ci sia in totale dello spazio disponibile (ovvero la somma di tutti gli spazi grigi in figura) ma essendo non contiguo non permette un ulteriore inserimento del processo in memoria.

Alla frammentazione interna si contrappone la **frammentazione esterna**, che si verifica nel caso dell'allocazione a partizione variabile (7.2.3). Facendo appunto riferimento alla figura 7.9, può essere che una volta inserito il processo 9 all'interno della memoria si siano generati due buchi che non sono abbastanza grandi al fine di contenere un altro processo. Si è fatto uno studio dove si è notato che con l'allocazione a partizione variabile e il **first-fit**, ogni N blocchi in memoria, se ne perdono la metà a causa della frammentazione esterna: di conseguenza all'incirca 1/3 della memoria è inutilizzato (*50-percent rule*).

Una soluzione al problema della frammentazione esterna è il **compacting** (compattazione), illustrata in figura 7.11. Poniamo che il processo B termini e venga rimosso dalla memoria; si creerebbe un buco sopra e sotto il processo C che non consentirebbe a processi di grandi dimensione di essere inseriti all'interno della memoria. Allora, al posto di lasciare così com'è la memoria, si sceglie di compattare tutti i processi nella parte bassa in memoria al fine di avere un'unica zona di memoria con soli processi (e senza buchi) e una parte in memoria completamente a disposizione dei nuovi processi. Sembra che questa soluzione ottimale ma in realtà comporta alcuni diversi problemi. Primo tra tutti è che bisogna capire come **spostare** dei **processi** in memoria: fino ad ora venivano inseriti e rimossi ma mai spostati

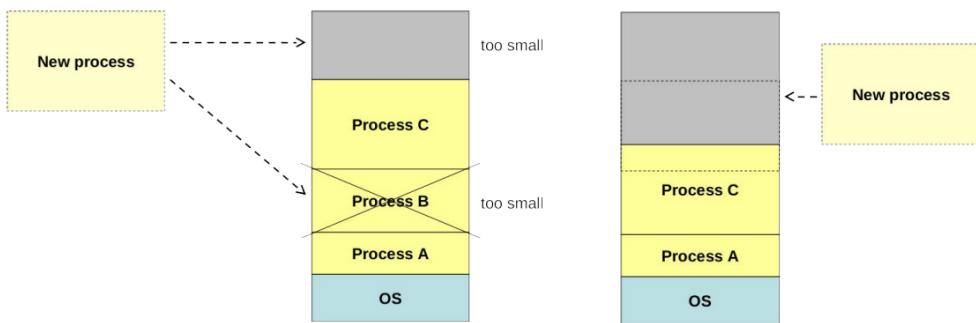


Figura 7.11: Processo di compacting, una soluzione alla frammentazione esterna.

da una zona all'altra. In secondo luogo, se si ripete questo processo per milioni di locazioni in memoria, compattare molti processi diventa un **problema computazionale** non indifferente, soprattutto se si fa ogni volta che un processo in memoria viene rimosso. È quindi evidente che la soluzione sia tutto fuorché ottimale e che quindi i problemi di frammentazione, con questo tipo di allocazioni, persistono.

7.3 Paginazione (*paging*)

Ecco quindi che introduciamo questo nuovo tipo di allocazione, molto più moderno ed elegante, ovvero la **paginazione**. Il concetto alla base di questa tecnica è quello di dividere la memoria in locazioni di dimensione fissa \bar{n} , chiamate **frames**, e dividere lo spazio logico dei processi in blocchi della stessa dimensione \bar{n} . Generalmente, nei sistemi operativi moderni, $\bar{n} \in [512 \text{ Bytes}, 16 \text{ MBytes}]$. In questo modo è quindi possibile dividere i processi in **pagine** e associare ad ogni pagina di ogni processo in esecuzione un frame in memoria, proprio come illustrato nella figura 7.12. Questa tecnica risolve subito il problema della frammentazione esterna e dei buchi. Osservando infatti la figura 7.13, nel momento in cui il processo B termina l'esecuzione e viene rimosso dalla memoria, dovrebbero rimanere dei buchi tra il processo A e il processo C. Questo buchi però vengono immediatamente colmati dal processo D le quali pagine vengono mappate nei frames che prima erano occupati da B e dal frame seguente al processo C. In questo modo, proprio perché le pagine e i frames sono della stessa dimensione si incastrano perfettamente tra di loro e non lasciano nessun buco.

Se andiamo a dividere lo spazio logico in pagine e la memoria viene divisa in frames, abbiamo bisogno di un supporto che mappa ciascuna pagina del processo ad un frame in memoria. Questo supporto è chiamato **page table** (discussa approfonditamente nel paragrafo 7.4) e si occupa appunto della

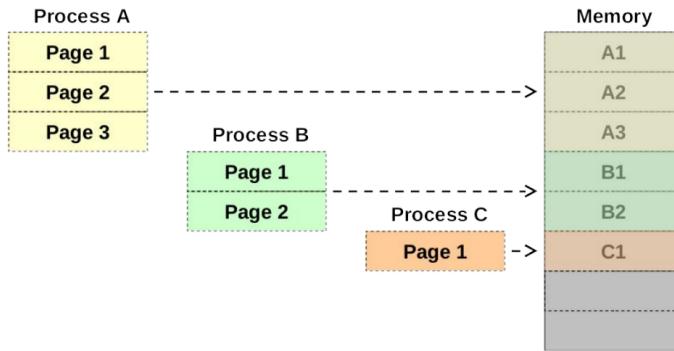


Figura 7.12: Il funzionamento ad alto livello della paginazione.

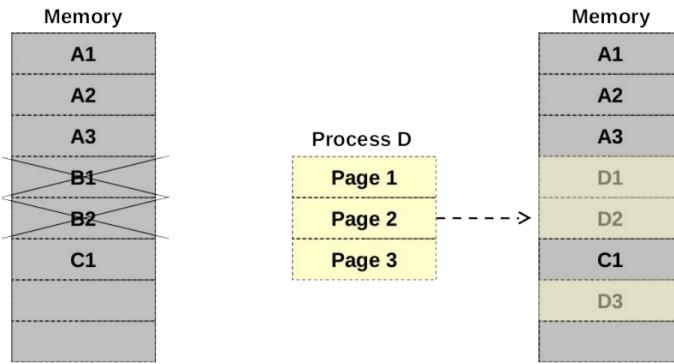


Figura 7.13: La rimozione di un processo e l'inserimento di un altro con la paginazione.

traduzione di indirizzi logici (pagine) in indirizzi fisici (frames). Si osserva che la page table è una struttura associata ad ogni singolo processo. Di conseguenza due processi hanno due page table differenti.

7.3.1 Frammentazione interna

Osserviamo infine un'ultima cosa: pur avendo eliminato il problema della frammentazione esterna, la **frammentazione interna** persiste: molto spesso infatti un processo non avrà esattamente la dimensione di n pagine, anzi, avrà generalmente una dimensione che è minore della dimensione occupata da n pagine ma comunque più grande di quella occupata da $n-1$ pagine. Di conseguenza l'ultima pagina non sarà mai completamente piena ma avrà dello spazio non occupato. Questo spazio

non potrà nemmeno essere utilizzato da altri processi dato che come sappiamo la dimensione della pagina è fissa.

Facciamo un esempio per fissare il concetto. Abbiamo un processo di 10 KBytes che deve essere allocato in memoria. Poniamo che la dimensione delle pagine (e quindi dei *frames*) è di 4 KBytes. Al fine di inserire il processo in memoria sono quindi necessarie 3 pagine, che forniscono uno spazio di $3 * 4 \text{ KBytes} = 12 \text{ KBytes}$ che è maggiore rispetto ai 10 KBytes del processo: sono quindi sprecati $12 - 10 = 2 \text{ KBytes}$ che non potranno mai essere occupati da altri processi. Al caso migliore lo spreco è di 1 Byte (ovvero la dimensione di 1 cella in memoria), mentre al caso peggiore lo spreco è di $\bar{n} - 1 \text{ Bytes}$. È ragionevole affermare che invece al **caso medio** si sprezza indicativamente la metà di un frame.

Una soluzione a cui si potrebbe intuitivamente pensare è quella di ridurre sempre di più la dimensione delle pagine al fine di sprecare sempre meno memoria. Riducendo però la dimensione delle pagine (e quindi dei frame) significa che abbiamo a che fare con un numero maggiore di pagine: se dimezziamo la dimensione di una pagina avremmo quindi il doppio delle pagine da indirizzare: è evidente che più si riduce la dimensione di un frame, la complessità aumenta. La soluzione migliore non è quindi avere delle pagine estremamente piccole ma riuscire a trovare un compromesso tra la dimensione della pagina e la complessità che ne deriva.

Calcolo della frammentazione interna

Un secondo esercizio che può ritornare utile è quello del calcolo della frammentazione interna durante la paginazione. Poniamo di avere la dimensione della pagina di 2048 Bytes (e quindi lo spazio per l'offset è di 11 bit) e che la dimensione del processo in esecuzione sia 72766 Bytes. Possiamo ora calcolare il numero di pagine occupate dal processo: $72766/2048 = 35.53$. Questo risultato ci dice che il numero di pagine completamente riempite sono 35 e che la 36esima verrà utilizzata ma solo in parte generando quindi della frammentazione interna. I byte sprecati sono infatti $(36 \cdot 2048) - 72766 = 962$ bytes inutilizzati.

7.3.2 Traduzione degli indirizzi

Al fine di convertire le pagine del processo in frames in memoria è quindi necessario un processo di traduzione. Tale traduzione è effettuata attraverso un'**interpretazione** dell'indirizzo logico. In

particolare, tale indirizzo è diviso in due parti: la prima parte (p), quella dei bit più significativi (ovvero quelli più a sinistra), indica l'indirizzo della pagina mentre la seconda parte (d) indica l'*offset* della cella dall'indirizzo della pagina. La page table si occupa di convertire il numero della pagina nel numero del frame in memoria, mantenendo l'offset invariato.

Osserviamo più nel dettaglio questo processo basandoci sull'illustrazione 7.14. Prima di tutto il processo fornisce l'indirizzo logico, che, come abbiamo appena visto, è composto da due parti. La parte più significativa, p , indica il numero della pagina, in particolare rappresenta il numero dell'indice della tabella dove è contenuto il frame. Una volta che l'accesso alla page table è stato effettuato, si preleva

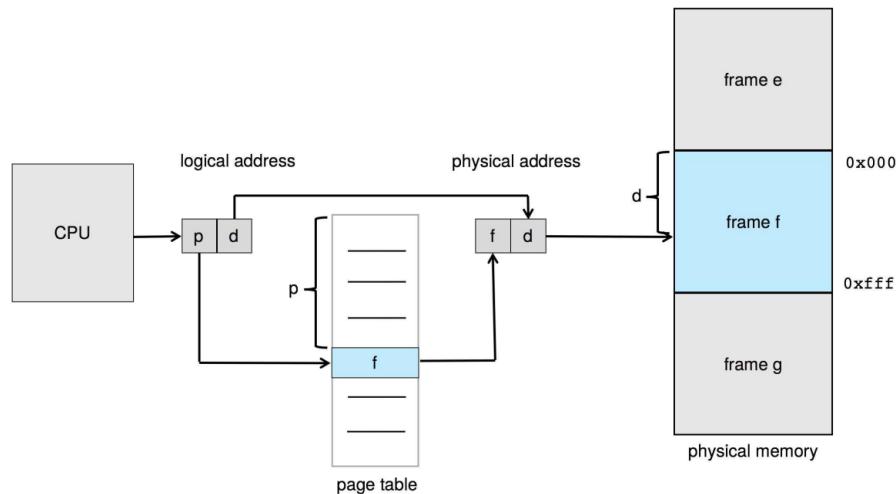


Figura 7.14: Il processo di conversione da indirizzo logico a indirizzo fisico.

il frame f e lo si sostituisce nell'indirizzo logico che diventa un indirizzo fisico. A questo punto non ci resta che andare in memoria nel frame f e aggiungerci l'offset con valore d .

Calcolo di un indirizzo tradotto

Supponiamo di avere lo spazio totale di indirizzo logico di 2^{16} , e quindi un indirizzo logico è 16 bit. Ci viene dato inoltre la grandezza di una pagina, che è 2^{12} . In altre parole, con questa informazione, sappiamo che i bit necessari per l'offset sono 12. A questo punto sappiamo che i bit disponibili per indirizzare le pagine logiche sono $16 - 12 = 4$; di conseguenza abbiamo $2^4 = 16$ pagine logiche disponibili.

A questo punto viene fornito l'indirizzo logico 0011 0000 1011 1001. Poniamo inoltre che siamo a conoscenza di tutti i valori presenti nella page table del processo. In quale indirizzo fisico viene mappato l'indirizzo dato? La soluzione è molto semplice: l'indirizzo fornito è ovviamente lungo 16 bit; sappiamo inoltre che durante la conversione vengono modificati solo i bit del numero della pagina, che in questo caso sono i primi quattro verso sinistra, ovvero 0011. A questo punto scorriamo sulla page table alla cella numero $0011_2 = 3_{10}$ e scopriamo che la pagina viene mappata nel frame 1110. Procediamo quindi con una banale sostituzione, ottenendo l'indirizzo fisico di valore: 1110 0000 1011 1001. L'illustrazione 7.15 è esemplificativa di questo processo.

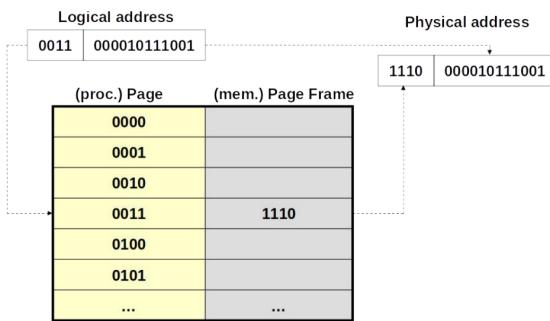


Figura 7.15: Esempio di una paginazione.

7.3.3 Allocazione nei frames liberi

Nel momento in cui si carica un nuovo processo in memoria, i frames che devono essere riempiti sono, naturalmente, quelli vuoti. Questi sono presenti in una lista (la **free-frame list**) che è detenuta dal sistema operativo al fine di tenere traccia dei frames disponibili. Di conseguenza quando deve essere inserita una pagina in memoria, il sistema operativo prende un frame disponibile dalla free-frame list che viene associato alla pagina. Questa associazione viene quindi inserita nella page table del processo. Nella figura 7.16 è illustrato questo processo. Inizialmente arriva un nuovo processo che contiene 4 pagine. Queste 4 pagine, in quanto nuove non hanno ad esse associato in frame, e di conseguenza non sono presenti nemmeno sulla tabella delle pagine. Di conseguenza, nel momento in cui la richiesta viene accettata dal sistema operativo, alle 4 pagine vengono associati 4 frames liberi che erano presenti nella lista. Dopo l'allocazione infatti notiamo che i frame che erano liberi ora contengono le 4 pagine e

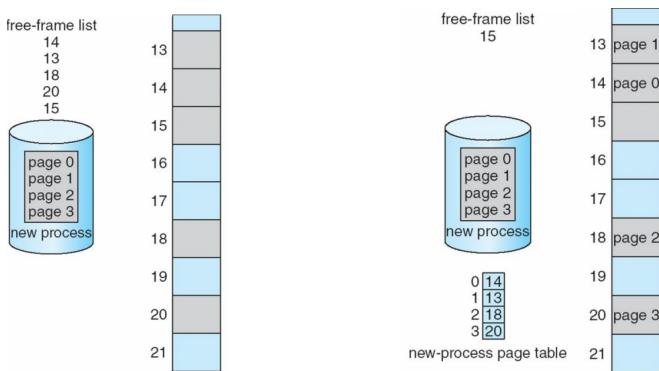


Figura 7.16: Il processo di allocazione delle pagine in nuovi frames liberi.

non sono più presenti nella free-frame list. Ancora più importante è però notare che nella pagina delle tabelle del processo si sono aggiunte le nuove associazioni.

7.4 Page table

Entriamo ora più nel dettaglio, e discutiamo in maniera più approfondita la, ormai nota, tabella delle pagine. Come sappiamo questa tabella non è unica in tutto il sistema operativo ma è presente una page table per ogni processo. Tipicamente la tabella ha dimensioni così significative che essa stessa è mantenuta in memoria. Quindi per ogni processo in esecuzione, esiste una sua page table associata che è presente in memoria. In particolare, della page table, il sistema operativo tiene traccia di due importanti valori (vedi figura 7.17), al fine di riuscire ad associare ogni processo alla sua tabella:

- ◊ *Page-Table Base Register (PTBR)*, che contiene l'indirizzo alla prima cella in memoria da cui parte la tabella;
- ◊ *Page-Table Length Register (PTLR)*, che si occupa di immagazzinare la lunghezza della tabella in memoria.

Ovviamente, nel momento in cui avviene un *context switch* (vedi paragrafo 2.2.1), i due registri vengono modificati. In questo modo si è in grado di associare ad ogni processo la sua personale page table.

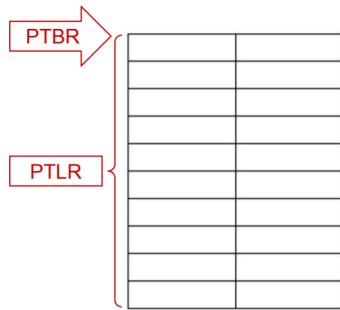


Figura 7.17: I due registri necessari per tenere traccia della page table in memoria.

7.4.1 Translation Look-aside Buffer (TLB)

Il problema di avere una tabella in memoria è che per accedere ad un indirizzo fisico in memoria, è necessario effettuare due accessi in memoria:

1. Accesso alla page table in memoria per tradurre l'indirizzo logico in un indirizzo fisico;
2. Accedere alla locazione dell'indirizzo fisico fornito dalla tabella delle pagine.

Così facendo il tempo per accedere ad un frame in memoria è raddoppiato. La soluzione a questo problema è fornita dalla **TLB**, che è una speciale **cache** la quale mantiene la porzione della page table che è usata più spesso. In questo modo la CPU non deve sempre accedere alla memoria bensì accede alla TLB che è molto più rapida e quindi consente di risparmiare tempo. Questa cache è infatti implementata fisicamente a livello di CPU e i suoi tempi di accesso sono estremamente ridotti comparati a quelli della memoria. Si osserva inoltre che, se la velocità di accesso è elevata, lo spazio di memorizzazione è molto basso: generalmente infatti le TLB contengono dalle 64 alle 1024 righe.

Ora che abbiamo inserito la TLB nel processo di traduzione da pagina logica a fisica, vediamo come questo è cambiato. L'illustrazione 7.18 rappresenta l'intero processo di traduzione. Una volta che arriva l'indirizzo logico dalla CPU si andrà innanzitutto a controllare se il numero della pagina logica è presente all'interno della TLB. Nel caso sia presente, si ha un **TLB hit** e si ottiene subito il frame associato alla pagina; di conseguenza si può subito accedere in memoria. Si osserva che la verifica della pagina all'interno della TLB viene effettuata in **parallelo**, di conseguenza tale verifica è praticamente istantanea. Se invece la pagina non è presente nella TLB, si verifica il cosiddetto **TLB miss**. A questo punto è necessario effettuare un primo accesso in memoria al fine di trovare il frame nella page table e, dopo di che, effettuare un secondo accesso al frame allocato.

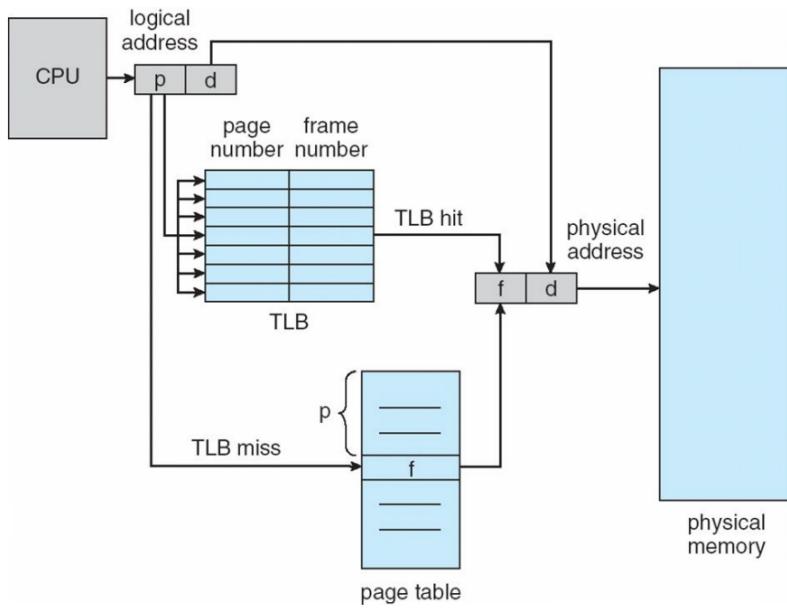


Figura 7.18: Il processo di traduzione effettuato con l'ausilio della TLB.

Va inoltre specificato che a differenza delle page table, la TLB è una unica per tutta il sistema ed è quindi "condivisa" da tutti i processi. Di conseguenza è presente un identificativo, chiamato *address-space identifier (ASID)*, che indica se la particolare entry della TLB fa parte o meno del processo che ne fa richiesta. Questo identificativo serve per fare in modo che i processi accedano solo nei loro indirizzi. Se questa funzionalità non fosse implementata, una soluzione comunque lecita è quella di resettare la TLB ogni volta che un context switch (2.2.1) si genererebbe però dell'overhead.

Si osserva infine che è anche possibile far sì che alcune entry rimangano permanentemente dentro la TLB in quanto, magari, sono delle pagine che vengono richieste quasi sempre. In questo caso si parla infatto di *wired down entries*.

Analisi delle performance con la TLB

Definiamo che **hit ratio** la percentuale secondo la quale si ha un TLB hit. Poniamo di avere un hit ratio di 80%: 80 volte su 100 si verificherà un TLB hit. Supponiamo inoltre che il tempo di accesso alla memoria sia 10 ns (nanosecondi). Da queste informazioni possiamo evincere che se si ha un TLB hit,

il tempo di accesso sarà 10 ns, altrimenti il tempo di accesso si duplicherà diventando 20 ns. A questo punto è possibile calcolare l'**EAT**, ovvero *Effective Access Time*:

$$EAT = 0.8 \cdot 10 + 0.2 \cdot 20 = 12 \text{ ns}$$

Passando però ad un hit ratio più realistico, 99%, il tempo effettivo di accesso diventa:

$$EAT = 0.99 \cdot 10 + 0.01 \cdot 20 = 10.1 \text{ ns}$$

7.4.2 Bit di validità

Uno dei meccanismi dedicati alla protezione della memoria è quello di implementare un bit di validità, chiamato **valid-invalid** bit. Alla page table viene aggiunta una colonna che contiene tale bit di validità. In questo modo, ogni entry della page table avrà un bit che indicherà se quell'associazione sia valida o meno. In particolare il bit si occupa di segnalare se la pagina si trova nello spazio di indirizzi logici del processo. Osservando la figura 7.19, notiamo che all'interno della page table le pagine 6 e 7 sono invalide proprio perché non sono presenti nello spazio di indirizzi del processo (che si ferma alla quinta pagina).

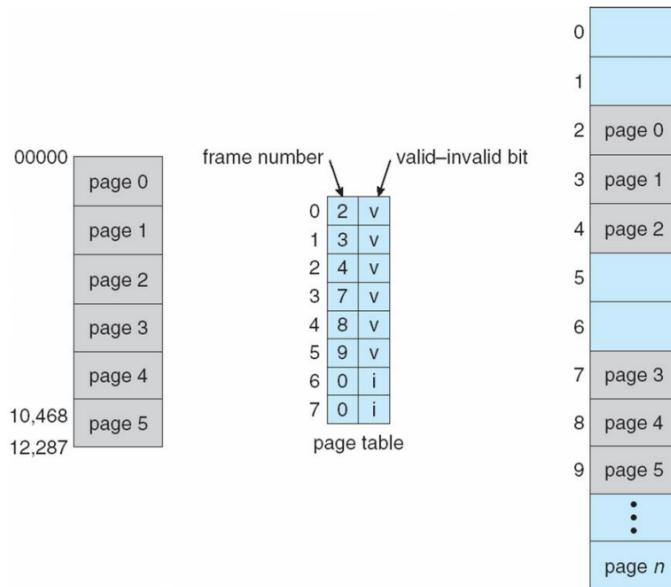


Figura 7.19: Implementazione del valid-invalid bit sulla tabella delle pagine.

Il problema delle macchine moderne

La page table inserita in maniera continua in memoria con l'ausilio della TLB sembra una soluzione ottima. E lo era un tempo, quando gli indirizzi erano al più di 16 bit. Con l'avvento però di indirizzi da 32 e 64 bit la situazione è però peggiorata perché la dimensione delle pagine è però sempre la stessa e non può cambiare. Poniamo di avere delle pagine da 4KB, che quindi necessitano di 12 bit per l'indirizzamento, rimangono $32 - 12 = 20$ bit per indirizzare tutte le pagine. Ciò significa che è possibile avere fino a 2^{20} pagine da indirizzare, ovvero avere una tabella con più di un milione di righe. Nei casi ancora più moderni, quando si ha a che fare con indirizzi da 64 bit, si ha la possibilità di avere $2^{52} \approx 4 \cdot 10^{15}$ ovvero 4 biliardi di pagine.

Ricordiamo che alla base della paginazione c'è la necessità di non dover allocare in maniera continua nulla in memoria, di conseguenza è giusto che anche la page table abbia lo stesso trattamento. Ecco quindi che nascono diverse tecniche al fine di riuscire a spezzettare anche la page table all'interno della memoria.

7.4.3 Page table gerarchica

La prima soluzione che si è pensata è quella di creare una page table che indirizzi un altro numero di page tables che a loro volta indirizzano i frames in memoria, proprio come rappresentato nell'illustrazione 7.20. Come cambia ora il processo di traduzione tra una pagina e un frame? Partiamo da un indirizzo a 32 bit e poniamo che le pagine abbiano la dimensione di 4 KB e che quindi necessitino di 12 bit per essere indirizzati (d). A questo punto i restanti 20 bit vengono a loro volta divisi in due parti: la prima parte p_1 punta ad un elemento della *outer page table* e la seconda parte p_2 punta ad un elemento della tabella puntata da p_1 (figura 7.21). In questo modo però puntualizziamo che, senza l'ausilio di una TLB cache, al fine di raggiungere un frame in memoria è necessario fare ben 3 accessi ad essa, triplicando così il tempo necessario per accedere al frame.

Nelle architetture a 64 bit invece le soluzioni possono essere 2. Avendo 52 bit a disposizione si sceglie di creare una *outer page table* molto grande, indirizzata con 42 bit, e delle tabelle di secondo livello indirizzate da 10 bit. Una seconda soluzione può essere invece l'inserimento di un terzo livello di page table (**three-level paging scheme**) dove, per esempio, si ha l'outer page table indirizzata con 32 bit, e la tabella secondaria e terziaria indirizzate con 1 bit ciascuna. Si osserva che in questo caso il numero di accessi in memoria quadruplica al fine di accedere ad un frames. Proprio perché la struttura

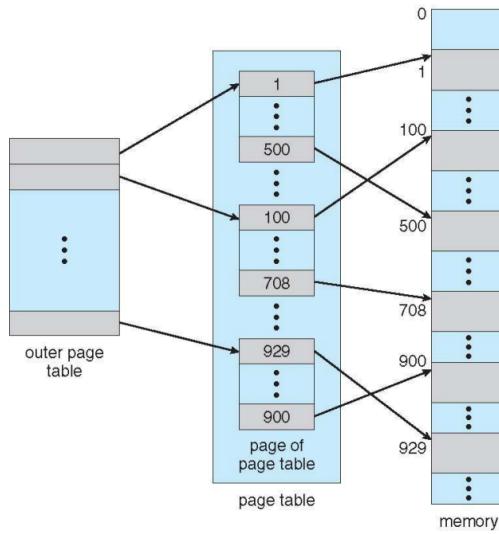


Figura 7.20: Struttura gerarchica della page table.

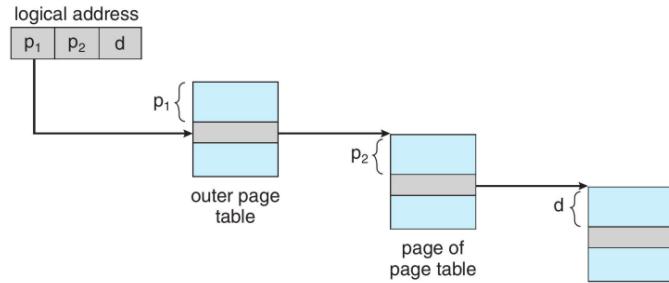


Figura 7.21: L'accesso in memoria attraverso una struttura gerarchica della page table.

gerarchica non si presta molto bene ad architetture con indirizzi maggiori di 32 bit, si sono progettate altre soluzioni.

7.4.4 Page table con tabella hash

Al fine di provare a risolvere il problema che è emerso utilizzando una struttura gerarchica si è pensato di implementare una **hashed page table**: ciò significa convertire il numero della pagina attraverso una *hash function* che restituisce una chiave. Tale chiave è l'indice di una tabella che contiene il frame che stiamo cercando. Questo meccanismo è illustrato in modo chiaro nella figura 7.22 sottostante. La

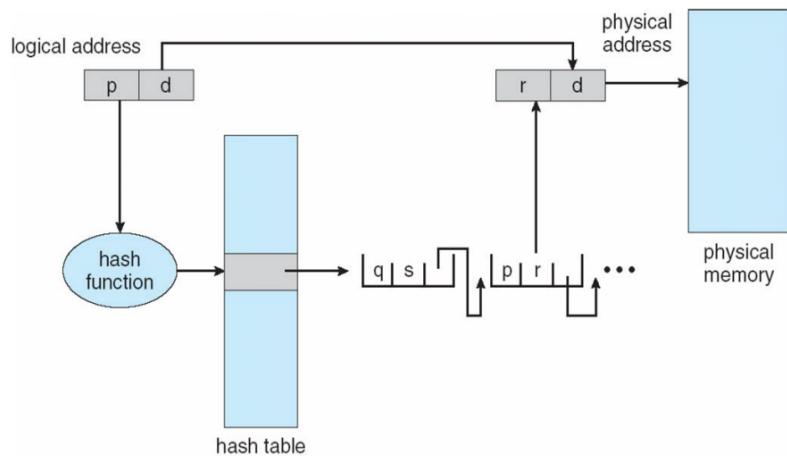


Figura 7.22: Accesso alla memoria tramite una page table implementata attraverso una hash table.

hash table, a seconda del range delle chiavi fornito, può essere infatti di dimensione molto ridotta. Ricordiamo inoltre che la hash function per valori di pagine diverse può restituire le stesse chiavi: proprio per questo motivo ad ogni entry della hash table in realtà è presente una lista concatenata che contiene tutte le entry che hanno avuto una collisione. Nel caso della figura 7.22, il valore della pagina *p* era situato alla seconda posizione della lista.

Introduciamo anche la **clustered page tables** che è una struttura più moderna e migliorata rispetto alla *hashed page table*. In questa struttura, ogni entry della page table, non corrisponde ad una pagina bensì ad un **cluster** di frames, che può arrivare fino a 16. In questo modo è possibile fare riferimento a frames che possono essere sparsi all'interno della memoria. Così facendo i casi con frames sparsi in diverse locazioni sono più facilmente gestibili. Questa struttura è quindi più efficiente rispetto alla classica implementazione con la tabella hash.

7.4.5 Page table invertita

Una terza soluzione è la cosiddetta tabella delle pagine invertita. Fino ad ora abbiamo detto che ogni processo ha la sua propria tabella delle pagine. Questo è necessario perché ogni processo ha il suo indirizzamento logico a cui può corrispondere un insieme di frames nella memoria. Il problema è che generalmente lo spazio di indirizzo logico non è mai utilizzato completamente. Questo perché tipicamente diverse entry all'interno della page table del processo sono invalide e quindi inutilizzate,

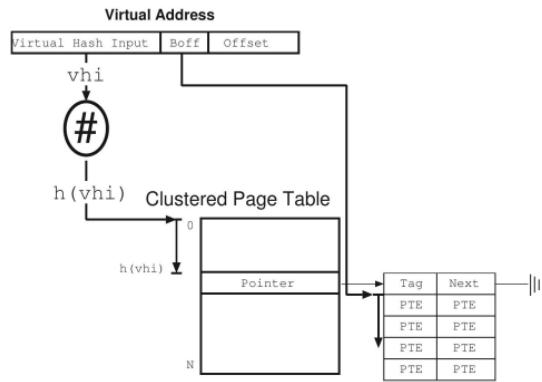


Figura 7.23: Rappresentazione di una *clustered* page table.

ma sono comunque presenti.

Allora perché non fare l'opposto? Invece di avere tante tabelle, di notevoli dimensioni che puntano a frames fisici, si è scelto di utilizzare una **tabella unica** che è indicizzata dai frames fisici reali, e contiene all'interno di ogni elemento la pagina che il frame contiene. Ciò significa che al posto di avere tante tabelle, una per ogni processo, ci ritroviamo con un'unica tabella dove saranno contenuti i frames della memoria fisica e le pagine corrispondenti al processo. Inoltre, nella tabella, sarà necessario aggiungere, per ogni pagina, il numero identificativo del processo, il `pid`, visto nel capitolo 2, altrimenti non si riuscirebbe a risalire a quale processo appartiene quella pagina. Possono infatti essere presenti due pagine logiche 0110, ma una appartiene al processo 314 e l'altra magari appartiene al processo 159.

La figura 7.24 mostra il procedimento necessario per accedere ad un frame in memoria. Prima di tutto è necessario effettuare una **ricerca** sulla tabella invertita l'identificativo del processo (`pid`) e la pagina cercata. Se tale ricerca ha buon fine allora il frame in memoria è presente ed è quindi possibile effettuare l'accesso.

Discutiamo, infine, alcuni vantaggi e svantaggi della tabella delle pagine invertita. Il vantaggio più evidente è che lo spazio occupato dalla tabella invertita è nettamente minore rispetto alla somma dello spazio occupato da ciascuna tabella per i processi in esecuzione. D'altro canto però il tempo di **ricerca lineare** nella tabella invertita genera dell'overhead. Questo però può essere ridotto implementando una tabella di hash per limitare la ricerca all'interno della tabella invertita. È inoltre possibile migliorare le prestazioni tramite l'ausilio della TLB cache.

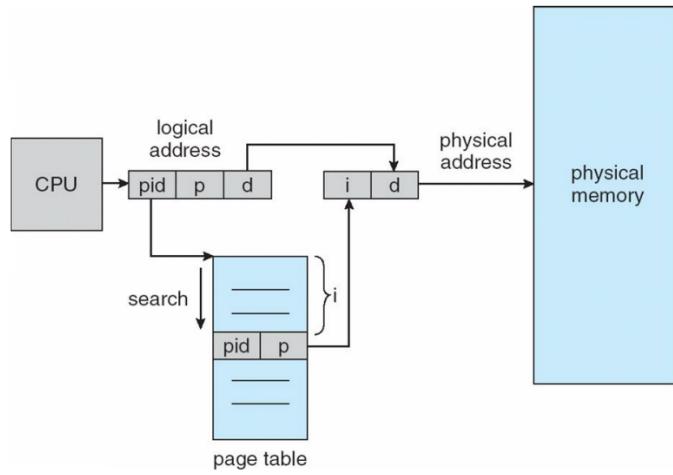


Figura 7.24: Accesso in memoria attraverso la tabella delle pagine invertita.

7.5 Swapping

Per quanto il metodo per paginare la memoria sia efficiente, ci saranno sempre dei casi in cui è necessario rendere dello spazio disponibile in memoria, rimuovendo uno o più processi. Introduciamo ora una tecnica che serve proprio per questo, si chiama *swapping* (su disco o, in generale, sul *backing store*), che è il procedimento secondo il quale si prende un processo dalla memoria e lo si salva temporaneamente sul disco rigido. Immaginiamo di avere in memoria 100 processi ed aver esaurito lo spazio. per aggiungere il 101 esimo è necessario per forza rimuovere dei processi. Si sceglie infatti di prendere quei processi che sono in attesa (per esempio di un I/O) e che sono fermi in memoria e portarli sul disco, proprio come illustrato nella figura 7.25. Il tempo di *swap-in* e *swap-out* rimane comunque un tempo significativo dato che il *backing store* è una memoria ad alta capacità ma molto lenta. È quindi necessario tenere conto anche del **tempo di trasferimento** nel momento in cui si scambiano i processi. Ovviamente sarà presente anche una **ready-queue** che contiene tutti i processi presenti nel disco che non sono più in attesa di un evento e che sono quindi pronti per essere inseriti in memoria. Naturalmente al tempo di trasferimento del processo *in / out* dal disco va sommato anche il tempo necessario per effettuare il context switch (2.2.1), ovvero ricaricare lo stato del processo nella CPU. Sono presenti anche delle varianti dello swapping, dove si va a scegliere quale processo portare fuori in base alla priorità: è un concetto molto simile allo scheduling con priorità che abbiamo visto nel

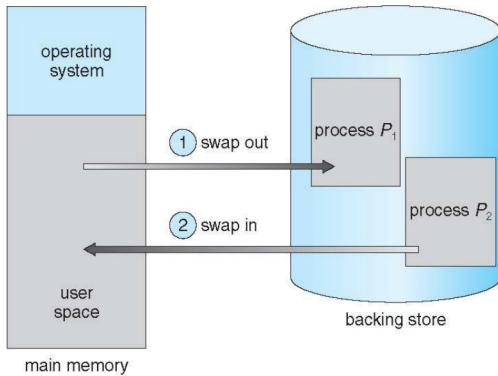


Figura 7.25: Processo di *swap-in* e *swap-out* dal backing store.

paragrafo 4.4.

7.5.1 Swapping con paginazione

Lo swapping classico, tipicamente non è più utilizzato nei sistemi operativi moderni. Spesso infatti non si fa lo swapping dell'intero processo ma solamente di alcune pagine di quel processo. Ecco quindi che parliamo di *swapping with paging*, rappresentato in figura 7.26. La scelta delle pagine da rimpiazzare sarà delegata agli algoritmi di *page replacement* (vedi paragrafo 8.3).

7.5.2 Swapping nei dispositivi mobili

Nei sistemi operativi per *mobile devices* la questione è un po' più complicata in quanto effettuare uno swapping è molto dispendioso, sia da un punto di vista energetico, ma soprattutto perché i supporti di memorizzazione dei dispositivi mobili spesso hanno un numero limitato di accessi in scrittura e dopo di ché le performance diminuire. Di conseguenza si cerca sempre di evitare la scrittura sul backing store.

Sono presenti ovviamente dei metodi alternativi. Per esempio **iOS** al fine di liberare memoria domanda a dei processi in esecuzione se qualcuno può volontariamente rilasciare la memoria altrimenti può forzatamente terminare un processo in esecuzione. D'altra parte, **Android** sceglie di terminare applicazioni che generalmente sono dormienti o in background. Nel momento in cui l'applicazione è terminata lo stato dell'applicazione viene comunque memorizzato al fine di essere già pronta per ripartire.

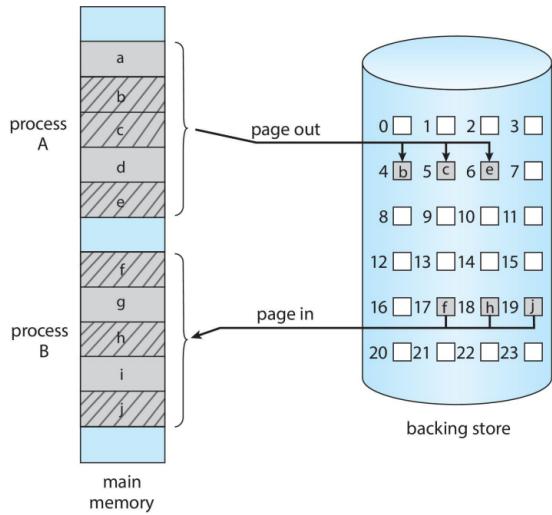


Figura 7.26: Rappresentazione del processo di swapping solo su determinata pagine di un processo.

7.6 Segmentazione (*segmentation*)

Introduciamo ora un nuovo modello di allocazione della memoria che è tuttora utilizzato ma non è tanto diffuso quanto la paginazione: stiamo parlando della segmentazione. Quando abbiamo parlato della paginazione, avevamo più volte specificato che la dimensione della pagina è fissa. Nel *segmentation* invece noi permettiamo che la dimensione sia variabile. Non parliamo delle pagine che hanno la stessa dimensione, ma di **segmenti** di dimensione variabile. Osservando la figura 7.27, notiamo che i processi sono appunto divisi in segmenti di dimensione variabile. La dimensione di un

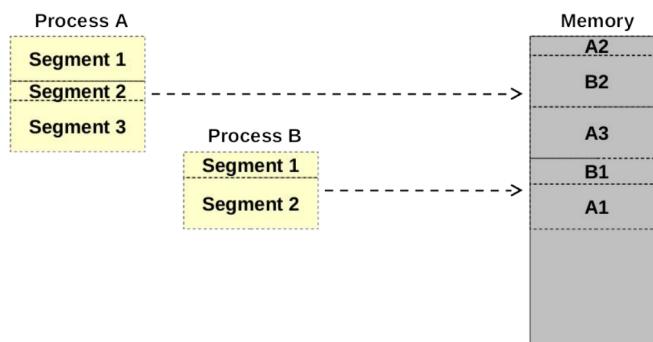


Figura 7.27: Il funzionamento ad alto livello della segmentazione.

segmento è decisa in base all'organizzazione e alla logica del programma stesso: potrebbe essere che alcuni segmenti contengono il `main`, altri contengono delle librerie, altri ancora che contengono le variabili.

7.6.1 Segment table

Al fine di tener traccia di questi segmenti non è più necessaria una tabella delle pagine, bensì una **tabella dei segmenti**. Questa contiene tre campi:

1. I primi bit dell'indirizzo logico corrispondono al **numero di segmento**;
2. La **dimensione** del segmento che si trova in memoria;
3. L'indirizzo della prima cella del segmento, ovvero il **base address**.

Osserviamo la figura 7.28 e discutiamo nel dettaglio il processo di traduzione da un indirizzo logico ad uno fisico. Quando un indirizzo logico viene fornito, questo è diviso in due parti: il numero del segmento e l'offset. A questo punto si prende il numero del segmento, che è l'indice della segment

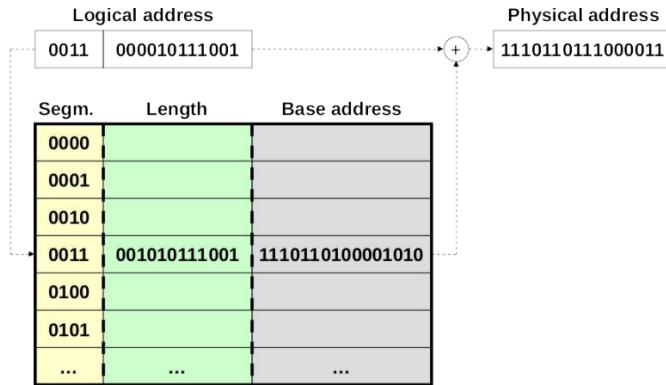


Figura 7.28: processo di traduzione di un indirizzo logico a fisico mediante la segment table.

table. A questo punto si controlla che l'offset sia minore della dimensione del segmento in memoria, in modo tale da garantire che il processo non invada celle di altri processi. Una volta che la verifica è andata a buon fine, se prende l'indirizzo base e lo si **somma** con l'offset in modo da ottenere l'indirizzo in memoria del segmento. Si ricorda che nella paginazione il frame veniva concatenato con l'offset, qui invece si somma l'offset all'indirizzo base.

7.6.2 Modello ibrido

In molti casi è preferibile combinare la paginazione con la segmentazione. Si va così a creare un **page-segmented system**. Osservando la rappresentazione 7.29, cerchiamo di capire come questi tipi di sistemi funzionino. Innanzitutto l'indirizzo logico (virtuale) è diviso in tre parti: la prima indica la segment table, il secondo indica la page table e il terzo invece rappresenta l'offset. Inizialmente, quando arriva l'indirizzo logico, si accede alla segment table: al suo interno è contenuto l'indirizzo iniziale della page table. A questo punto, utilizzando la seconda parte dell'indirizzo logico si accede alla

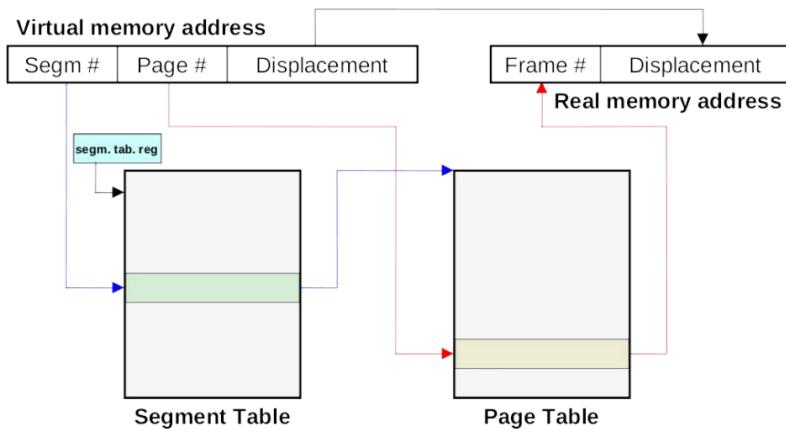


Figura 7.29: Rappresentazione di un modello che combina paginazione e segmentazione.

riga della tabella che contiene l'indirizzo del frame in memoria. Questo frame va quindi concatenato con l'offset.

8 MEMORIA VIRTUALE

In questo capitolo cerchiamo di capire com un sistema operativo gestisce la memoria attraverso le tecniche che abbiamo visto nel capitolo precedente (7). Ci occuperemo prima di tutto del concetto di *demand paging* e alcune tecniche di ottimizzazione come il *copy-on-write*. In secondo luogo passeremo a discutere di tutti gli algoritmi di rimpiazzo delle pagine (*page replacement*) in memoria. Discuteremo infine il fenomeno del *thrashing* e di alcune soluzioni che adottano i moderni sistemi operativi.

8.1 Introduzione

Sappiamo che il codice, al fine di essere eseguito, deve essere presente nella memoria. Quando parliamo di memoria virtuale, parliamo infatti della separazione tra indirizzo logico e fisico. Fino ad ora abbiamo discusso di processi che, anche se divisi in pagine o in segmenti, venivano comunque caricati tutti in memoria. Da adesso introduciamo la possibilità di poter eseguire un processo anche se le sue pagine non sono tutte in memoria, anzi, possono risiedere sul file system (11) o sul *backing store*. Infatti quando un processo è eseguito, non è necessario che l'intero codice del programma sia in memoria, è sufficiente solo un sottoinsieme. Osservando infatti la figura 8.1, possiamo notare che le pagine nella memoria virtuale non sono tutte presenti nella memoria fisica, bensì alcune risiedono nei *backing store*.

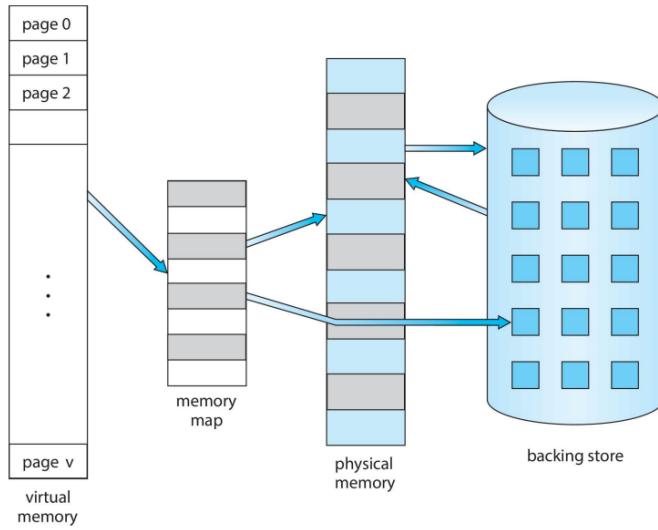


Figura 8.1: Alcune pagine del processo sono in memoria, altre sono sul disco.

8.1.1 Spazio degli indirizzi virtuali

Quello che nel capitolo precedente abbiamo definito come lo spazio degli indirizzi logici in realtà si chiama più propriamente spazio di indirizzi virtuali. Questo è lo spazio degli indirizzi che il programma vede: per ogni programma quindi questo spazio parte dall'indirizzo zero e consiste in un insieme di indirizzi continui. L'implementazione di questo spazio virtuale può essere effettuata attraverso delle tecniche di paginazione o segmentazione chiamate come *demand paging* e *demand segmentation*.

Il fatto di riuscire ad implementare uno spazio di indirizzi virtuali contigui ci permette di implementare delle strutture che dal punto di vista virtuale sono quelle viste dal processo (come la figura 2.1 nel capitolo 2).

8.1.2 Memoria condivisa

Tramite la dinamicità fornita dalla memoria virtuale è possibile mappare delle zone in memoria condivise e quindi avere dei riferimenti in processi diversi alla stessa zona in memoria (8.2). In queste zone di memoria possono essere condivise delle librerie, come nel caso delle `dll`, viste nel paragrafo 7.1.4, oppure utilizzare questa zona per della comunicazione tra processi, ovvero la *IPC*, *Inter Process Communication* (2.3).

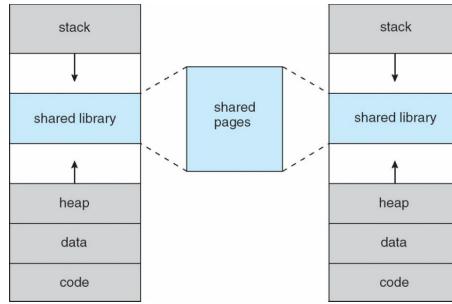


Figura 8.2: La memoria virtuale fornisce la possibilità di condividere delle librerie.

8.2 Demand Paging

Al fine di essere in grado di trasformare tutti i *virtual address space* in indirizzi fisici caricando solo le pagine che vengono utilizzate e non tutte le pagine del processo si utilizza la tecnica del **demand paging**, illustrato in figura 8.3. Ciò significa che se abbiamo dei processi A e B che richiedono delle particolari pagine (contenenti codice o dati), si andranno a caricare quelle pagine su domanda dal *secondary storage* (come l'hard disk) alla memoria principale. Dovremo quindi essere in grado di

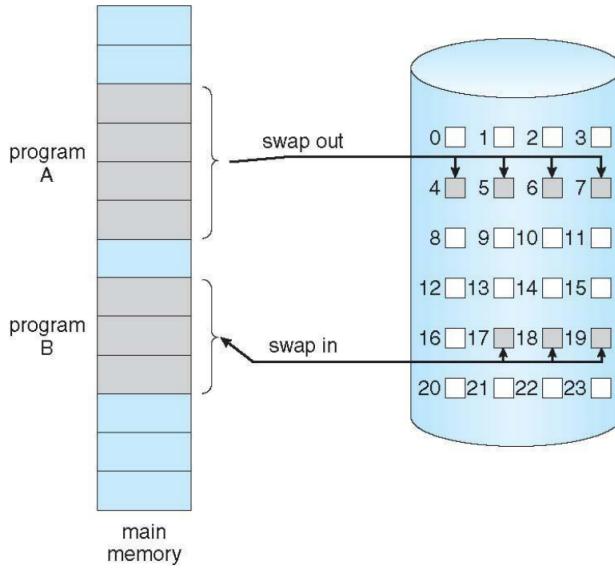


Figura 8.3: Illustrazione del funzionamento del demand paging nel caso di due processi, A e B.

gestire questa richiesta per caricare dalla memoria secondarie le pagine oppure per scaricare le pagine dalla memoria principale, ovvero la memoria RAM. Al fine di riusciri, ci appoggeremo a dei supporti hardware come il **bit di reference** che indica se la pagina di cui abbiamo bisogno è in memoria o sullo storage secondario. Nel caso in cui la pagina si trova già in memoria si dice infatti che la pagina è **memory resident**, ovvero che è immediatamente disponibile.

8.2.1 Page fault

Per garantire che in memoria ci sia sempre il set di pagine richiesto si avrà bisogno del supporto hardware, in particolare si fa riferimento alle funzionalità della MMU. In particolare, questa fornisce la possibilità di settare il *valid/invalid bit* - nella tabella delle pagine - il quale indica se la pagina che andiamo a richiedere è già presente in memoria o meno. Se la pagina non è in memoria si genera un interrupt chiamato **page fault** che deve essere ovviamente gestito dal sistema operativo.

Osserviamo la figura 8.4. Abbiamo a che fare con un processo che richiede 8 pagine che però non sono presenti tutte in memoria: dalla *page table* possiamo notare che la pagina A è in memoria e si trova nel frame 4, la pagina C è associata al frame 6 e infine la pagina F è associata al frame 9. Tutte le altre pagine

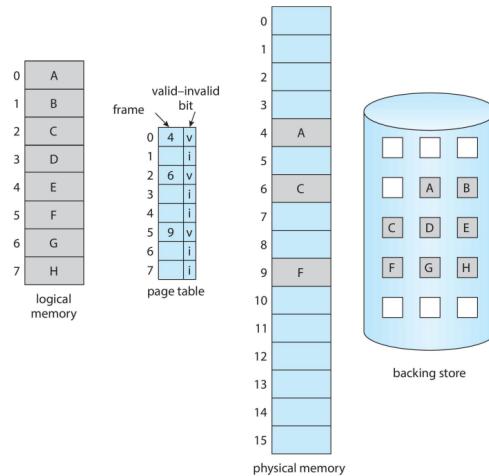


Figura 8.4: Generazione di un page fault.

non sono presenti in memoria e quindi il loro *invalid bit* è impostato a "v". Per quanto riguarda le altre

pagine invece si verificano dei *page fault* dato che non si trovano in memoria ma solo nel *backing store*: possiamo infatti notare che l'*invalid bit* di queste pagine è settato a "i".

Come si comporta il sistema operativo adesso? Basandoci sull'illustrazione 8.5, possiamo osservare che ci sono 6 steps da compiere. Dopo aver richiesto la pagina (punto 1) e, mediante la page table,

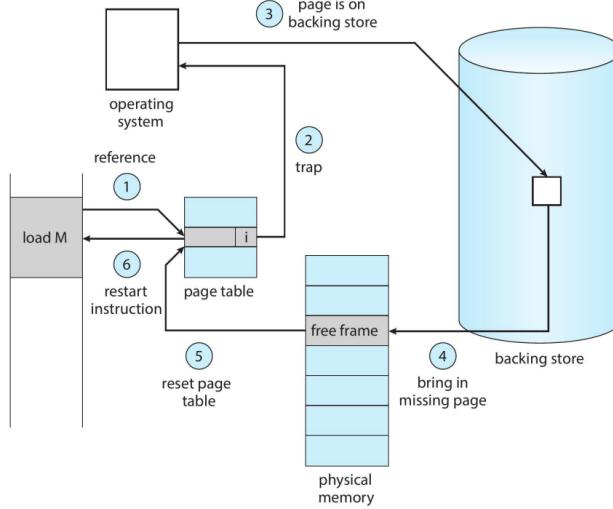


Figura 8.5: Gli steps necessari per prelevare la pagina dallo storage secondario

aver constatato che non è presente in memoria (punto 2), il sistema operativo cattura il page fault e va a cercare il frame sulla memoria secondaria (punto 3). Assumendo che ci siano sempre degli spazi liberi in memoria, il sistema operativo ne sceglie uno e inserisce il frame che ha trovato sul *backing store* (punto 4). A questo punto aggiorna l'*invalid bit* sulla *page table* e specifica in quale frame si trova la pagina virtuale (punto 5). Infine l'istruzione viene fatta ripartire (punto 6): in questo modo l'istruzione può essere eseguita e il processo può continuare.

8.2.2 Pure demand paging

Appena iniziamo l'esecuzione di un processo, questo deve ancora essere caricato in memoria. Ebbene, il demand paging puro consiste nel caricare solo le pagine del processo che vengono richieste. In altre parole il processo viene caricato in memoria frammento dopo frammento. Questo, per quanto istintivo possa sembrare, è in realtà estremamente **inefficiente**: significherebbe aumentare generare un numero elevatissimo di *page fault* e di conseguenza il sistema operativo dovrebbe passare per il *backing store*

molte volte. Fortunatamente però i processi godono di una proprietà chiamata **locality of reference**: questo significa che la maggior parte del codice e dei dati di un processo è raggruppata tutta assieme e non sono sparpagliati nella memoria.

Un approccio sicuramente più efficiente è quello di cercare di caricare il maggior numero di pagine possibili in memoria al fine di minimizzare il numero di page fault e quindi di accessi alla memoria secondaria.

Gestione della memoria libera

Fino ad ora abbiamo assunto che ci siano sempre dei frame liberi in memoria ma naturalmente non è così (vedi paragrafo 8.3). Il sistema operativo possiede quindi una lista dei frame che sono liberi, la cosiddetta **free-frame list**, che viene utilizzata per identificata dove poter inserire la nuova pagina. Nel caso non ci siano elementi liberi è necessario liberare dello spazio in memoria attraverso gli algoritmi di rimpiazzo (paragrafo 8.3).

Nel momento in cui dei frame vengono liberati tramite questi algoritmi va ad azzerare il contenuto utilizzato precedentemente, in modo tale da garantire che il nuovo processo non vada a leggere dati che appartengono al vecchio processo. Questa tecnica è chiamata **zero-fill-on-demand**.

8.2.3 Performance e ottimizzazione

Come abbiamo visto, nel momento in cui viene catturato un page fault il tempo perso a sostituire la pagina desiderata è notevole. Ci sono molte cause che portano le tempistiche ad essere così elevate però, generalmente le principali sono le tre seguenti:

1. La *routine* che gestisce il page fault;
2. Il tempo per leggere la pagina dal disco;
3. Il tempo necessario per far ripartire il processo (ricaricare i dati e i registri nel momenti in cui la pagina è in memoria).

Per valutare le performance del demand paging ci si basa su una variabile chiamata **page fault rate**, che è la frequenza con cui un page fault avviene. Quando questa vale 0, non si verificano mai questi interrupt, quando è 1 si verificano ad ogni nuova esecuzione che viene eseguita. Ovviamente, il nostro obiettivo è quello di minimizzare il più possibile questo fattore. Questo valore lo utilizziamo per

calcolare l'**EAT** (*Effective Access Time*), ovvero il tempo di accesso reale, effettivo. Sia p il *page fault rate*, per calcolare l'EAT si utilizza la seguente formula:

$$(1 - p) \cdot \text{Memory Access Time} + \\ p \cdot (\text{Page fault Overhead} + \text{Swap Page Out} + \text{Swap Page In})$$

Al fine di riuscire a diminuire l'EAT si può cercare di ottimizzare alcuni aspetti. Si può cercare di diminuire il tempo di *swap-out* e di *swap-in* dei frame nella memoria oppure si può cercare di copiare l'intero processo sullo swap space. Si possono sfruttare anche altre proprietà dei processi: se le pagine che abbiamo in memoria che devono essere sostituite non sono state modificate è sufficiente eliminare le pagine al posto di andarle a salvare sul disco (tanto sono uguali).

Infine, una tecnica molto importante ed ingegnosa è quella del **Copy-On-Write** (figura 8.6) dove, nel momento in cui un processo padre viene *forkato* (vedi 2.2.2), il processo figlio condivide le stesse pagine del padre: è molto più efficiente che copiare le stesse pagine due volte in memoria. Le pagine continuano ad essere condivise fino a che tali pagine non devono essere modificate: ecco perché si chiama in questo modo, una pagina viene copiata solo nel momento in cui si deve scrivere su di essa.

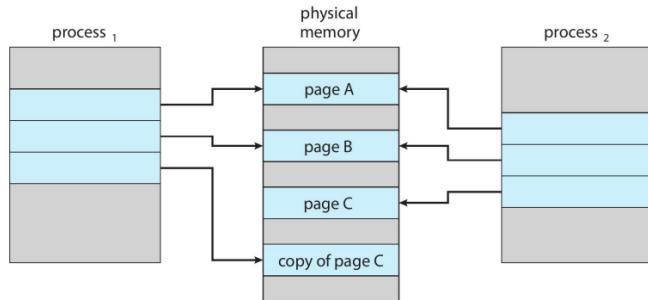


Figura 8.6: Dato che P1 deve scrivere sulla pagina C, questa viene copiata.

8.2.4 Prepaging

Controposto al *pure demand paging*, dove, dopo aver caricato il processo all'inizio, si carico le pagine che necessita attraverso i page fault. Ovviamente, c'è anche la possibilità di non caricare solamente una pagina del processo, bensì caricare un insieme di pagine al fine di diminuire la frequenza di page fault e quindi risparmiare più tempo

8.3 Page replacement

Fino ad ora abbiamo dato per scontato che ci fosse sempre dello spazio libero in memoria. Spesso invece non è così e il sistema operativo si trova nella condizione di dover rimuovere una pagina dalla memoria per sostituirla con quella che è stata appena richiesta. Come scegliere la pagina da scartare? È proprio questo il compito di un **algoritmo di rimpiazzo**. Attraverso i *replacement algorithms* ben studiati è quindi possibile avere una grande memoria virtuale in una memoria fisica di dimensioni ridotte.

Nella pratica le operazioni generali di un algoritmo di rimpiazzo sono (vedi anche figura 8.7):

1. Selezionare un frame vittima dalla memoria e portarlo nel *backing store* (*swap out*);
2. Cambiare il bit nella page table da *valid* a *invalid*;
3. Copiare la nuova pagina dallo storage secondario alla memoria *swap-in*;
4. Aggiornare l'*invalid* bit nella page table a *valid*.

Osserviamo anche che in questi casi è aggiunto un altro bit, chiamato **dirty bit** che segnala se una pagina in memoria è stata modificata o meno. Di conseguenza se la pagina non è stata modificata viene rimossa, altrimenti deve essere copiata nella memoria secondaria.

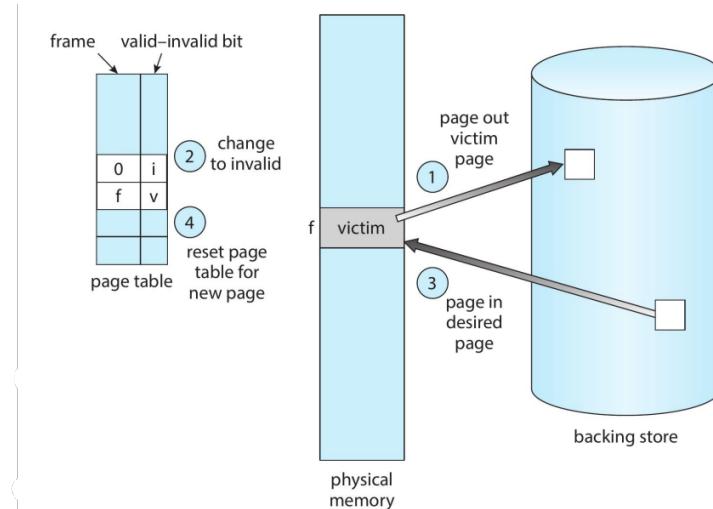


Figura 8.7: Funzionamento generico di un algoritmo di rimpiazzo.

Ricordiamo che un algoritmo di *page replacement* deve tenere conto di diversi aspetti. Innanzitutto bisogna capire quanti e quali frame devono essere assegnati per ciascun processo? È meglio avere una distribuzione equa oppure proporzionato all'importanza o alla priorità che hanno? In secondo luogo il compito di un algoritmo di rimpiazzo rimane quello di **ridurre il page fault rate**.

8.3.1 FIFO

Il primo algoritmo che vediamo, come per lo scheduling e per altre occasione, è un algoritmo simile ad una coda: *First In First Out*. In altre parole la vittima che si sceglie è il frame che è stato inserito in memoria meno recentemente.

Prendiamo in considerazione la stringa 7 0 1 2 0 3 0 4 2 3 0 3 3 2 1 2 0 1 7 0 1, dove ogni numero rappresenta la pagina che si sta richiedendo. Supponiamo inoltre, per semplicità, che la nostra memoria sia in grado di contenere solamente 3 frames. Simuliamo una le richieste e contiamo il numero di *page fault* che si sono verificati durante le richieste. Commentiamo il comportamento di

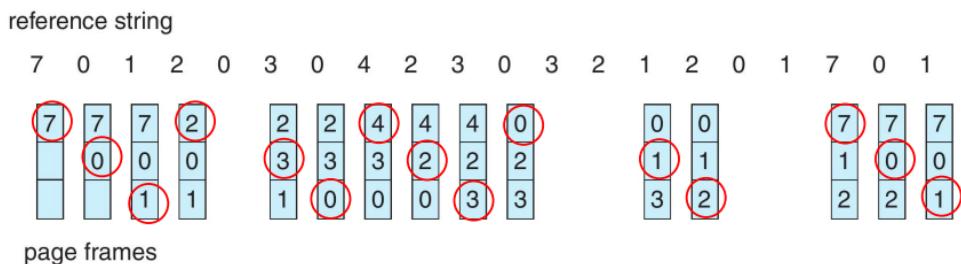


Figura 8.8: Comportamento dell'algoritmo di rimpiazzo FIFO.

questo algoritmo basandoci sulla figura 8.8. Con le prime 3 pagine si verifica un *page fault* in quanto la memoria da vuota deve riempirsi. Al quarto *timestamp*¹, quando si ha bisogno della pagina 2, questa non è presente in memoria, si verifica un *page fault* e si scambia pagina 2 con pagina 7 questo perché, tra le tre pagine presenti, è stata la prima ad entrare e quindi è la più "vecchia". Osserviamo che al quinto *timestamp*, si richiede la pagina 0 che però è già presente in memoria e quindi non si verifica alcun *page fault*. Proseguendo così per tutta la stringa possiamo notare che su 20 richieste si verificano 15 *page fault*.

¹Con *timestamp* si intende l'istante di tempo che stiamo esaminando.

Osserviamo un avvenimento curioso, chiamato **Belady's anomaly**. In modo contro-intuitivo, all'aggiungere di più frames si incrementa il numero di page fault anziché andarla a diminuire. Dando un'occhiata alla figura 8.9 notiamo passando da 3 a 4 frames il numero di page fault cresce.

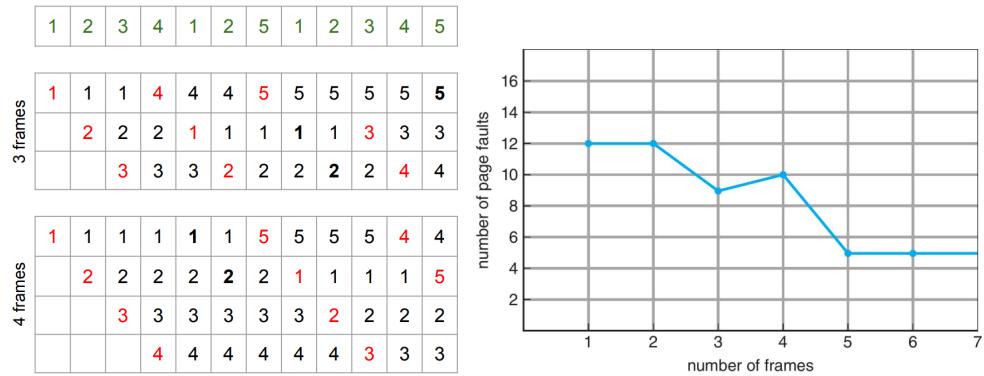


Figura 8.9: Una piccola illustrazione dell'anomalia di Belady.

8.3.2 Algoritmo ottimale

Il problema dell'algoritmo FIFO, oltre all'anomalia, è che nel momento in cui rimuoveva un frame dalla memoria non andava a domandarsi quanto questo frame è stato utilizzato e quindi senza tenere conto della storia dei frame. A livello teorico, un algoritmo ottimale rimuoverebbe il frame che verrà utilizzato il più tardi possibile nel futuro. Questo algoritmo è ovviamente utopico dato che conosce tutta la stringa. Nella realtà, ovviamente, non si è in grado di prevedere il futuro però si è in grado di fare delle buone stime.

Riprendiamo in considerazione la stringa `7 0 1 2 0 3 0 4 2 3 0 3 3 2 1 2 0 1 7 0 1` e cerchiamo di capire come l'algoritmo ottimale rimpiazzi i frames all'interno della memoria (figura 8.10). Dall'immagine notiamo che le prime tre richieste sono page fault dato che la memoria da vuota deve riempirsi. Al quarto *timestamp* notiamo che anche in questo caso la pagina 7 viene rimossa perché è quella che viene utilizzata più tardi (verso la fine praticamente). Dopo di che, alla richiesta della pagina 0 non viene rimosso nulla perché verrà utilizzata nel futuro prossimo (più precisamente dopo due *timestamp*). Proseguendo, quando viene richiesta la pagina 3 viene rimossa la pagina 1 dato che rispetto alla pagina 2 e 0 è quella che verrà utilizzata più tardi. Osserviamo che dalle 15 page fault dell'algoritmo FIFO riusciamo ad arrivare a 9.

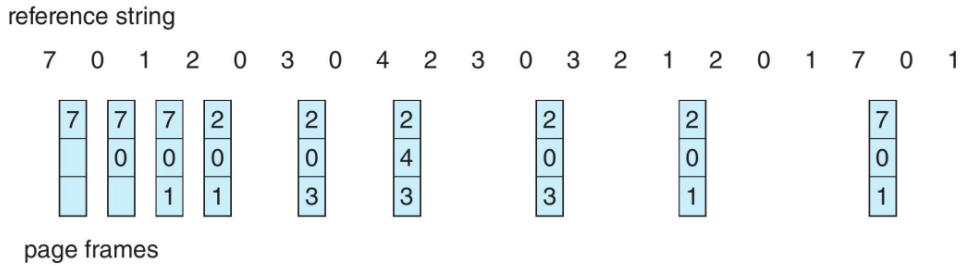


Figura 8.10: Funzionamento teorica dell'algoritmo ottimale.

Dato che questo algoritmo non è realistico, questo serve come **caso limite** o caso ideale verso cui vogliamo che i nostri algoritmi vertano. È impossibile raggiungerci però ci si cerca di avvicinare il più possibile. Inoltre, sappiamo che un algoritmo, basandosi su questa stringa, non potrà avere meno di 9 page fault in quanto il limite è proprio stabilito dall'algoritmo ottimale.

8.3.3 LRU

Dato che non è possibile guardare nel futuro, questo algoritmo guarda il passato, la storia degli utilizzi delle pagine e fa una stima per cercare di capire quale saranno le prossime richieste. È proprio questo il funzionamento dell'algoritmo **LRU**, che sta per **Least Recently Used**, ovvero il frame che è stato utilizzato meno tra quelli presenti nella memoria.

Prendiamo ancora una volta in considerazione la stringa 7 0 1 2 0 3 0 4 2 3 0 3 3 2 1 2 0 1 7 0 1 e vediamo come questo algoritmo si comporta (figura 8.11). Come nei due casi precedenti, le

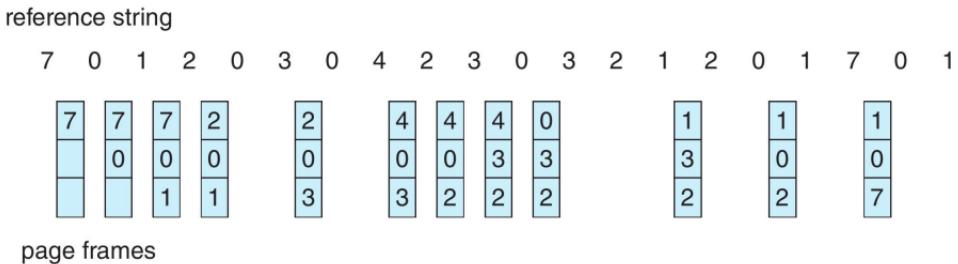


Figura 8.11: Funzionamento dell'algoritmo LRU.

prime quattro richieste sono page fault e la quinta no. Al sesto *timestamp*, quando viene richiesta la pagina 3 l'algoritmo LRU sceglie di scartare il frame 1, questo perché è stato quello utilizzato meno

recentemente rispetto a 0 e 2. A questo punto, la richiesta della pagina 0 non produce alcun page fault però la richiesta successiva, quella per la pagina 4 genera un interrupt che provoca la rimozione della pagina 2 in quanto è stata quella utilizzata meno rispetto alle pagine 0 e 3. Procedendo in questa maniera scopriamo che il numero di page fault è 12, che sono 3 in meno rispetto all'algoritmo FIFO ma rimangono 3 in più rispetto all'algoritmo ottimale.

Com'è possibile implementare questo tipo di algoritmo? Non necessariamente abbiamo a che fare con dei *timestamp* ma possiamo utilizzare un **contatore** mantenuto non dal sistema operativo bensì dall'hardware (come l'MMU), che incrementa il contatore ogni volta che c'è una reference a quella pagina. In questo modo, per scegliere quale pagina deve essere rimpiazzata, si controlla questo contatore e la pagina con il valore minore verrà scartata. Una seconda tecnica per implementare l'algoritmo LRU è attraverso uno **stack**: ogni volta che una pagina viene utilizzata viene impilata in cima allo stack e di conseguenza nel momento in cui si deve scegliere una pagina da rimpiazzare, quella che si troverà nel *bottom* dello stack verrà scelta e scartata. Osserviamo che la seconda soluzione, a differenza del contatore, è una soluzione software piuttosto che hardware.

8.3.4 Second-chance (clock)

Questo algoritmo si basa sull'utilizzo di una **lista circolare** che contiene i riferimenti alle pagine che sono presenti in memoria e ogni qualvolta che ci deve rimpiazzare una pagina in memoria con una nuova pagina appena richiesta si va a scorrere questa lista circolare. Ad ogni pagina caricata in memoria è associato un **bit**. Questo bit è impostato come segue:

- ◊ Il bit viene impostato a 1 ogni volta che la pagina in memoria è utilizzata dal processo;
- ◊ Il bit è impostato da 1 a 0 nel momento in cui la pagina presente non coincide con la pagina richiesta;

L'algoritmo quindi, nel momento in cui è necessario effettuare un rimpiazzo, scorre la lista fino a che non trova un bit a zero, e la pagina associata a quel bit viene rimossa. Tutti i bit a 1 che l'algoritmo ha incontrato prima di arrivare allo 0 sono impostati a zero. In questo modo, se la pagina non viene utilizzata a breve, al prossimo controllo dell'algoritmo questa verrà rimossa dalla memoria.

Facendo quindi riferimento alla figura 8.12, osserviamo che in memoria sono presenti 6 pagine in memoria. In particolare osserviamo che tutte le pagine tranne la numero 3 hanno il bit impostato ad 1: ciò significa che tutte le pagine, eccetto la terza, nel momento in cui è necessario rimpiazzare una

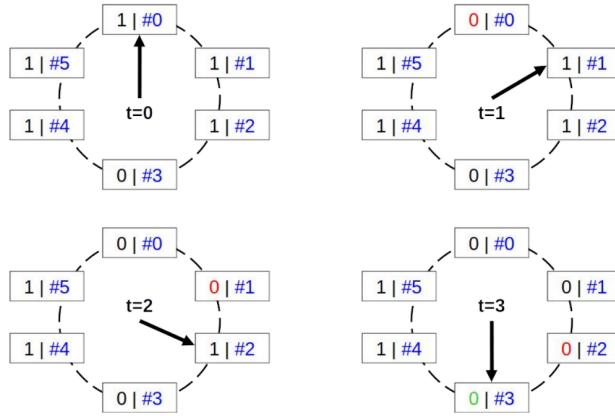


Figura 8.12: Il funzionamento del second-chance algorithm.

pagina l'algoritmo parte dalla pagina 0 e comincia a scorrere. La pagina 0 è stata utilizzata poco prima perché il bit è a 1: di conseguenza l'algoritmo imposta il bit a zero e procede alla pagina seguente. Sia per la pagina 1 che per la pagina 2 l'algoritmo le lascia in memoria ma il bit diventa zero. Alla pagina numero 3 però il bit è già a zero, di conseguenza tale pagina verrà rimossa al fine di lasciare spazio a quelle nuove.

È infine possibile migliorare questo approccio inserendo un secondo bit, chiamato **modify bit** che indica se la pagina in memoria è stata modificata o meno. Si creano quindi quattro combinazioni possibili, e in base a queste l'algoritmo effettua delle azioni sulla pagina (`reference`, `modify`):

- ◊ $(0, 0)$ indica che la pagina non è stata utilizzata recentemente e nemmeno modificata: è quindi la scelta migliore.
- ◊ $(0, 1)$ segnala che la pagina non è stata utilizzata di recente ma è stata modificata, di conseguenza, nel momento in cui la si rimpiazza, è necessario salvare le modifiche nel backing store.
- ◊ $(1, 0)$ indica che la pagina non è stata modificata ma è stata utilizzata di recente.
- ◊ $(1, 1)$ indica che la pagina è sia stata utilizzata di recente che modificata.

8.3.5 Algoritmo counting

Oltre all'algoritmo LRU sono presenti anche altri tipi di approcci che si basano sul numero di utilizzi di una pagina. Le altre due opzioni che vediamo sono algoritmi che vanno a contare il numero di referenza fatte ad ogni singola pagina. A differenza quindi dell'algoritmo second-chance non si ha un bit che indica se la pagina è stata utilizzata di recente ma si ha un vero e proprio **contatore** che tiene traccia del numero di referenze alla determinata pagina in memoria.

Con questa informazione si possono quindi andare a rimuovere le pagine che sono state usate di meno, come nel caso dell'algoritmo **LFU** (*Least Frequently Used*). Con questo approccio però può capitare che una pagina rimanga molto in memoria, anche se inutilizzata: poniamo il caso di una pagina nuova che entra in memoria e viene usata molto spesso. Questo significa che il contatore cresce subito e rimane alto. Anche se la pagina dopo molto tempo non è utilizzata, l'algoritmo non la rimuoverà mai dato che il contatore è molto alto. Si è quindi deciso di avere un check periodico dove si va a dimezzare il valore del contatore alle pagine che non sono utilizzate molto. In questo modo, prima o poi quella pagina in memoria se continua ad essere inutilizzata verrà rimossa appena il contatore diventa sufficientemente piccolo.

Un secondo approccio è quello opposto al LFU: stiamo infatti parlando del **MFU** (*Most Frequently Used*), dove si va a rimuovere la pagina che ha il maggior numero di utilizzi.

8.3.6 Ottimizzazione

Oltre alle performance che si ottengono attraverso un algoritmo di page replacement si può cercare di ottimizzare l'algoritmo anche in altri momenti durante la procedura di rimpiazzo di una pagina. Per esempio è possibile implementare un insieme di **frame** sempre **liberi** (una sorta di buffer) al fine di allocare subito la pagina richiesta nel momento in cui si verifichi un page fault. Solo dopo aver soddisfatto la richiesta ci si preoccupa di rimuovere una pagina dalla memoria. In questo modo il processo viene immediatamente servito e non viene sprecato tempo prezioso al fine di scambiare le pagine dalla memoria la backing store.

È possibile inoltre implementare una **lista delle pagine modificate**. Oltre al modify bit (8.3.4), questa scelta è opportuna in quanto le pagine modificate contenute al suo interno verranno periodicamente salvate all'interno del backing store. In questo modo, una volta che sono state salvate la lista sarà vuota e il modify bit ritornerà a zero.

8.4 Allocazione dei frames

Facciamo ora anche qualche considerazione riguardante il numero minimo e massimo di frames da allocare a ogni singolo processo. Ci sono diversi modi di allocare i frames per ciascun processo: poniamo di far partire 5 processi e che abbiamo a disposizione 100 frames. Possiamo scegliere di allocare i frames in maniera **uniforme**, dove ogni processo ha a disposizione 20 frames oppure, eventualmente, allocarne 15 per processo e tenere una *pool* di 25 frames liberi. Altrimenti è possibile effettuare un'**allocazione proporzionale**: il numero di frames allocati per processo dipende dalla dimensione e dalla priorità di quest'ultimo. È infatti ragionevole che un processo di grandi dimensioni richiede un numero maggiore di frames rispetto ad un processo di dimensioni minori.

8.4.1 Allocazione globale e locale

Quando andiamo a scegliere il frame da liberare al fine di fare spazio a nuove pagine, si possono scegliere due approcci.

Con l'approccio **globale** il frame da liberare viene scelto tra tutti i frame compresi in memoria, non importa a quale processo appartiene. In generale questo approccio è più performante rispetto al secondo e per questo è più utilizzato.

Con l'approccio **locale** invece, per ogni processo viene definito un insieme di frames da dove è possibile andare a pescare. In questo caso invece le performance sono ridotte in quanto è possibile che ci siano frame liberi al di fuori dell'insieme scelto per il processo ma questi non potranno essere usati proprio perché non compresi nel frame. In questo caso però il vantaggio è che è più consistente nei tempi di risposta dato che sono presenti dei limiti ben definiti per allocare i frames.

8.4.2 Richiesta delle pagine

Al fine di riuscire ad implementare il *global page replacement*, ovvero le classiche politiche di rimpiazzo, è quello di impostare dei limiti inferiori e superiori (*thresholds*) al numero di frames liberi in memoria. Osservando la figura 8.13, osserviamo che nel momento in cui il numero di frame disponibili scende sotto il limite inferiore, il sistema operativo esegue un algoritmo di page replacement al fine di liberare abbastanza frames da raggiungere il limite superiore. Nel caso in cui il sistema operativo non riesca a liberare la memoria, può scegliere di cambiare algoritmo di page replacement, magari scegliendone

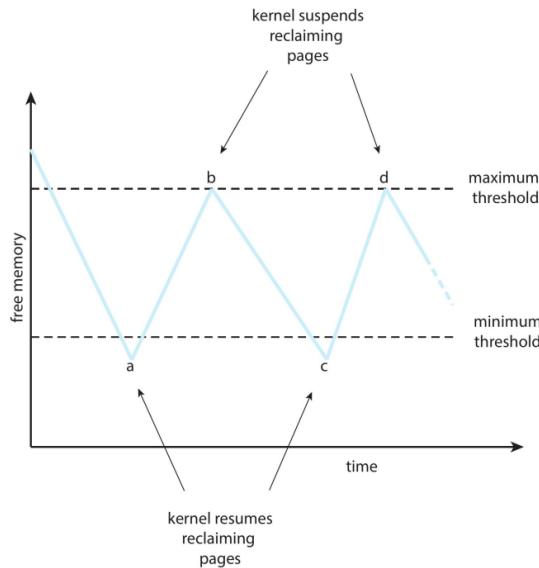


Figura 8.13: Questa tecnica assicura la presenza di frames liberi in memoria al fine di soddisfare eventuali nuove richieste.

uno più *strict*, oppure scegliere proprio di terminare qualche processo, potenzialmente quelli che consumano più memoria.

8.5 Thrashing

Il thrashing è un aspetto importante, soprattutto perché diversi metodi, ormai superati, di gestione della memoria, non sono riusciti a gestire questo problema. La figura 8.14 contiene uno schema dove viene mostrato l'utilizzo della CPU in relazione al numero di processi presenti in memoria. Teoricamente l'obiettivo è quello di aumentare il numero di processi il più possibile al fine di aumentare l'utilizzo della CPU il più possibile. Notiamo in realtà che ad un certo punto l'utilizzo della CPU cala drasticamente. Questo generalmente capita quando il numero di processi in memoria è molto alto e quindi il numero di frames disponibili per processo diminuisce. Di conseguenza aumenta il numero di *page fault* (8.2.1) e quindi è necessario effettuare uno *swapping* (7.5). Lo swapping però non è effettuato dal processore, di conseguenza l'utilizzo della CPU diminuisce. Se l'utilizzo diminuisce allora il sistema operativo pensa che il processore sia pronto ad eseguire altri processi. Questi processi

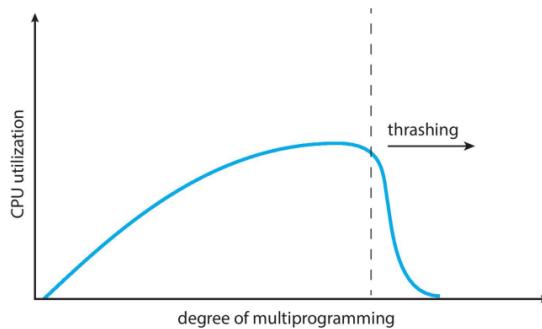


Figura 8.14: Il fenomeno del thrashing.

vengono quindi caricati in memoria, però non ci sono frames disponibili per loro e di conseguenza è necessario rimpiazzare delle pagine mediante uno swapping. Si entra quindi in un loop dove la CPU non viene utilizzata e tutti i processi non vengono eseguiti dato che si devono swappare tra di loro.

8.5.1 Modello Working-set

Al fine di cercare di evitare fenomeni di thrashing si può utilizzare una caratteristica che hanno i programmi, chiamata **locality model**. Questa caratteristica dei programmi dice che tipicamente questi sono strutturati in modo tale che istruzioni e dati risiedi, sono raccolti, in zone di memoria **adiacenti**. Cerchiamo quindi di trovare un apporccio che sfrutti al meglio questo principio di località dei programmi.

Iniziamo definendo con Δ la finestra temporale dove avviene un certo numero di riferimenti alle pagine. Chiamiamo inoltre **WS** in *working-set*, ovvero l'insieme di pagine che sono utilizzate all'interno di Δ (vedi figura 8.15, dove la window è di 10 accessi). Definiamo ora un'altra variabile, la *Working-set size*, **WSS**, che rappresenta il numero di pagine utilizzate durante la finestra Δ . In particolare possiamo dire che $WSS(t_1) = 5$ e che $WSS(t_2) = 2$. Con queste informazioni possiamo constatare che:

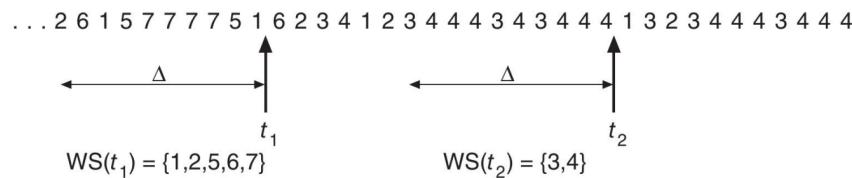


Figura 8.15: Il working-set.

size, **WSS**, che rappresenta il numero di pagine utilizzate durante la finestra Δ . In particolare possiamo dire che $WSS(t_1) = 5$ e che $WSS(t_2) = 2$. Con queste informazioni possiamo constatare che:

- ◊ Se Δ è molto piccolo allora non sarà possibile comprendere tutte le pagine che fanno parte della locality;
- ◊ Se Δ è molto grande è molto probabile che verranno comprese, oltre alla locality del processo, anche locality che non sono attive, andando quindi a sprecare memoria.
- ◊ Se $\Delta \rightarrow \infty$ si andrebbero a considerare tutte le pagine del processo.

Chiamiamo ora con $D = \sum WSS_i$, ovvero la somma della dimensione dei working set di tutti i processi. Se D è maggiore del numero dei frames disponibili allora si è in una situazione di thrashing. Ogni sistema operativo tiene traccia di questi WSS_i al fine di evitare situazioni di thrashing, di conseguenza se vede che ci si sta per avvicinare al thrashing si occupa di terminare o mettere in pausa un processo.

8.5.2 Frequenza dei page fault (PFF)

Un'alternativa alla WSS è quella di andare a manipolare la frequenza dei page fault. Questo approccio, a differenza del WSS offre una soluzione meno drastica e cerca di prevenire il verificarsi del thrashing (figura 8.16). In particolare, se la frequenza dei page fault è molto alta significa che ci sono diversi processi che stanno accedendo alla memoria, di conseguenza è necessario ridurre il numero di processi. Viceversa, se la frequenza di page fault è molto bassa, ciò significa che è possibile andare

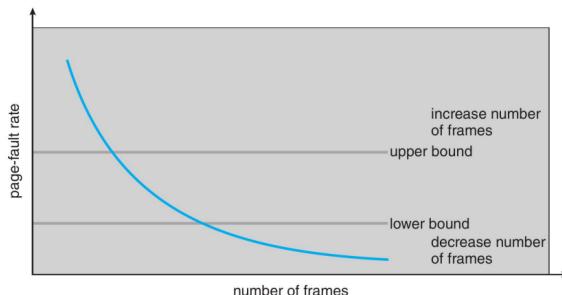


Figura 8.16: Il controllo della frequenza dei page fault

ad aggiungere processi in memoria.

Legame tra WS e PFF

Come è possibile vedere anche dalla figura 8.17, esiste una relazione tra il working-set e il page fault rate. Notiamo che sono presenti dei picchi di frequenza che piano piano si riducono. Questi picchi

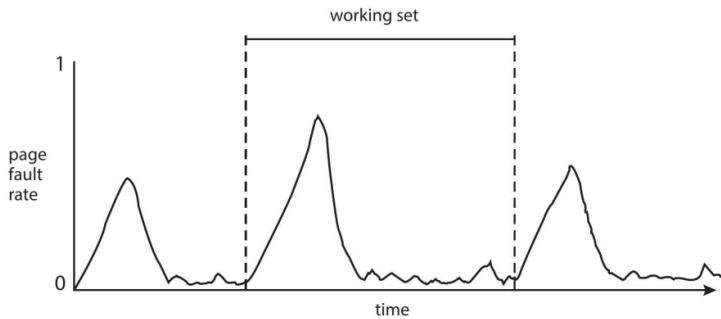


Figura 8.17: La relazione tra working-set e page fault rate

rappresentano il cambiamento di locality: quando il processo cambia da una locality ad un'altra, per un certo tempo si genereranno page fault fino a che tutti i frames necessari sono caricati in memoria.

8.6 Allocare la memoria del kernel

Fino ad ora abbiamo visto la gestione della memoria in **modalità utente**. Ovviamente sono presenti anche dei modi dedicati al fine di allocare la memoria da parte del kernel del sistema operativo. Essendo infatti una parte molto importante e delicata, l'accesso in memoria non può essere rallentato a causa di altri processi: ha infatti un accesso privilegiato ed efficiente alla memoria. Sono presenti due strategie al fine di riuscire ad allocare la memoria da parte del kernel: il *buddy system* e la *slab allocation*.

8.6.1 Buddy system

Con questo approccio la memoria è allocata da un segmento di dimensione fissata che consiste di pagine contigue (figura 8.18). La memoria viene allocata utilizzando il **power-of-2-allocator**: se, per esempio, abbiamo 256KB di segmento in memoria e il kernel ne richiede 31KB, i 26KB vengono divisi per due fino a che non si arriva al blocco ideale per la richiesta effettuata dal kernel. Se la richiesta fosse stata 33KB, si sarebbe dovuta allocare in un blocco da 64KB. Lo svantaggio sta proprio qui, dove lo spazio in memoria raramente è utilizzato ottimamente in quanto tipicamente le richieste del kernel non sono potenze di due: è il famoso fenomeno della frammentazione interna (7.2.4).

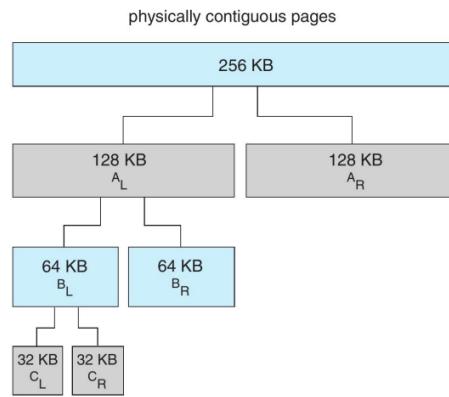


Figura 8.18: L'approccio di allocazione buddy system con l'allocatore *power-of-2-allocator*.

8.6.2 Slab allocation

Il secondo approccio è l'allocazione a "lastre". Questo metodo raggruppa delle pagine in memoria fisicamente contigue in **slab**; inoltre, gruppi di slabs costituiscono una **cache** (figura 8.19). Ad ogni

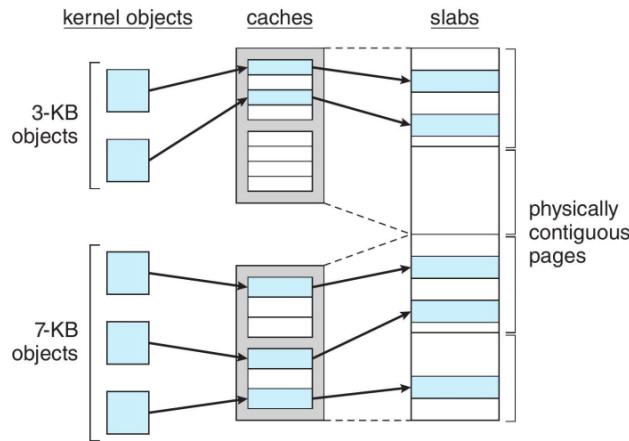


Figura 8.19: La slab allocation del kernel.

struttura dati che il kernel utilizza è associata una cache. Sarà quindi presente una cache per gestire la lista dei processi (lista dei PCB 2.1.1), una cache per gestire la lista delle pagine e così via. Una volta che il kernel definisce le strutture che vengono usate, vengono allocate le cache volte a contenerle ed, eventualmente, ne incrementa la dimensione nel caso in cui il numero di istanze nelle strutture dati

aumenti. Uno dei benefici di questa tecnica è che non si verifica la frammentazione ma comunque l'accesso rimane veloce.

9

MEMORIA DI MASSA

In questo capitolo iniziamo ad occuparci della gestione della memoria secondaria. Discuteremo innanzitutto il tipo di periferiche con il quale andremo a lavorare per poi passare ad alcuni algoritmi per lo scheduling e l'utilizzo ottimale di tali periferiche. Passeremo poi ad approfondire alcuni aspetti sulla gestione di questi dispositivi e del loro spazio di archiviazione (in particolare dello *swapping*, discusso già nel paragrafo 7.5). Infine discuteremo su alcuni metodi per l'immagazzinamento di molti dati, come le strutture RAID.

9.1 Tipi di memoria secondaria

Iniziamo con il distinguere due tipi di supporti di dati più comuni: stiamo parlando dei dischi rigidi, chiamati anche HDD, e delle schede basate su una tecnologia moderna chiamate NVM

9.1.1 HDD

L'HDD, che sta per *Hard Disk Device*, è quello che comunemente viene chiamato **disco rigido**. Questa unità è composta da diversi dischi magnetici, uno sopra l'altro. Facendo riferimento all'immagine 9.1, possiamo notare che ogni disco è formato da delle **tracce**, ovvero il cerchio a distanza \bar{R} dal centro. Ogni traccia è composta da diversi **settori**. L'insieme delle tracce a distanza \bar{R} dal centro del disco è detto **cilindro**. Al fine di riuscire a leggere le informazioni dei dischi, questi girano e una "testina" (*arm*, ce

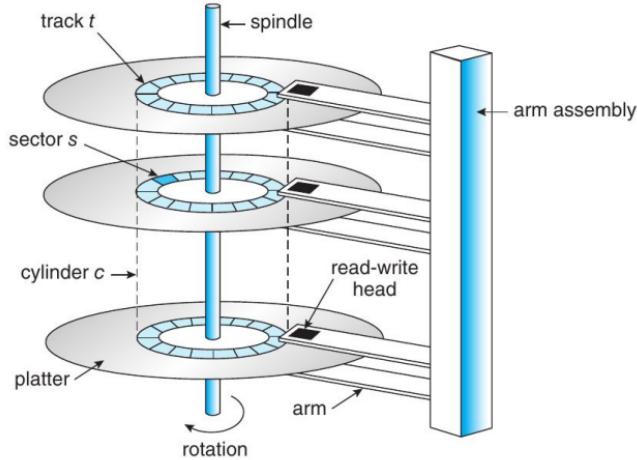


Figura 9.1: Schema ad alto livello di un disco rigido.

n'è una per ogni disco) si appoggia sulla traccia e legge l'informazione.

Essendo uno spostamento fisico, la lettura dal disco è abbastanza lenta in quanto è necessario che la testina raggiunga la traccia del disco necessaria, a questo punto il disco gira fino a che non si è raggiunta l'informazione cercata. Il tempo di posizionamento sulla traccia (o settore) è detto **seek time** mentre il tempo dovuto alla rotazione del disco è detto **rotational latency**. Possiamo quindi cercare di calcolare la performance di un HDD attraverso delle medie aritmetiche:

- ◊ *average access time*: il tempo medio di accesso si calcola sommando l'*average seek time* e l'*average latency*;
- ◊ *average I/O time* si calcola sommando il valore trovato nel punto precedente a al tempo di trasferimento e all'*overhead* di un potenziale controllore elettrico/periferica:

$$AVG_Access_Time + \frac{Amount_to_Transfer}{Transfer_rate} + Controller_Overhead.$$

Esercizio

Prendiamo in considerazione la situazione di dover trasferire un blocco da 4KB su un disco che gira a 7200 RPM (*rotation per minute*) con un *average seek time* di 5 millisecondi e un *transfer rate* di 1Gigabit/sec con un *overhead* del controllore di 0.1 millisecondi. Seguendo la formula del secondo punto, dobbiamo trovare *average Access Time* e *Transfer Time* dato che il *Controller Overhead* già lo conosciamo.

- ◊ *Average Access Time* = *average Seek Time* (=5ms) + *Average Latency*. Per calcolare quest'ultima si fa: $\frac{1}{2} \cdot \frac{60}{7200} \cdot 10^3 = 10^3 \cdot \frac{30}{7200}$. Di conseguenza l'*Average Access Time* diventa: $5 + 4.17\text{ms} = 9.17\text{ms}$.
- ◊
$$\text{Transfer Time} = \frac{\text{Amount to Transfer}}{\text{Transfer rate}} = \frac{4KB}{1Gb/sec} = \frac{4KB}{\frac{1}{8}GB/sec} = \frac{32KB}{1024^2 KB/sec} = 0.031\text{ms}$$

A questo punto sommiamo i due risultati all'*overhead* e otteniamo 9.301ms.

9.1.2 NVM

I dispositivi NVM (*NonVolatile Memory*, ovvero memoria non volatile) sono anche chiamati **SSD**, che sta per *Solid-State Disk*. Questi dispositivi hanno lo stesso compito dei dischi rigidi, ovvero quello di immagazzinare grandi moli di dati. In questo caso cambia il metodo di implementazione di tali dispositivi: non sono infatti presenti parti meccaniche e quindi sono molto più veloci nell'accesso e quindi anche più affidabili nell'**accesso random**.



Questa velocità molto elevata va però a scapito sia della capacità di *storage* che anche della durabilità del dispositivo. Questo perché le SSD sono molto veloci in lettura ma presentano enormi difficoltà e complicazioni in scrittura. Non è infatti possibile sovrascrivere un'informazione ma è possibile solo cancellarla e poi scrivere la nuova informazione. Inoltre la cancellazione è effettuata per **blocchi** di pagine e non pagine singole, di conseguenza, per sovrascrivere una pagina è necessario cancellare l'intero blocco e riscriverlo con la pagina modificata correttamente. La cancellazione inoltre usura il dispositivo: ecco che si cerca di misurare la "vita" di una SSD attraverso i **DWPD** (*Drive Writes Per Day*), ovvero le scritture effettuate in un giorno.

Generalmente ci si trova in una situazione dove in un blocco di pagine sono presenti dei dati che sono validi e dei dati che non lo sono (come in figura 9.2). Alcuni dati infatti saranno aggiornati mentre degli altri dati sono più vecchi e non hanno più valore, di conseguenza sono dei candidati per essere rimpiazzati nel caso in cui fosse necessario dello spazio: è quindi presente un algoritmo di **garbage collection** che permette proprio di rimuovere delle pagine datate. Cosa succede però se tutti i blocchi hanno una o più pagine scritte? È necessario di un buffer dove il garbage collector sposta le pagine temporaneamente in modo da liberare lo spazio per poi ricopiare le pagine valide. Tipicamente una parte della NVM (circa il 20%) è lasciata a disposizione del buffer per il garbage collector (*overprovisioning*). Nel momento in cui arriva una nuova richiesta in scrittura e non c'è spazio,

valid page	valid page	invalid page	invalid page
invalid page	valid page	invalid page	valid page

Figura 9.2: Un blocco con delle pagine valide e delle pagine non valide.

il garbage collector andrà a prendere le pagine e salvarle nel buffer, andrà a cancellare il blocco per poi ricopiare le pagine nuove nel blocco.

9.1.3 Memoria volatile

Sono presenti casi, dove al fine di implementare dei file systems, sono necessario le memoria volatili, ovvero memoria che non mantiene salvati i dati quando manca l'alimentazione (ovvero quando il computer si spegne). Stiamo parlando dei **RAM drives**, che sono implementati attraverso memorie mantenute in vita solo quando l'alimentazione è disponibile. Tali memorie possono essere utilizzate per varie operazioni sfruttando il tempo di accesso molto rapido, come l'utilizzo di file temporanei.

9.1.4 Nastro magnetico

Un'altra tecnologia meno comune nei computer di tutti i giorni ma che comunque rimane importante per altri dispositivi, come i server, è il cosiddetto nastro magnetico. Questa è un'altra forma di memorizzazione su cui grandi moli di dati vengono immagazzinati, soprattutto per fare un **backup**. In questi nastri l'accesso ai dati è effettuato in modo **sequenziale**: se dobbiamo accedere ad una particolare zona del nastro è necessario scorrere tutto il nastro fino a che non si raggiunge la zona desiderata. È ovviamente molto inefficiente per un accesso di tipo random ma è efficiente per accessi sequenziali; proprio per questo motivo è ottimo per i backup.

9.1.5 Dispositivi di memorizzazione esterna

Discutiamo ora, brevemente, altre periferiche le quali possono essere collegate temporaneamente attraverso dei connettori (dei *bus*) al computer. Le periferiche più famosi sono l'ATA, in particolare il SATA. Per invece per quanto riguarda i dispositivi NVM la loro velocità di trasferimento ha stimolato la creazione di connettori con standard più veloci

9.2 Indirizzamento

Nonostante le tecnologie sono implementate diversamente l'una dall'altra, dal punto di vista del computer i dispositivi sono molti simili, anche se i dispositivi fisici sono completamente diversi. Indipendentemente dal fatto che un file venga salvato su una SSD oppure su un HDD, questo verrà salvato sia su una periferica che sull'altra in quanto il computer non nota differenze tra un dispositivo e l'altro. È un concetto molto simile all'indirizzo logico e fisico per i processi, dove questi ultimi vedevano solo i loro indirizzi relativi senza sapere quale spazio in memoria sarebbe stato occupato. Analogamente il sistema operativo salva il file all'interno del dispositivo senza sapere come questo sia fatto. Sarà infatti compito del dispositivi tradurre i segnali dal sistema operativo in operazioni da effettuare nella specifica periferica.

Per esempio, la scrittura su un HDD potrebbe essere rappresentata logicamente come la scrittura su un array unico. Questo array è definito inserendo in maniera contigua tutti i cilindri di una traccia per poi proseguire con il cilindro della traccia più interna. IN questo modo tutti i dati presenti sui vari livelli del disco rigido sono rappresentati mediante un array.

Avere una conoscenza di come il dispositivo funziona è comunque utile per il sistema operativo al fine di riuscire ad ottimizzare le operazioni da effettuare sulla periferica di archiviazione. Queste informazioni sono salvata nella **LBA**, ovvero il *Logical Block Address*.

9.3 HDD scheduling

Tipicamente non c'è mai solo un processo che vuole accedere al disco ma ce ne sono di diversi. Di conseguenza il sistema operativo mantiene delle code che contengono i processi in attesa per l'accesso al disco. Nel momento in cui sono presenti delle code di attesa, sicuramente saranno presenti degli

algoritmi che scelgono quale processo andrà ad effettuare l'accesso al disco. In questo caso si può sfruttare il fatto che spesso indirizzi LBA vicini corrispondono anche a indirizzi fisici vicini tra loro.

9.3.1 FCFS

Il primo algoritmo che andiamo a vedere è, di consueto, un algoritmo FIFO. Infatti, proprio come nel CPU scheduling (4), FCFS sta per *First Come First Served*. Con questo algoritmo si ha quindi una coda

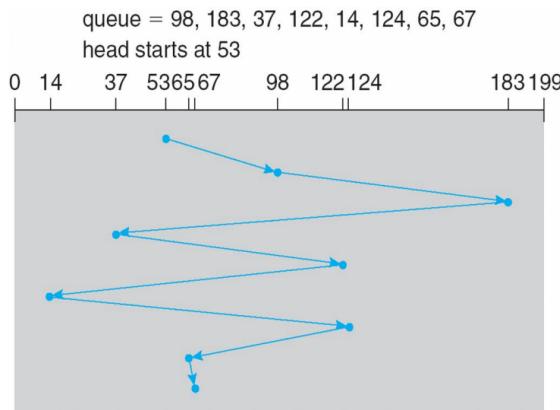


Figura 9.3: Funzionamento dell'algoritmo FCFS.

FIFO dove il primo processo che entra nella coda (e quindi fa la richiesta di accesso) sarà anche il primo ad uscire. Osserviamo la figura 9.3 e ipotizziamo che arrivi una serie di richieste che corrisponda ad andare nella traccia 98, poi nella traccia 183 e così via fino alla traccia 67. Ipotizziamo di partire dalla traccia 53: andare a soddisfare le richieste in ordine FIFO significa spostare la testina lungo il disco e spostarsi da una traccia all'altra. Possiamo notare subito che questo algoritmo non tiene conto in alcun modo della prossimità delle tracce tra di loro, ma si basa solamente sull'ordine di entrata. Calcolando il numero di tracce che ha dovuto scorrere per completare la coda, scopriamo che il totale è di 640 tracce.

9.3.2 SSTF

Un possibile miglioramento è il cosiddetto SSTF, che sta per *Shortest Seek-Time First*: è un algoritmo simile al SJF del CPU scheduling (4.2.2), dove si va a scegliere il processo che ha un burst time più basso. Analogamente in questo caso, si va a scegliere il processo meno distante rispetto alla posizione della testina, di conseguenza si va a soddisfare la richiesta che ci permette di saltare meno tracce.

Osservando la figura 9.4, notiamo che l'ordine di esecuzione è diverso dal precedente in quanto dalla traccia 53 si passa alla 65, poi alla 67 e così via. In questo caso la distanza totale percorsa (in tracce) è di 236, che è un terzo rispetto a quella del FCFS. Proprio come nel caso dell'algoritmo SJF ci possono

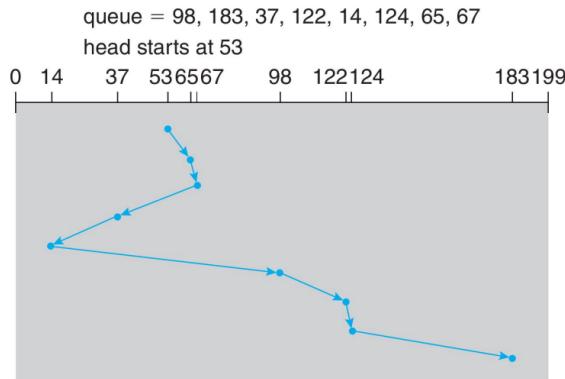


Figura 9.4: Funzionamento dell'algoritmo SSTF.

essere dei problemi di **starvation**, ovvero che arrivino sempre richieste vicine tra loro che vengono immediatamente soddisfatte trascurando una richiesta che è più lontana che potenzialmente non verrebbe mai eseguita. Ciò nonostante l'algoritmo rimane migliore rispetto al FCFS però, dato che si possono verificare questo tipo di situazioni, rimane un algoritmo non ottimale.

9.3.3 SCAN e C-SCAN

Al fine di evitare la starvation si è implementato l'algoritmo SCAN. Questo algoritmo è molto semplice: la testina va avanti e indietro lungo il disco e, mentre lo fa, soddisfa tutte le richieste che arrivano (figura 9.5). Il problema di questo algoritmo è che si ha uno soddisfacimento non equo delle richieste. Poniamo di trovarci nella traccia 10 e stiamo per arrivare alla 0 per poi tornare indietro. Se arriva una richiesta alla traccia 160 e dopo arrivano delle richieste alla traccia 31, 41, e 59, queste richieste verranno soddisfatte prima della richiesta alla traccia 160 la quale deve attendere più tempo anche se è arrivata prima delle altre.

Per questo motivo si è implementato un altro algoritmo, molto simile, chiamato **Circular-SCAN** (C-SCAN). Questo algoritmo, al posto di andare avanti e indietro come l'algoritmo SCAN, una volta che arriva alla fine, alla traccia 199, non torna indietro soddisfacendo altre richieste ma va subito alla traccia

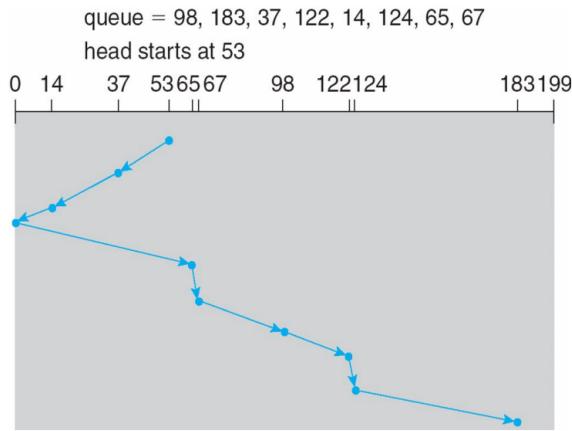


Figura 9.5: Funzionamento dell'algoritmo SCAN.

0 e ricomincia (figura 9.6). In questo caso però, anche se le richieste sono soddisfatte in maniera più equa, il tempo di ricerca totale è nettamente maggiore rispetto a quello dell'algoritmo SCAN.

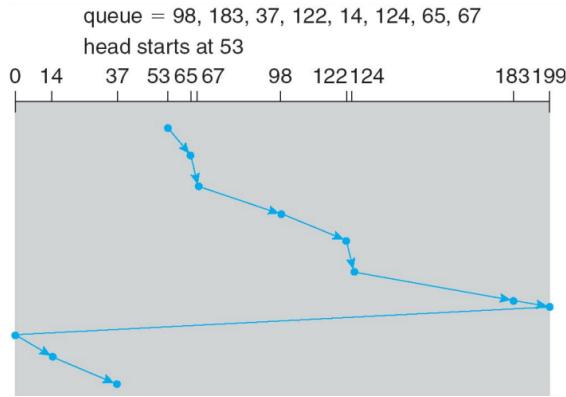


Figura 9.6: Funzionamento dell'algoritmo C-SCAN.

9.3.4 Scelta dell'algoritmo

La scelta dell'algoritmo da utilizzare dipende sicuramente dal tipo di applicazione che si usano. Quelli discussi in questo capitolo sono gli algoritmi utilizzati più comunemente nei sistemi operativi. Generalmente SSTF è molto più utilizzato rispetto al FCFS mentre gli algoritmi SCAN e C-SCAN sono utilizzati nei sistemi con un alto carico di dati da memorizzare.

10 SISTEMA INPUT/OUTPUT

Occupiamoci ora di un altro argomento molto importante. Stiamo parlando del sistema Input/Output che, nella maggior parte dei casi, è fondamentale per l'utilizzo dei computer. Spesso in fatto lo si utilizza non per la sua capacità computazionale ma per scrivere documenti, leggere dei *files* da dispositivi, connettersi e navigare su internet oppure guardare un video. Tutte queste operazioni non hanno a che vedere con la potenza di calcolo del computer ma dalla capacità che ha quest'ultimo di integrare e connettere dispositivi esterni come periferiche di memorizzazione, schermi, tastiera, mouse, dispositivi per la connessione internet e molto altro. È proprio di questo che ci occuperemo in questo capitolo, di come il sistema operativo sia in grado di integrare questi dispositivi e riuscire a gestirli in maniera ottimale.

10.1 Componenti hardware

Passiamo ora all'analisi dei principali componenti HW. Come accennato in precedenza, possiamo spaziare da periferiche per la memorizzazione di dati, per la trasmissione oppure per la comunicazione con l'uomo (monitor, tastiera, mouse, etc).

Sono quindi presenti delle **porte**, ovvero i punti di connessione tra il dispositivo esterno e il computer. Internamente al calcolatore abbiamo dei **bus** che servono per la comunicazione tra la periferica e il computer o tra due periferiche. Tra i bus citiamo il bus **PCI/PCIe**, comunemente usati per la comunicazione tra schede e periferiche e sono ad elevate prestazioni (velocità e mole di dati). In

secondo luogo abbiamo gli *expansion bus* che servono per connettere periferiche più lente e le **SAS** (*Serial Attached SCSI*) che sono comuni per i dischi. Osserviamo dalla figura 10.1 che ci sono dei bus che sono direttamente collegati alla memoria tramite PCI mentre l'expansion bus è utilizzato per delle periferiche più lente.

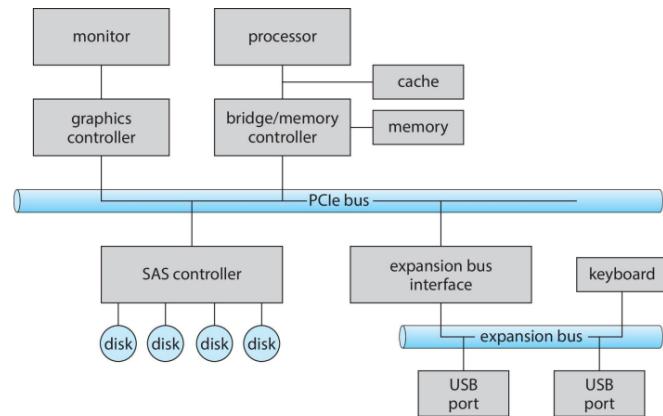


Figura 10.1: Illustrazione sui collegamenti tra periferiche.

Sempre dall'immagine 10.1 si osserva che i collegamenti sono generalmente gestiti da dei **controller**, che possono essere sia delle schede dedicate che integrate nel processore che interfacciano un componente con un altro.

10.2 Tecniche di comunicazione

La comunicazione viene attraverso delle istruzioni che sono scambiate con un set di registri che possono essere di quattro tipi:

- ◊ *Data-in* e *data-out* dove si possono rispettivamente ricevere e scrivere dati;
- ◊ *Stato*, sono dei registri di stato che indicano se, per esempio, la comunicazione è stata effettuata correttamente o meno;
- ◊ *Controllo* che servono effettivamente per inizializzare o modificare la comunicazione.

Questi registri, tipicamente vanno da 1 a 4 Byte. Spesso, al posto di fornire solamente un accesso al signolo registro, la periferica fornisce dei buffer Fifo dove viene incrementata la dimensione dei dati che è possibile leggere o scrivere.

La trasmissione può essere effettuata mappando l'I/O del dispositivo nello spazio di indirizzamento del sistema operativo. In questo modo è possibile trasmettere comandi o dati sui registri andando a scrivere o leggere su indirizzi che sono uguali agli indirizzi di memoria.

10.2.1 Polling

Scaviamo più a fondo: come avviene la comunicazione input/output tra computer e una periferica? Il primo metodo che andiamo a vedere è quello dell'utilizzo di bit di diverso tipo:

- ◊ *busy* bit, che indica se la periferica è libera o meno;
- ◊ *read* o *write* bit;
- ◊ *command-ready* bit che controlla se ci sono dei comandi pronti ad essere eseguiti;

Segue un esempio di comunicazione con questa tecnica. Si controlla inizialmente se la periferica è occupata o meno attraverso il *busy* bit. Nel caso la periferica è libera, imposta il modo di comunicazione attraverso il *read/write* bit per poi impostare un comando e segnalare al controller che tale comando è pronto ad essere eseguito. A questo punto il controllore prenderà in considerazione il comando, imposta il *busy* bit ad 1 e inizia l'esecuzione di tale comando.

Osserviamo che con il *busy* bit si può verificare una situazione di *busy-waiting*: il sistema operativo non può sempre controllare che una determinata periferica sia libera o meno. Se su 10000 volte, solo una volta il dispositivo è libero, si perde molto tempo per effettuare le 9999 richieste.

10.2.2 Interrupt

Una soluzione sicuramente più ottimale comprende l'utilizzo di **interrupt**. In altre parole il sistema operativo effettua una richiesta ad una periferica, questa, appena sarà disponibile, la prenderà in considerazione e la completerà. Una volta completata, la periferica informa il sistema operativo attraverso un interrupt. Osserviamo che il sistema operativo non rimane fermo in attesa ma fa altre task per poi essere informato del soddisfacimento della richiesta. Una volta arrivato l'interrupt, questo è gestito dall'**interrupt handler** che è una *routine*. Possiamo dividere gli interrupt in due categorie:

- ◊ Mascherabili (**maskable**) ovvero interrupt non critici e che possono essere messi in attesa;

- ◊ Non mascherabili (**nonmaskable**) che hanno una priorità maggiore (device non trovato) e che quindi devo essere serviti il prima possibile dalla CPU.

Gli interrupt non mascherabili, in quanto critici, sono divisi in molte sotto categorie, ognuna della quali deve essere gestita in maniera particolare. Ecco quindi che si è creata una tabella, l'**interrupt vector**, che per ogni tipo di interrupt assegna il numero dell'interrupt e l'interrupt *handler* che lo andrà a gestire. Dalla tabella in figura 10.2 osserviamo un esempio di interrupt vector: i primi 31 sono quelli

vector number	description
0	divide error
1	debug exception
2	null interrupt
3	breakpoint
4	INTO-detected overflow
5	bound range exception
6	invalid opcode
7	device not available
8	double fault
9	coprocessor segment overrun (reserved)
10	invalid task state segment
11	segment not present
12	stack fault
13	general protection
14	page fault
15	(Intel reserved, do not use)
16	floating-point error
17	alignment check
18	machine check
19–31	(Intel reserved, do not use)
32–255	maskable interrupts

Figura 10.2: Esempio della tabella di interrupt (Intel Pentium).

non mascherabili e devono essere gestiti all'istante. Notiamo che tra i primi 31 è presente il *page fault* (8.2.1).

Si sottolinea, infine, che alcuni sistemi operativi come il *Solaris*, sono in grado di gestire gli interrupt in maniera concorrente e parallela.

10.2.3 DMA

Un uso intelligente dell'interrupt si ha nel DMA (*Direct Memory Access*), che è quel controllore ormai integrato nei processori sessi che aiuta a gestire la comunicazione con *devices* che richiedono un grosso flusso di scambio dati. Non siamo parlando di mouse e tastiera bensì di dischi rigidi o comunque in genere uno *storage* secondario.

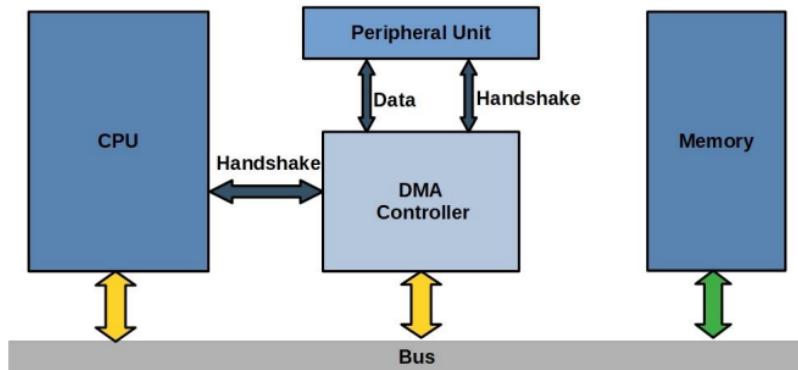


Figura 10.3: Schema di funzionamento del DMA.

Come sappiamo, la CPU non usa i dati da qualsiasi periferica ma solo ed esclusivamente dalla memoria¹. Di conseguenza un dato di un dispositivo esterno, al fine di essere utilizzato dal processore, deve innanzitutto passare dalla memoria. Conviene quindi lasciare questo compito al DMA che prende temporaneamente il controllo del bus tra CPU e memoria e gestisce la sincronizzazione con la periferica trasferendo i dati verso la periferica o verso la memoria (vedi figura 10.3). Nei sistemi di DMA più avanzati viene implementato anche lo scambio di dati tra periferica secondaria e memoria virtuale: si parla infatti di **DVMA**.

Più nel dettaglio, come fa il sistema operativo a sfruttare la potenza del DMA? Per prima cosa è necessario scrivere il comando su un blocco di memoria. In questo blocco di memoria, devono essere specificate le seguenti informazioni:

- ◊ *Sorgente* (a quale periferica prelevare i dati) e la *destinazione* dei dati (su quale blocco di memoria scriverli);
- ◊ Indicare se l'accesso è in lettura o scrittura;
- ◊ Dimensione di byte che devono essere letti, o scritti.

Dopo di che, all'interno del **DMA controller**, viene fornita la locazione in memoria del comando precedentemente scritto. A questo punto il DMA interpreta il comando fornito dal sistema operativo e gestisce il bus come richiesto. Il controllo del bus viene a tutti gli effetti rubato dalla CPU (*cycle stealing*) ma viene creato un collegamento diretto - e veloce - tra periferica e memoria. Una volta terminato il trasferimento, il DMA, attraverso l'interrupt, segnala il tutto al sistema operativo.

¹Si intende, ovviamente, la memoria RAM.

10.3 Gestione software

Abbiamo discusso dell'hardware, abbiamo parlato della comunicazione tra il sistema operativo e le periferiche, ora saliamo di un livello e capiamo come l'input/output viene gestito a livello software dal *kernel* del sistema operativo. Un approccio abbastanza diffuso è l'utilizzo di moduli dedicati a una o più periferiche: questi moduli sono detti **device-drivers**. Questi costituiscono un livello del *kernel* che mette in comunicazione la parte software con la parte hardware del kernel.

10.3.1 Caratteristiche delle periferiche I/O

Nella gestione delle periferiche via software è importante tenere conto del tipo di periferica e del tipo di operazioni che tale periferica ci consente di fare. Inoltre è necessario tenere conto di alcune proprietà delle periferiche:

1. Se la comunicazione è a caratteri o a byte singoli;
2. Se l'accesso di tipo sequenziale oppure *random*;
3. Se si tratta di una comunicazione di tipo sincrono o asincrono (o eventualmente entrambi);
4. Se si ha a che fare con periferiche condivise tra componenti oppure dedicate;
5. Se la periferica è *read-only*, *write-only* oppure *read-write*;
6. Se la velocità della periferica è ridotta (USB, bluetooth) oppure sostenuta (dal bus PCIe oppure le 100Gbit ethernet).

10.3.2 Tipi di device-drivers

I device-drivers devono quindi essere in grado di gestire la maggior parte dei meccanismo al fine di comunicare con il controllore della periferica. Possiamo dividere i device-drivers in quattro categorie, a seconda delle periferiche a cui si riferiscono: *block devices* (dischi rigidi), *character devices* (mouse e tastiera), *memory-mapped I/O* (scheda grafica) e *networking*.

Block devices In questi *drivers* la comunicazione avviene tramite bytes. I comandi tipici per interfacciarsi con i dispositivi ad essi associati (come i dischi rigidi) includono la lettura, la scrittura e la ricerca (*seek*, ovvero il movimento della testina lungo i cilindri, vedi paragrafo 9.1.1). Nel caso più comune si ha un *file system* dedicato (capitolo 11); altre volte invece si ha un accesso **raw**, ovvero

lascia la possibilità all'applicazione stessa di accedere alla periferica in modo diretto (desiderabile da applicazioni che conoscono completamente l'architettura); infine si può avere un' accesso **diretto** dove il sistema operativo garantisce dei privilegi all'applicazione ma mantiene comunque un controllo per quel *device* per qualche altro processo. È molto comune l'uso di *memory-mapped file*, ovvero associare dei file in aree del disco in modo tale che accedere ad un file equivale ad accedere ad una locazione in memoria. Questi drivers, dato che si interfacciano con il disco, sono quelli che fanno maggiormente uso del **DMA**.

Character devices Questi driver sono invece associati a tastere o mouse che sono più lente e quindi richiedono uno scambio di dati meno consistente. Spesso la comunicazione in lettura e scrittura avviene attraverso dei comandi dedicati come `get` oppure `put`.

Networking Visto l'importanza di tali drivers, questi hanno una classe a parte. L'interfaccia software con cui si comunica con tali periferiche viene chiamata **socket**: interfaccia il sistema operativo a tali *devices*. Oltre alla comunicazione networking tra *devices*, il socket può essere utilizzato per *inter-process communication*, la comunicazione tra processi diversi all'interno della stessa macchina.

Clock e timer Sono molto utili per mantenere la data del sistema oppure per generare degli interrupt periodici (utilizzati nel Round Robin, paragrafo 4.3.2). Attraverso dei protocolli dedicati, come l'*NTP*, sincronizzare diversi computer su una rete per fare in modo che questi siano allineati tutti con la stessa data, correggendo eventuali sfasamenti.

10.3.3 Device sincroni e asincroni

Un'importante caratteristica dei *devices* che abbiamo elencato è se questi siano sincroni o asincroni, in particolare, per quanto riguarda i sincroni, se questi sono **blocking** e **nonblocking**.

Blocking una comunicazione di questo tipo si ha quando, dopo una richiesta di lettura o scrittura su una periferica ed, effettivamente blocchiamo l'esecuzione del codice su quell'istruzione, fino a che la richiesta non è soddisfatta. Questo tipo di comunicazione è ottima se ad ogni richiesta è quasi garantita una risposta; altrimenti si ricade nel *busy-waiting*.

Nonblocking In questo caso il device non fornisce tutti i dati che sono stati richiesti ma fornisce i dati che sono disponibili al momento della richiesta (alcuni magari stanno per essere sovrascritti da un altro processo). Immaginiamo un *device* con un *buffer* e ogni volta che riceve una richiesta, fornisce solo il contenuto del buffer.

Asincrono Nella modalità asincrona, la gestione del *device*, al posto di effettuare continui check, si utilizzano gli interrupt. In questo caso il processore continua a lavorare fino a che non riceve l'interrupt.

10.4 Task del kernel

Nella gestione dell'I/O il kernel deve essere in grado di fornire funzione per accodare e soddisfare richieste di processi diversi associati a periferiche diverse che possono essere condivise o meno. Ovviamente, quando si ha a che fare con situazioni come questa, dove sono presenti code di processi, si avranno sicuramente degli algoritmi di **scheduling** al fine di ottimizzare le tempistiche e soddisfare in modo equo le richieste dei processi verso le periferiche. Ad esempio, nel caso della comunicazione asincrona, il kernel si affida alla **device-status table** (figura 10.4) che è una tabella che contiene lo stato delle periferiche. Per ogni periferica, e quindi per ogni riga della tabella, il kernel collega la coda dei

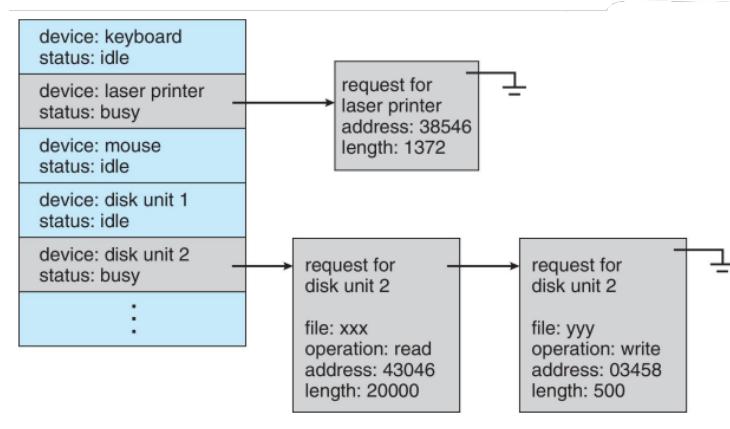


Figura 10.4: Device-status table che, per ogni periferica, ha collegata la coda contenente i processi in attesa.

processi che sono in attesa del *device*.

Ricordiamo che tutte le comunicazioni tra periferiche e il computer avvengono tramite il sistema operativo: di conseguenza il kernel deve anche implementare dei metodi di *buffering* in modo da equilibrare la velocità della periferica con la velocità di CPU. A volte si usano anche dei metodi più sofisticati come il *double buffering* che ritorna utile nel caso di periferiche con diversa velocità: in un buffer processa i dati con velocità ridotta rispetto al secondo in modo da non rallentare la velocità della periferica.

10.4.1 Gestione degli errori

naturalmente il sistema operativo deve essere in grado di gestire anche gli **errori** e di uso errato da parte delle periferiche. Tipicamente quando andiamo a fare delle chiamate per accedere ad una periferica, avremo a che fare con dei messaggi di errore: se vale 0, l'accesso è andato a buon fine, altrimenti il numero identifica l'errore che si è verificato. Il sistema operativo deve essere in grado di gestire questi errori, possibilmente tenendone traccia attraverso dei **file di log**.

Tutta la gestione dell'I/O è fatta in maniera **privilegiata** dal kernel per cui se l'utente vuole accedere ad una periferica, per farlo dobbiamo utilizzare delle *system calls* che fanno la richieste al kernel che a sua VOLA richiede al device. In particolare la *system call* genera un interrupt che viene catturato dal sistema operativo, viene interpretato, prende i dati richiesti dall'utente e li restituisce. C'è sempre un passaggio dal kernel per accedere ad una periferica.

10.4.2 Strutture dati

Ovviamente il kernel fornisce anche delle strutture dati. Le strutture più significative sono tabelle per la gestione di file aperti, connessioni di rete e anche verificare lo stato di alcuni devices. Sono incluse anche strutture dati che tengono traccia di buffers, allocazioni di memoria o se determinati blocchi sono stati sovrascritti o meno (*dirty blocks*).

Si osserva che su sistemi operativi Linux e Unix è possibile accedere alle periferiche in maniera concettualmente simile all'accesso del file. Su Windows, invece, la gestione della comunicazione I/O è effettuata principalmente mediante dei **messaggi** che concettualmente sono più semplici, ma sono fonte di *overhead* oerchè sono più complessi da gestire

Gestione dell'energia

La gestione dell'energia non è un argomento inherente - e per questo non approfondito - alla gestione dell'I/O ma rimane legato. Ci limitiamo ad osservare che nei sistemi moderni ci sono dei meccanismi come le **ACPI** (*Advanced Configuration and Power Interface*) che è un **firmware** che fornisce *routines* utilizzate dal kernel per gestire al meglio non solo l'energia del dispositivo ma anche la gestione di nuove periferiche e la gestione degli errori.

Sappiamo che la gestione dell'energia è un *key factor* per i dispositivi mobile. In questi dispositivi il sistema operativo fornisce dei meccanismi aggiuntivi per la il *power management* in modo anche da utilizzare in maniera ottimizzata le risorse. Tra questi meccanismo si citano i **wake-locks**, di Android, che servono per fare in modo che il sistema operativo non vada in *sleep mode*. Si pensi ad un video che stiamo guardando sul nostro telefono; generalmente il telefono, se inutilizzato, si blocca dopo alcuni minuti: in questo caso però, dato che si sta guardando un video i *wake locks* fanno in modo che il sistema operativo non vada in *sleep-mode* mentre si sta guardando il video. Una volta terminato, i wake locks vengono rilasciati.

Performance

Per gestire tutte le periferiche, il sistema operativo, assieme all'hardware, deve effettuare un numero elevato di operazioni, anche solo per fare una banale operazione di lettura o scrittura. Ecco quindi che ha senso tenere conto di tutti questi meccanismo nel momento in cui si va a progettare un *device-driver*.

Delle buone pratiche per aumentare le performance sono:

- ◊ Ridurre il numero di *context-switches*: ogni volta che un processo si mette in attesa genera un context switch dove parte un altro processo che a sua volta può mettersi in wait etc;
- ◊ Ridurre la copia di dati: i processi devono sempre richiedere i dati tramite il kernel, significa che tali dati devono prima essere copiati nel kernel e poi passati all'utente;
- ◊ Ridurre il numero di interrupt;
- ◊ Cercare di usare il più possibile funzionalità come il DMA;

Generalmente, nella progettazione si parte dall'implementazione di alcuni algoritmi livello applicativo. Una volta testati possono essere passati al kernel in modo da renderli più efficiente fino a poi essere parte integrante del controller dove si va proprio a modificare il firmware stessi. Al caso ottimale si arriva alla soluzione hardware dove cresce l'efficienza ma decresce, ovviamente, il livello di flessibilità.

11

INTERFACCIA DEL FILE SYSTEM

In questo capitolo ci occupiamo del *file system*, ovvero metodo utilizzato dai sistemi operativi per gestire le memorie secondarie, in particolare i *files*. Definiremo quindi il concetto di file e i metodi per accedervi. In secondo luogo si discuteremo di alcune nozioni delle quali sentiamo parlare tutti i giorni come *folder* o *directory* andando a vedere che cosa queste rappresentino nella pratica. Faremo poi alcune considerazioni riguardo la protezione.

11.1 Concetto di file

Partiamo dall'inizio: che cos'è un file? Un file è uno spazio di indirizzi logici contigui dove all'interno si possono scrivere o leggere dei dati. Tali dati possono essere numerici, binari, dei caratteri oppure possono essere dei particolari dati che corrispondono ad un effettivo file eseguibile o programma il quale potrà essere caricato in memoria ed eseguito dal processore.

Spesso, soprattutto nei sistemi Windows, i files sono identificati da un'**estensione** la quale agevola l'interpretazione del contenuto del file ma non è strettamente necessaria per il suo funzionamento. Generalmente è una sequenza di tre o quattro caratteri che viene separata dal nome del file da un punto. Estensioni tipiche in Windows sono `.exe` `.com` se si parla di una file eseguibile, `.obj` `.o` è un file compilato prima di essere linkato oppure `.txt` `.doc` per i documenti.

11.1.1 Attributi e operazioni

Osserviamo inoltre che ogni file possiede delle proprietà, o **attributi**. Tipicamente questi attributi sono visualizzabili attraverso la sezione "Proprietà" del file oppure attraverso il terminale. Il **nome** è l'unica informazione leggibile dall'uomo e non è necessariamente univoco; l'**identificatore** invece è univoco (può essere un numero o un codice) e serve per distinguere ogni file precisamente all'interno del disco; il **tipo** indica il tipo di file (eseguibile, documento, etc) e inoltre segnala se tale file è supportato dal sistema operativo¹; la **locazione** possiede informazioni sulla posizione fisica e logica del file sulla periferica; la **dimensione** indica lo spazio che occupa il file e la **protezione** fornisce informazioni che riguardano l'accesso in lettura/scrittura al file.

Normalmente queste informazioni sono contenute in una struttura chiamata *directory* (vedi paragrafo 11.3) la quale è mantenuta su disco e caratterizza il file system.

I file supportano, ovviamente, diverse operazioni le quali sono generalmente fornite tramite **system calls**. Prima tra tutte è la **creazione** di un file seguita dalla **lettura** e la **scrittura** ad una determinata posizione del file. È inoltre possibile **riposizionare** il puntatore nel file: se lo si pensa come un array corrisponderebbe a puntare ad un altro indice. È consentita anche la **cancellazione** e il **troncamento**. Infine sono naturalmente presenti operazioni per **aprire** (in lettura/scrittura o entrambe) e **chiudere** il file.

11.1.2 Files aperti e file *locking*

Ogni volta che un file viene aperto il sistema operativo tiene traccia di questa apertura in una tabella apposita: l'**open-file table**. È anche presente un **file pointer** che indica dove andare ad operare nel file una volta che questo viene aperto. Tutte le volte che un file viene aperto, viene incrementato il contatore **file-open count**: questo valore non è altro che il numero di processi che hanno aperto quel file in particolare. Una volta che il processo è terminato e il file deve essere chiuso è necessario controllare che tutti i processi lo abbiano effettivamente chiuso al fine di rimuoverlo dalla *open-file table*. La **disk location** è una cache che tiene temporaneamente i dati per fornire un accesso più veloce: ogni volta che si legge o scrive sul file non si fa direttamente su disco ma su questa cache.

Così come discusso nella sincronizzazione tra processi (vedi capitolo 5) dove sono presenti i **lock**, utili per accedere ad una sezione critica di un processo senza avere interferenza da altri processi, anche nel

¹Un eseguibile Windows non può infatti essere eseguito da un dispositivo Linux o UNIX.

caso dei file è possibile utilizzare dei *lock* volti a garantire l'accesso esclusivo di un processo ad un file. La soluzione consigliata è quella di utilizzare un *lock* esclusivo in scrittura e uno condiviso in lettura del file. In questo modo si garantisce che un file sia scritto solo da un processo alla volta oppure che venga letto da più processi contemporaneamente.

I sistemi operativi possono implementare versioni più avanzate di questi lucchetti, alcuni di questi che sono obbligatori, ovvero che bloccano l'accesso a file e nessun altro può accedervi, altri invece sono così detti "suggeriti", dove la scelta di accesso è lasciata al sistema operativo.

11.1.3 Struttura del file

La struttura del file può essere pensata come una sequenza logica contigua di byte in memoria. Ciò nonostante tale sequenza può concettualmente rappresentare delle strutture dati che non sono dei semplici array ma possono essere, per esempio, degli alberi. Questa struttura è dettata dal tipo di file o tipo di applicazione.

11.2 Metodi di accesso

Una caratteristica importante del file è che ha una dimensione fissa, a differenza dei dati in memoria: questo semplifica sicuramente l'accesso al file. Due sono i metodi per l'accesso: uno è sequenziale l'altro invece è diretto (o *random*).

11.2.1 Accesso sequenziale

La tecnica di accesso sequenziale implica lo scorrimento del puntatore lungo tutto il file. Al fine di accedere ad un dato che si trova prima rispetto al puntatore è necessario fare una **reset** del puntatore per poi scorrere nuovamente il file (figura 11.1). Questo è un algoritmo di accesso molto semplice,

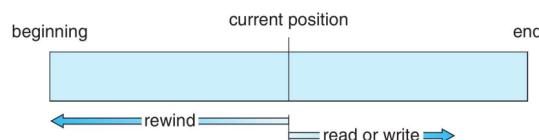


Figura 11.1: L'accesso sequenziale ad un file.

vagamente simile all'algoritmo scan per quanto riguarda i dischi (vedi paragrafo 9.3.3).

Avremo quindi delle istruzioni per leggere o scrivere il byte o la *word* successiva e un'istruzione per effettuare l'operazione di reset

11.2.2 Accesso diretto

Quando parliamo di un accesso diretto, o **random**, parliamo di meccanismo più complesso sia dal punto di vista software che hardware che però garantisce una velocità maggiore.

In questo caso abbiamo invece delle istruzioni più sofisticate in quanto è possibile esplicitare l'**n**-esimo byte da leggere o scrivere. È inoltre possibile simulare un accesso di tipo sequenziale utilizzando istruzioni come `read next` oppure `write next`.

11.3 Struttura della directory

Fino ad ora abbiamo parlato di file come elemento base per lavorare e memorizzare dati su memorie secondarie. Ora invece saliamo di un livello di astrazione e andiamo ad occuparci di che cos'è e come è strutturata una directory.

La directory è una struttura che supporta l'utilizzo dei files e consente l'effettivo accesso al file. Possiamo dire, in altre parole, che è una **collezione di puntatori** o nodi (figura 11.2) che contengono informazioni dei files.

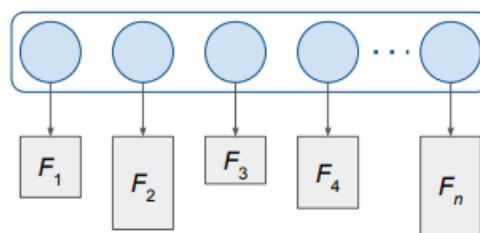


Figura 11.2: La struttura concettuale-minimale di una directory.

Così come è possibile effettuare delle operazioni su file (11.1.1), anche la directory supporta delle operazioni. Tra queste ci sono la **ricerca** di un file, **creare un nodo** che punta ad un file, **eliminare** un file, **listare** il contenuto della directory oppure rinominare e attraversa il file system. La directory garantisce l'**accesso efficiente** ai files (1), gestisce il modo in cui tali files possono essere nominati

(2) in modo tale che due utenti possono fare riferimento allo stesso file utilizzando nomi diversi. Si possono inoltre raggruppare i files (3) in base alle loro proprietà e al loro utilizzo. Questi tre punti saranno utilizzati in seguito per comparare le diverse strutture.

11.3.1 Directory a uno e due livelli

L'approccio più semplice per rappresentare la struttura di una directory è attraverso una singola struttura che contiene un puntatore a tutti i files presenti sul disco (vedi figura 11.3). Per quanto banale,

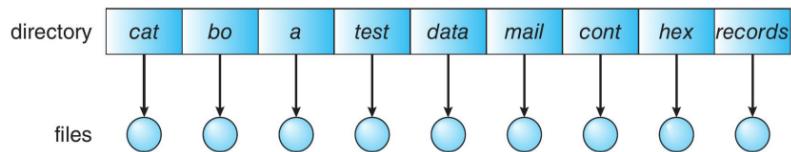


Figura 11.3: Struttura di una direcotory a livello singolo.

per determinati sistemi, come quelli *embedded*² che hanno a disposizione poche risorse è una struttura più che adeguata. Questa struttura si traduce in tutti gli effetti in una tabella che contiene un puntatore con relazione uno a uno da nome alla posizione del file sul disco. Con questa struttura non siamo ovviamente in grado di soddisfare i 3 punti elencati in precedenza.

Per riuscire a soddisfare i 3 punti precedenti è necessario aumentare il livello di complessità: una soluzione sensata è quindi quella di aggiungere un ulteriore livello (figura 11.4). Questa rappresentazione torna nel momento in cui dobbiamo rappresentare un sistema multiutente dove è presente una *master file directory* che contiene il nome di tutti gli utenti, ognuno dei quali ha una propria directory personale. In questo modo è possibile risolvere il problema del **naming**, ovvero avere file con lo stesso nome per utenti diversi, abbiamo una ricerca più efficiente ma comunque il **raggruppamento** non è ancora possibile.

11.3.2 Directory strutturata ad albero

Arriviamo quindi ad una struttura di directory che si avvicina molto a quelle che sono le directory di sistemi operativi moderni con cui normalmente lavoriamo. Queste presentano un numero

²I sistemi embedded sono sistemi integrati. L'esempio più banale è il controllore di un cancello elettronico oppure il processore di una lavatrice.

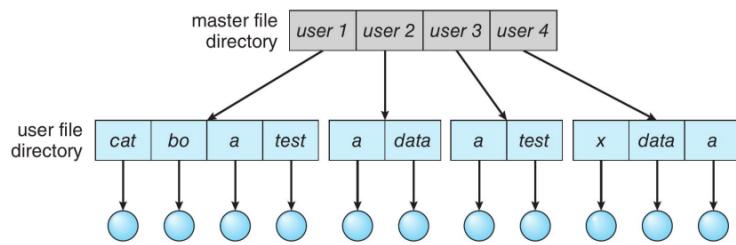


Figura 11.4: Struttura di una directory a due livello

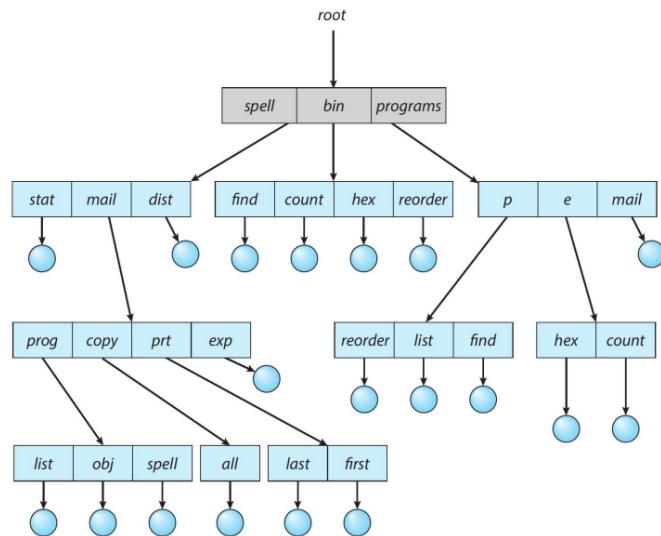


Figura 11.5: Rappresentazione di una directory strutturata ad albero.

potenzialmente infinito di strutture a diversi livelli: ecco che si ha una **struttura ad albero**. All'interno di tale struttura un bit che indica se la struttura è un file o una directory.

Il percorso per l'accesso ad un file può essere specificato in modo relativo o assoluto. Nel primo caso si specificano le operazioni da fare dalla posizione corrente mentre nel secondo caso è riferito solo alla *root*, ovvero il nodo principale da cui è strutturato il file system.

11.3.3 Problema dei cicli

Nel lavorare con questo tipo di strutture si può incorrere nel **problema** della possibile creazione di **cicli**. Per esempio se abbiano a che fare con una struttura ad albero e creiamo un **link** a directory o files possiamo avere a che fare con cicli. Osserviamo la figura 11.6 possiamo notare che due sotto-directory diverse puntano entrambe ad un file che si trova nella directory `list`. Questo ciclo deve essere gestito dal sistema operativo. Se infatti eliminiamo l'intera directory `avi` la directory `jim/book` punterà ad

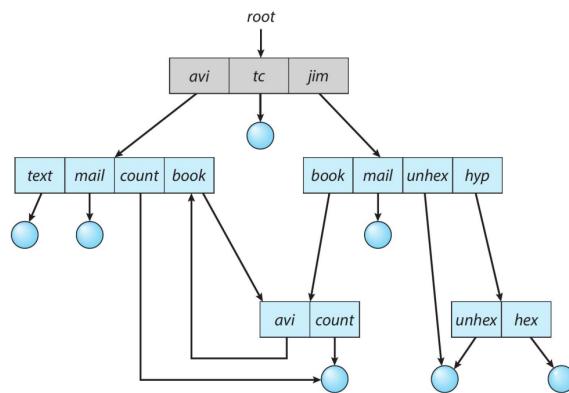


Figura 11.6: Rappresentazione generica di una directory.

una directory inesistente in quanto eliminata: stiamo parlando dei **dangling pointers**. Come possiamo risolvere questa soluzione? Ci sono diversi modi. Primo tra tutti è l'eliminazione di tutti i link oppure, il file viene eliminato solo nel momento in cui tutti i riferimenti a tale file sono rimossi³.

Ci sono quindi diversi modi per gestire i cicli all'intero di queste strutture.

1. Evitare i cicli: ogni volta che si crea un riferimento che genera un ciclo, proibirne la creazione;
2. Ad ogni link creato, verificare se viene creato un ciclo gestibile oppure se può creare dei problemi;
3. Permettere la creazione di cicli, ma evitare loop infiniti di ricerca di file attraverso qualche contatore che impedisca la ricerca ciclica di un file;
4. Attraverso la **garbage collection**: viene eseguito un processo a parte periodicamente che elimina i *dangling pointers*.

³Molto simile al concetto dello `shared_pointer` di C++.

riferimento
RAID

11.3.4 Disco

Il disco rigido è organizzato e strutturato per fornire, ovviamente, la possibilità di immagazzinare files. Normalmente è diviso in **partizioni** che possono essere di diverso tipo a seconda dell'uso. Ci possono essere delle partizioni di tipo **RAID** (vedi paragrafo), di tipo **raw**, cioè che possono essere utilizzate direttamente da un database che gestisce a modo suo l'informazione, di tipo **swap** che permettono un accesso veloce (come cache) o anche **formattate** con un determinato file system e di conseguenza inizializzati con una particolare struttura che varia a seconda del sistema operativo o a seconda dell'utilizzo che ne vogliamo fare⁴.

Normalmente la partizione di disco che contiene il file system viene chiamata anche *volume*. All'interno di esso troviamo una *device directory*, ovvero una directory principale che contiene le informazioni generali sul contenuto del volume.

11.4 Protezione

Come sui file c'è l'attributo per la loro protezione, anche nel caso delle directory emerge la questione di come proteggere sia il contenuto del file che dell'intera directory. Uno dei primi obiettivi che è necessario raggiungere è quello di riuscire a proteggere i files o le directory create dell'utente. Si vorrebbe che l'utente proprietario di un determinato file o directory abbia un accesso privilegiato rispetto ad altri utenti. Per gestire la proprietà del file è quindi necessario gestire azioni in lettura, scrittura, esecuzione, rimozione, aggiunta o *list*.

11.4.1 Access list in Unix/Linux

Nei sistemi operativi Unix/Linux tipicamente si identificano tre tipi di accesso: *read* (R), *write* (W) ed *execute* (X); e tre classi di utente: *owner*, gruppo e una classe pubblica. Per gestire gli accessi si usa quindi la terza RWX binaria. Per esempio, se all'*owner* è assegnato 111 = 7 (attraverso il comando chmod) ha accesso sia in lettura, che in scrittura che in esecuzione del file questo perché i 3 bit che rappresentano RWX sono tutti a 1. Se invece all'accesso pubblico assegniamo 100 = 4, questi avranno accesso solo in lettura.

⁴Per esempio, nei sistemi Windows abbiamo partizioni di tipo NTFS.

12 IMPLEMENTAZIONE DEL FILE SYSTEM

In questo capitolo ci occuperemo di discutere in dettaglio il modo in cui viene implementato il file system. Capiremo che cos'è la struttura di un file system e che tipo di operazioni supporta. Discuteremo inoltre come sono implementate le directories e i diversi metodi di allocazione del file system: questi aspetti sono già stati citati nel capitolo precedente ma in questo capitolo scaviamo più nel profondo e cerchiamo di capire come questi siano effettivamente creati. Parleremo infine della gestione dello spazio libero e di performance.

12.1 Struttura e operazioni del file system

La periferica fisica, come il disco, ci permette di leggere e scrivere i dati. Tipicamente è composto di **blocchi** che vanno da 512 a 49 bytes. Il file system risiede sulla memoria secondaria e che fornisce l'interfaccia utente per memorizzare tali dati mappando il loro indirizzo logico in un indirizzo fisico sul disco. Inoltre il file system deve anche essere in grado di fornire lo spazio sulla feriferica in modo efficiente.

12.1.1 Livelli

Il file system non è un singolo blocco ma lo si può considerare come un'struttura organizzata a diversi livelli. Osserviamo la figura 12.1 e discutiamo i diversi livelli. Il **controllo I/O** viene effettuato attraverso i *device-drivers* (già discussi nel paragrafo 10.3) e attraverso gli *interrupt handlers* (vedi paragrafo 10.2.2). Questo si occupa di tradurre comandi del tipo `read drive3, cylinder 14, traccia 15, settore 92` nella locazione di memoria 6535. Salendo di un livello incontriamo il **basic file system** che gestisce comandi del tipo `read block 12` di un determinato indirizzo logico. Questo gestisce anche eventuali buffer che possono essere eventualmente utilizzati per la trasmissione della periferica alla memoria e di cache, utili per mantenere una copia dei blocchi utilizzati più frequentemente. Dopo di che troviamo il **modulo di organizzazione dei file** che è in grado di interpretare la struttura dei file e il loro indirizzo logico. Si occupa di tradurre l'indirizzo logico in indirizzo fisico e gestisce il *free space* (vedi 12.3). Infine, poco più sotto del livello di applicazione, troviamo il **logical file system** che si occupa di gestire i *metadata* e strutture di alto livello. Per esempio è responsabile della traduzione dei nomi dei file in un ID, identificativo numerico che viene utilizzato dai livelli sottostanti. Mantiene inoltre il *file control block* (12.1.3), chiamato anche **inode** in UNIX. Oltre a ciò gestisce anche la struttura delle directory (discusse nel paragrafo 11.3) e la protezione (paragrafo 11.4).

Questa stratificazione del file system è stata ideata per ridurre la complessità di tutte le operazioni da effettuare però ha lo svantaggio perché introduce un **overhead** che in alcuni casi può decrementare l'effettiva performance. Ecco perché in alcuni applicativi non è presente uno specifico file system per memorizzare dati per evitare di andare attraverso questi layers. Queste operazioni sono infatti lasciate all'effettivo *database* che gestisce i dati.

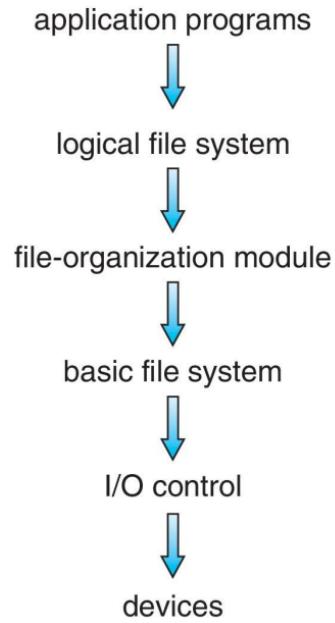


Figura 12.1: I livelli del file system.

Tipi

Possiamo trovare una grande vastità di file systems implementati in maniera diversa. Alcuni di questi sono dedicati a periferiche removibili come nel caso della ISO 9660, utilizzata per i CD. Altri invece variano a seconda del sistema operativo dato che sistemi operativi diversi spesso sono pensati per operazioni diverse. Ecco quindi che in Unix troviamo **UFS**, in Windows, storicamente FAT (12.2.2), e adesso NTFS mentre in Linux troviamo **ext3** ed **ext4**. Citiamo anche il **ZFS**, trovato in Solaris e GoogleFS che sono più recenti.

12.1.2 Strutture per le operazioni

Naturalmente il file system deve supportare operazioni per lavorare con i dati. Per fare ciò necessita di contenere delle strutture dati per agevolare e implementare tali operazioni. Possiamo dividere le strutture in due tipo: quelle che risiedono sul disco e quelle che vivono in memoria.

Sul disco troviamo tipicamente dei blocchi dedicati al booting ovvero i **boot control block** che servono appunto fare partire il sistema operativo: fare eseguire il minimo necessario per poi far partire il kernel e tutto il resto. È inoltre presente il **volume control block** che contiene informazioni sul volume: il numero totale di blocchi sul device oppure il numero di blocchi liberi. Infine è presente la **directory** - o almeno parte di essa - che, come sappiamo (11.3), è utilizzata per organizzare i files: contiene quindi anche gli FCB.

Nella memoria invece sono presenti le strutture per implementare le system calls. Troviamo quinti la **mount table**, che contiene le parti che sono collegate al sistema operativo. È presente anche la **in-memory directory structure**, ovvero la memoria che si occupa di detenere le informazioni riguardanti le directory utilizzate più recentemente. Nella memoria possiamo anche trovare **system-wide open-file table** ovvero la tabella di sistema dei file aperti che contiene una copia degli FCB (12.1.3). Come c'è la tabella di sistema è presente anche una tabella con i FCB dei files aperti per ogni processo. Infine sono anche presenti dei buffers per agevolare le operazioni di lettura e scrittura.

12.1.3 File Control Block (FCB)

Il *File Control Block*, già menzionato molte volte all'interno di questo capitolo, è una struttura mantenuta dal sistema operativo che, per ogni file, contiene diverse informazioni riguardanti tale file. Dalla figura 12.2 possiamo notare che in genere sono memorizzati i permessi del file (come le liste di

accesso, 11.4.1), informazioni temporali (come la creazione o l'ultimo accesso), proprietà, dimensione

file permissions
file dates (create, access, write)
file owner, group, ACL
file size
file data blocks or pointers to file data blocks

Figura 12.2: La struttura del *File Control Block*

e altre informazioni riguardanti il reperimento di informazioni del file come puntatori.

completa
slide 10 e 11
(minuto 18 - 24)

12.2 Metodi di allocazione

Arriviamo al punto più critico: come avviene effettivamente allocato lo spazio sul disco per contenere i file? Ci sono tre tecniche principali - allocazione contigua, concatenata e indicizzata - e le approfondiremo tutte in questo paragrafo.

12.2.1 Allocazione contigua

Nell'allocazione contigua il disco viene diviso in un numero di blocchi che possono essere da 512 byte o 4096 byte. I file sono quindi allocati in **locazione contigua**, una vicina all'altra. Ricordiamo che sia nel caso di dischi elettromeccanici ma anche negli SSD blocchi vicino corrispondono ad una velocità maggiore in accesso. Osservando l'illustrazione 12.3 possiamo notare che non è nemmeno difficile da implementare: è necessario solamente memorizzare la locazione del blocco iniziale e la dimensione/lunghezza del file: per esempio il file `mall` parte dalla locazione 19 e ne occupa 6; possiamo infatti notare che i blocchi, dal 19 al 24, sono ingrigiti in quanto occupati.

Questa soluzione però porta con sé dei problemi non indifferenti. Innanzitutto è necessario conoscere la lunghezza del file, ma non è sempre possibile saperlo, soprattutto se si tratta di un file che magari viene modificato (per esempio un documento). Ancora più grave è la **frammentazione esterna**: questo significa che all'interno del disco è presente dello spazio ma questo spazio non è contiguo. Per esempio,

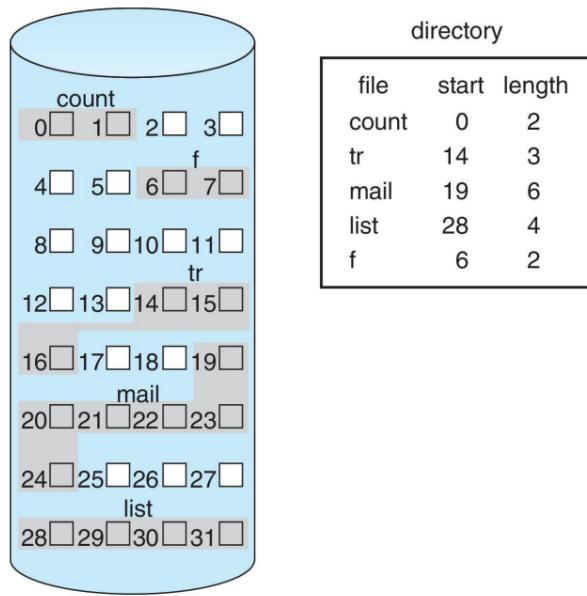


Figura 12.3: Rappresentazione dell'allocazione contigua su disco.

sempre facendo riferimento alla figura 12.3, se dovessimo inserire un file di lunghezza 7 questo non sarebbe possibile anche se sono presenti più di 7 blocchi liberi nel disco. Si possono adottare delle tecniche per fare una **compattazione** dove, per esempio, si trasferiscono tutti i file in un altro *storage* per poi ricopiarli uno dopo l'altro in modo tale da renderli tutti attaccati. Questa però rimane una soluzione non ottimale in quanto nel caso di grandi *data centers* non è possibile bloccare il sistema per fare questi trasferimenti.

Un possibile miglioramento di questa tecnica si basa sull'uso di **extent**, non a caso questa versione migliorata è chiamata *extent-based system*. Un extent è un insieme di più blocchi contigui sul disco. Possiamo dire che è un approccio semi-contiguo nel senso che è contiguo sicuramente per tutta la lunghezza di un extent. Questa soluzione è implementata nel *Veritas file system*.

12.2.2 Allocazione concatenata

In questa tecnica, al posto di utilizzare una disposizione contigua si sceglie di disporre il file in una lista di blocchi non contigui i quali possono essere sparsi all'interno del disco. Come possiamo osservare dalla figura 12.4 questo comporta la necessità di avere un puntatore in un blocco n che punta al blocco

$n + 1$. Di conseguenza, all'interno di ogni blocco, è necessario utilizzare dello spazio per memorizzare

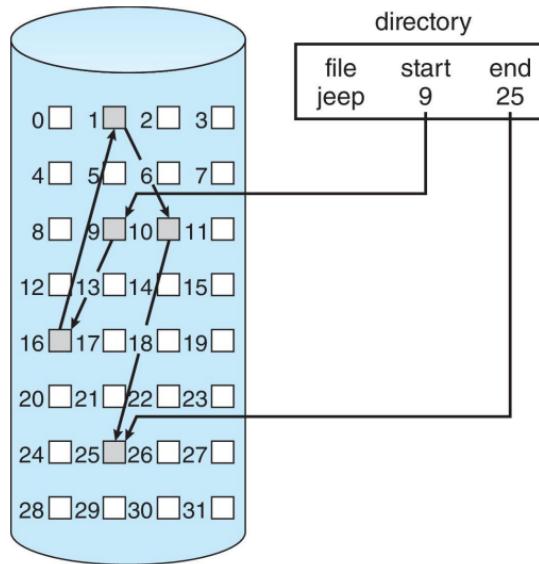


Figura 12.4: Rappresentazione dell'allocazione concatenata.

il puntatore al blocco successivo. Per esempio, se i blocchi occupano 512 bytes e il sistema dispone di indirizzi a 4 bytes (32 bit) allora in ogni blocco si possono memorizzare $512 - 4 = 508$ bytes di file. Non è un valore eccessivamente grande ($\frac{1}{100}$) ma è bene tenerne conto.

Il vantaggio è che la frammentazione esterna presente nell'allocazione contigua non è più un problema. D'altro canto però non è un'ottima tecnica se si vuole avere un accesso **random** in quanto al fine di arrivare al blocco n si devono effettuare $n - 1$ accessi in lettura dato che è necessario scorrere tutta la lista. Per migliorare l'accesso diretto si può *clusterizzare* i blocchi: al posto di avere dei blocchi isolati si possono avere dei gruppi di blocchi contigui cosicché da scorrerli una volta raggiunti. Questa soluzione però aumenta la **frammentazione interna** nel caso in cui lo spazio allocato da quel *cluster* non venga utilizzato completamente. Infine, l'ultimo problema di questa soluzione si ha nel caso in cui un **blocco** sia **corrotto** (a livello di bug oppure anche fisicamente). Questo perché non sarebbe più possibile raggiungere la parte della lista seguente al blocco.

Un esempio di questo tipo di allocazione è la storica **FAT** (*File-Allocation Table*, figura 12.5) che era un tempo utilizzata nei sistemi MS-DOS e si trova tutt'ora in qualche applicazione. Questa è una tabella dedicata che è detenuta all'inizio del volume del disco ed ogni elemento di tale tabella contiene l'indice

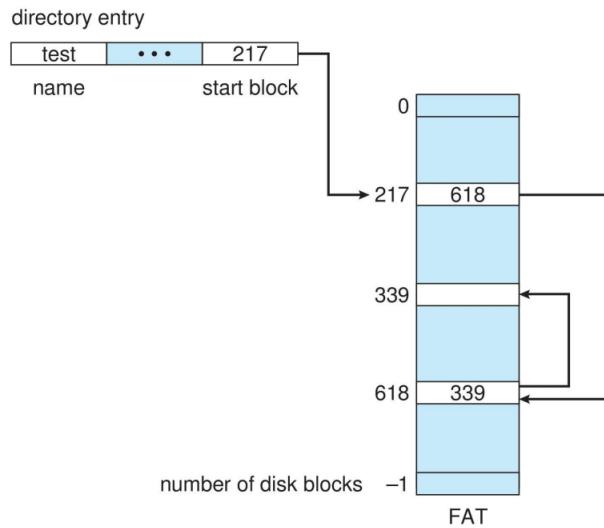


Figura 12.5: La tabella FAT.

del blocco successivo. Quindi per identificare quali sono i blocchi di un determinato file è sufficiente mantenere l'informazione nella directory del blocco iniziale e poi nella *file table* c'è la posizione dell'elemento successivo e così via fino a che non si raggiunga la fine del file. Con questa strategia, per effettuare l'accesso su un file non è necessario scorrere gli effettivi blocchi sul file bensì è sufficiente scorrere la FAT che rappresenta la struttura di quei blocchi.

12.2.3 Allocazione indicizzata

L'ultimo metodo che abbiamo elencato è il metodo di allocazione indicizzata. Anche questa tecnica si basa sui puntatori ma cerca di utilizzarli in maniera efficiente. Come è possibile notare nell'illustrazione 12.6, al posto di salvare su ogni blocco un puntatore (o indirizzo) al blocco successivo, è più efficiente memorizzare tutti gli indirizzi tutti i blocchi del file su un unico blocco, chiamato **index block**. Così facendo, con n blocchi, $n - 1$ blocchi sono utilizzati per la pura memorizzazione mentre l'ultimo è utilizzato per collegarli tutti. Dunque non si ha più una struttura sequenziale come nell'allocazione concatenata ma si ha una struttura simile a quella di una ragnatela. Per files molto grandi, inoltre, è possibile utilizzare uno blocco fra gli $n - 1$ restanti e utilizzarlo anch'esso per indicizzare altri $m - 1$ blocchi.

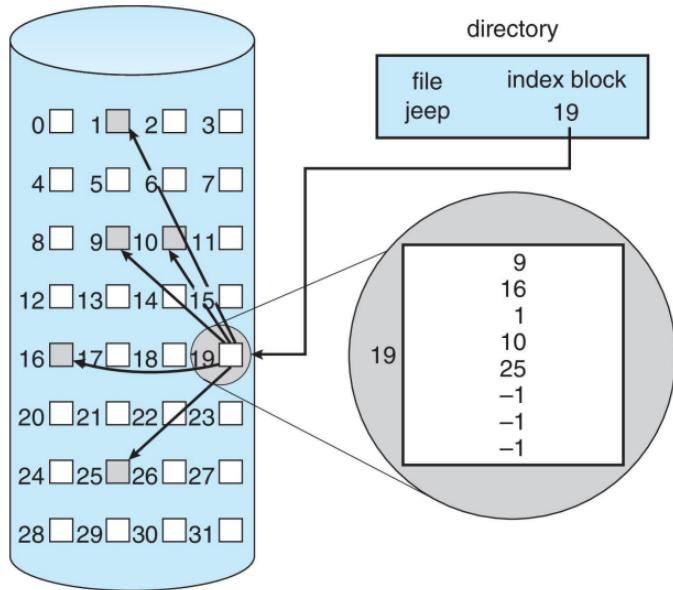


Figura 12.6: Rappresentazione dell'allocazione indicizzata.

In questo caso il problema è che se il file è molto piccolo, gran parte dell'index block è sprecato: si incombe quindi nella frammentazione interna. Ridurre la dimensione dei blocchi per evitare la frammentazione interna non è una buona idea perché elimina il problema ma nel caso di files di grandi dimensioni si necessitano molti più index blocks. Ecco quindi che si è scelto di migliorare questa tecnica attraverso diverse versioni. Nello **schema concatenato** si utilizzano dei blocchi di piccole dimensioni andando quindi a creare più blocchi indice collegati tra loro (strutturalmente diventa una sequenza di piccole ragnatele). Nello schema con **indice multi-livello** il blocco indice di primo livello contiene gli indirizzi a blocchi indice di secondo livello aumentando notevolmente lo spazio di indirizzamento.

Infine c'è lo **schema combinato** che mescola le versioni precedenti ed è quello che tuttora viene utilizzato in Unix e Linux dove abbiamo a che fare con l'**inode** che, come visto in 12.1.1, non è altro che il *File Control Block* di Linux. Osservando la figura 12.7, osserviamo che all'interno dell'inode sono presenti 15 puntatori:

- ◊ I primi 12 puntatori fanno riferimento diretto ai *datablock* che si trovano sul disco, si chiamano infatti **direct blocks**. Di conseguenza se dobbiamo accedere ad un file con dimensione minore di 12 blocchi è possibili utilizzare questi puntatori;

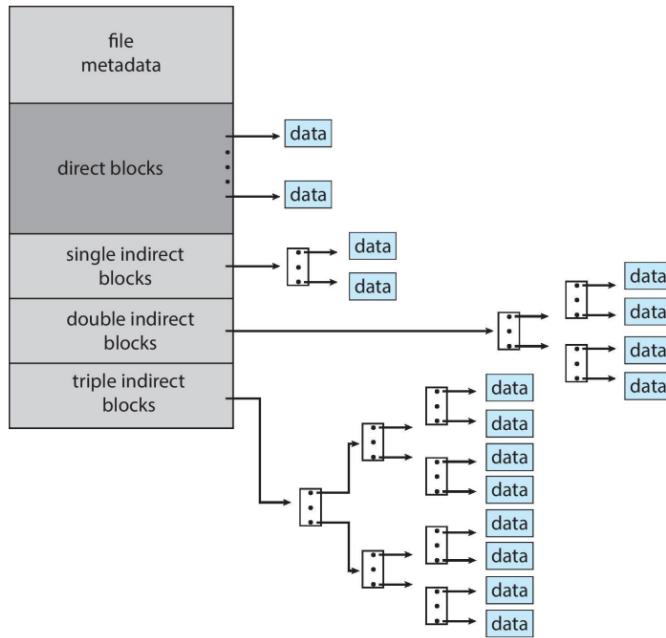


Figura 12.7: Rappresentazione dell'allocazione indicizzata - schema combinato.

- ◊ Il 13esimo puntatore è usato come **single indirect block**; ciò significa che contiene un puntatore ad un blocco puntatori. Di conseguenza per accedere ai dati è necessario accedere ai blocchi puntatore (che si trovano sul disco) e da lì essere indirizzato al dato effettivo (si fanno quindi 3 "salti", uno per il primo blocco, uno per il secondo e uno per leggere il dato);
- ◊ Il 14esimo e il 15esimo rappresentano il **double e triple indirect block** che hanno due e tre livelli di blocchi puntatore (che si trovano tutti sul disco).

Supponiamo che un blocco sia di 512 byte e gli indirizzi siano a 32 bit (4 byte), con il *single indirect block* possiamo indirizzare $\frac{512}{4} \cdot 512 = 128 \cdot 512$ bytes (64KB), con il *double* $128^2 \cdot 512$ bytes (8MB) e con il *triple* $128^3 \cdot 512$ bytes (1GB).

Il vantaggio di questo schema è che la complessità arriva solo nel caso in cui si lavora con file di grandi dimensioni. Se invece si lavora sempre con file sufficientemente piccoli (6KB) si avranno a disposizione i 12 puntatori iniziali che sono molto più rapidi rispetto agli ultimi 3. Si nota infatti che l'inode risiede nella memoria e di conseguenza per effettuare l'accesso ai blocchi puntati dai 12 puntatori l'unica accesso al disco che si effettua è proprio quello di prelevamento del file.

12.2.4 Performance

Come abbiamo notato è che il metodo migliore in assoluto non esiste ma è strettamente legato al tipo di accesso che effettuiamo. Se infatti l'accesso è sequenziale quasi tutte le tecniche vanno più che bene. Se invece l'accesso è di tipo random la scelta della tecnica gioca una ruolo significativo. Inoltre è importante aver chiaro che non è assolutamente detto che le tecniche più complesse siano le migliori: se prendiamo in considerazione un sistema di backup che memorizza dei file enormi (GB oppure TB) in modo contiguo, non è assolutamente necessario implementare una struttura indicizzata come l'inode, è sufficiente utilizzare l'allocazione contigua che è molto più semplice da implementare.

È bene inoltre puntualizzare che nel caso delle memorie non volatili (paragrafo 9.1.2) come le SSD, le tecniche che abbiamo appena elencato non sono compatibili dato che le SSD sono fisicamente organizzate in maniera diversa rispetto agli HDD.

12.3 Gestione dello spazio libero

Oltre alla memorizzazione dei dati, nella memoria secondaria è importante tenere traccia dei blocchi che sono liberi e utilizzabili in futuro. Tipicamente il sistema operativo mantiene la **free space list** che è una lista che tiene traccia dei blocchi (o cluster di blocchi) che sono liberi. Ci sono diversi metodi per implementare questa lista - come la free list o il grouping - e in questo paragrafo li discuteremo brevemente.

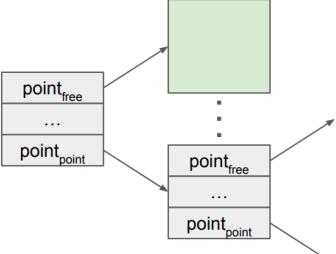
Sprechiamo prima due parole sull'soluzione più banale: il **bit vector** o, in alternativa, la *bitmap*. Questa è una sequenza di n bit che rappresentano gli n blocchi; se il bit i -esimo è impostato ad 1 significa che il blocco i -esimo è libero. Per esempio sei sono liberi i blocchi 1, 3, 4, 5, 9 allora il bit vector sarà 101110001.

12.3.1 Free list

Ritornano anche in questo caso utili le liste. La free list è una lista che tiene traccia di tutti i blocchi che sono liberi all'interno del disco. Ciascun blocco in realtà non è completamente vuoto ma contiene, naturalmente, il puntatore al blocco libero successivo. Quindi l'unica cosa di cui dobbiamo tener traccia è il puntatore al primo blocco (**head**), a differenza invece della bitmap che cresce al crescere dello spazio di memorizzazione. Nel momento in cui si ha bisogno di un blocco si va ad occupare

il primo e si aggiorna l'head al secondo blocco (che diventa quindi il primo). Ne consegue che, se abbiamo bisogno del primo blocco o dei primi n blocchi, non è assolutamente necessario scorrerla tutta, quindi il problema della complessità non è così significativo. Inoltre non è facile ottenere dello spazio contiguo in quanto questo tipo di implementazione si occupa, ovviamente, di blocchi sparpagliati.

12.3.2 Counting e grouping



Altre soluzioni lecite si ottengono attraverso un raggruppamento dei blocchi, come nel **grouping** (figura a lato): molto simile alla soluzione di indicizzazione multi-livello che abbiamo discusso precedentemente. Potremmo infatti avere un blocco dedicato che contiene tutti i blocchi liberi i quali, a loro volta, possono puntare a blocchi liberi. Il metodo **counting** invece, dato che spesso lo spazio è liberato in modo contiguo quindi è sufficiente tenere traccia del numero di blocchi successivi che sono liberi. Non è quindi sufficiente tenere traccia di un indirizzo e il numero di blocchi che, oltre a quell'indirizzo, sono liberi uno dopo l'altro. Nel caso in cui il contatore del numero di blocchi contigui è > 1 allora possiamo dire che l'implementazione di un contatore risulta più conveniente rispetto alla *free list*.

12.3.3 Altri metodi

Concludiamo menzionando altri metodi usati in altri file system che si occupano di risolvere altri problemi. Per esempio **ZFS** - presente in Solaris - divide lo spazio del disco in *metaslab* i quali hanno associato una **spacemap** che è un log per memorizzare le attività effettuate in quella particolare zona del disco. Osserviamo inoltre che questo file system è ottimo per la gestione di files di grandi dimensioni grazie anche ad un algoritmo basato sul counting (12.3.2)

Menzioniamo anche un'altra tecnica, chiamata **TRIM**, che è un nuovo meccanismo utilizzato per gestire le NVM. Ricordiamo infatti che sulle SSD non è possibile effettuare un'operazione di sovrascrittura (vedi 9.1.2) bensì è necessario prima cancellare il blocco per poi scrivere la nuova informazione e quindi è necessario utilizzare delle strategie diverse. In questo caso il metodo TRIM

riesce a ridurre la necessità di fare frequentemente *garbage collection* e minimizza la possibilità di eliminare grandi blocchi di dati