ANNO ACCADEMICO 2024/2025

Sistemi Operativi

Teoria

Dionesalvi's Notes

UNIVERSITÀ
DI TORINO

DIPARTIMENTO DI INFORMATICA

Capitolo 1 Introduzione Pagina 2

1.1 Prima Lezione 2
Architetture Single/Multi-Core — 2 • Tipi di Eventi — 2 • Gestione degli Eventi — 2
1.2 Struttura della Memoria 3
1.3 Gerarchia delle Memorie 3
1.4 Struttura di I/O 4
1.5 Multitasking e Time-Sharing 4
Time-Sharing — 5
1.6 Compiti del sistema operativo 5
Duplice modalità di funzionamento — 6 • Timer — 7 • Protezione della memoria — 7

Capitolo 2 Strutture dei Sistemi Operativi Pagina 8

2.1 Interfaccia del Sistema Operativo 8
Interprete dei comandi — 8
2.2 Interfaccia grafica 9
2.3 Programmi/servizi di Sistema 9
2.4 Chiamate di sistema (Syscall) 9
Chiamate di sistema: le "API" — 9
2.5 Gestione dei processi 10
2.6 Gestione dei file e del filesystem 10
2.7 Macchine Virutali 10

Capitolo 3 Gestione dei processi Pagina 11

3.1 Processi 11
Concetto di processo — 11 • Stato del processo — 12 • Processo Control Block (PCB) — 13
3.2 Scheduling dei processi 13
Il cambio di contesto (context switch) — 14 • Code di scheduling — 15 • CPU Scheduler — 16
3.3 Operazione sui processi 16
Creazione di un processo — 16 • Creazione di un processo in Unix — 17 • Passi dell'SO all'invocazione delle fork — 17 • Altro esempio — 18 • Osservazioni — 18 • Terminazione di un processo — 19
3.4 Comunicazione tra processi 19
3.5 Esempio: il problema Produttore-Consumatore 19
Inter-Processo Comunication (IPC) — 20

Capitolo 4 Threads Pagina 22

4.1 Processo Multi-Thread 23

4.2 CPU/Core multi-threaded 25
Limitazioni delle Architetture Superscalari — 25 • Simultaneous Multi-Threading (SMT) — 25 • Definizioni: Dual-Threaded vs Dual-Core — 26
4.3 Ma l'SMT è sempre vantaggioso? 26

Capitolo 5 Scheduling della CPU Pagina 28
5.1 Scheduling 28
Fasi di eleborazion e di I/O — 28 • Lo Scheduler della CPU — 28 • Il Dispatcher — 29
5.2 Criteri di Scheduling 30
5.3 Algoritmi di Scheudling 30
First Come First Served (FCFS) — 30 • Shortest Job First (SJF) — 31 • Osservazioni — 32 • Scheduling a Priorità — 32 • Scheduling Round Robin (RR) — 33 • Esempio — 34 • Scheduling a Code Multiple — 34 • Scheduling a Code Multilivello con retroazione (MFQS) — 35
5.4 Scheduling per sistemi multi-core 36
Esempio — 36
5.5 Esempi di sistemi operativi 36
Lo scheduling in Windows — 37 • Lo Scheduling in Linux — 38

Capitolo 6 Sincronizzazione dei Processi Pagina 39
6.1 Introduzione 39
Esempio: Produttore-Consumatore con n eleemnti — 39
6.2 Sezioni critiche 40
Problema della Sezione Critica — 40 • Sincronizzazione via Hardware — 42
6.3 Semafori 44
Uso dei semafori — 45 • Implementazione dei semafori — 46 • Riassunto — 47
6.4 Definizione di DeadLock 47
6.5 Definizione di Starvation 48
6.6 Deadlock & Starvation: (stallo e attesa indefinita) 48

Capitolo 7 Esempi di sincronizzazione Pagina 49
7.1 Produttori-Consumatori con memoria limitata 49
Codice del Produttore — 49 • Codice del Consumatore — 50 • Spiegazione — 50
7.2 Problema dei Lettori-Scrittori 50
Codice del Processo Scrittore — 50 • Codice del Processo Lettore — 51 • Spiegazione — 51
7.3 Problema di cinque filosofi 51
Dati Condivisi — 51 • Codice del Filosofo i (Soluzione Errata) — 51 • Problema di Deadlock — 51 • Soluzioni Migliori — 52

Capitolo 8 Stallo dei processi (deadlock) Pagina 53
8.1 Definizione 53
8.2 Situazioni simili anche nella realtà 53
8.3 Problema dei nastri 53
Dati Condivisi — 53 • Codice dei Processi P1 e P2 — 53 • Problema di Deadlock — 54
8.4 Un ponte ad una sola corsia 54
8.5 Modello del sistema 54
8.6 Caratterizzazione dei Deadklock 54

8.7 Metodi per prevenire dei Deadlock 55

Capitolo 9 Memoria centrale Pagina 56

9.1 Introduzione 56
9.2 Binding (associazione degli indirizzi) 57
Quando? — 58
9.3 Spazio degli indirizzi (Logici e Fisici) 60
9.4 Le librerie 61
Tipi di Librerie — 61 • Estensioni delle Librerie Dinamiche — 62
9.5 Tecniche di gestione della memoria primaria 62
9.6 Allocazione contigua della Memoria Primaria 62
Allocazione a partizioni multiple fisse — 63
9.7 Allocazione a partizioni multipli variabili 64
La frammentazione — 65
9.8 Paginazione della memoria 65
Metodo di base — 65 • Rilocazione Dinamica — 71 • Dimensione delle Pagine e Gestione della Tabella delle Pagine — 71 • Supporto Hardware — 72 • Pagine condivise — 73
9.9 Paginazione a più livelli 74
9.10 Paginazione a due livelli 75
Page Table Invertita (IPT) — 76
9.11 Il supporto alla paginazione nei vecchi processori Intel 77
9.12 Conclusioni 77

Capitolo 10 Memoria virtuale Pagina 78

10.1 Introduzione 78
10.2 Paginazione su richiesta (Demand Paging) 79
10.3 Demand Paging 80
L'area di swap — 82
10.4 Sostituzione delle pagine 83
Algoritmi di sostituzione delle pagine — 84 • Sostituzione delle pagine secondo l'ordine d'arrivo (FIFO) — 85 • Algoritmo LRU (Least Recently Used) — 87 • Algoritmo Seconda Chance con Dirty Bit — 89
10.5 Allocazione dei Frame 89
Thrashing (attività di paginazione degenere) — 91 • Cause del Thrashing — 91 • Come combattere il Thrashing — 92
10.6 Dimensioni delle pagine 93
Struttura dei programmi — 93
10.7 Gestione nei Sistemi Operativi 94
Windows — 94 • Solaris — 94

Capitolo 11 Memoria di massa Pagina 96

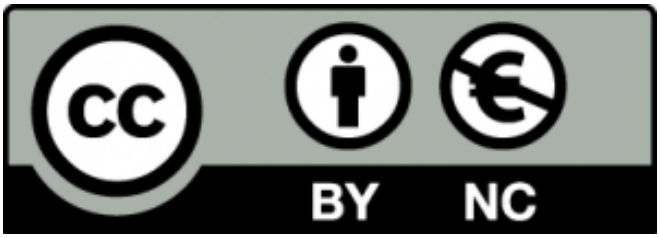
11.1 Disco Rigido 96
Struttura — 96 • Mappatura degli indirizzi — 97 • Scheduling dei dischi rigidi — 97
11.2 Formattazione del disco 98
Il Boot Block — 99
11.3 Gestione dell'area di SWAP 99
Gestione dell'area di Swap — 99

11.4 Sistemi RAID 99
Introduzione ai sistemi RAID — 100 • Caratteristiche principali di un sistema RAID — 100 • Idee principali alla base di RAID — 100 • Livelli di RAID — 100 • RAID di Livello 0 — 100 • RAID di Livello 1: Mirroring — 101 • RAID di Livello 01: Striping + Mirroring — 102 • RAID Livello 10 (Mirroring + Striping) — 103 • RAID Livello 4 (Striping con Parità) — 104 • RAID Livello 5 (Striping con Parità Distribuita) — 104 • RAID Livello 6 (Striping con Doppia Parità Distribuita) — 104 • Ricostruzione di Dati Persi con la Parità — 105
11.5 Memorie a Stato Solido (SSD) 106
Confronto tra Tipi di Memorie — 106 • Utilizzo degli SSD — 106

Capitolo 99 Esercizi Pagina 108
99.1 Capitolo 5 108
1 — 108
99.2 Esercizi pre-esame 108

Premessa

Licenza



1 Introduzione

1.1 Prima Lezione

Un Sistema Operativo (SO) agisce come intermediario tra l'utente e l'hardware, fornendo gli strumenti per un uso corretto delle risorse della macchina (CPU, memoria, periferiche). Ha due obiettivi principali:

- Dal punto di vista dell'utente: rendere il sistema facile da usare.
- Dal punto di vista della macchina: ottimizzare l'uso delle risorse in modo sicuro ed efficiente.

1.1.1 Architetture Single/Multi-Core

Negli anni 2000 si è passati da processori single-core a multi-core, con CPU dotate di più core in grado di eseguire istruzioni di programmi diversi simultaneamente.

1.1.2 Tipi di Eventi

- **Interrupt**: Eventi di natura hardware, rappresentati da segnali elettrici inviati da componenti del sistema.
- **Eccezioni**: Eventi di natura software, causati dal programma in esecuzione. Le eccezioni si dividono in:
  - *Trap*: Causate da malfunzionamenti del programma (es. accesso a memoria non autorizzato, divisione per 0).
  - *System Call*: Richiesta di servizi al SO, come l'accesso ai file.

1.1.3 Gestione degli Eventi

Quando si verifica un evento:

1. **Salvataggio dello stato della CPU**: Il Program Counter (PC) e i registri della CPU vengono salvati in appositi registri speciali per poter riprendere l'esecuzione successivamente.
2. **Esecuzione del codice del SO**: Il **PC** viene aggiornato con l'indirizzo del codice del **SO** che gestisce l'evento, memorizzato in una tabella detta *vettore delle interruzioni*. Questo vettore contiene puntatori a differenti routine di gestione eventi.
3. **Return**: Una volta gestito l'evento, il SO ripristina lo stato precedente e l'esecuzione del programma sospeso riprende.

Note:-

Nel Program Counter viene scritto l'indirizzo in RAM della porzione di codice del So che serve a gestire l'evento che si è appena verificato. All'accensione del computer, il SO stesso carica in aree della RAM che il SO riserva a se stesso le varie porzione di codice eseguibile che dovranno entrare in esecuzione quando si verifica un eccezione.

Osservazioni 1.1.1

Nei primi N indirizzi della RAM viene caricato una array di puntatori noto come **vettore delle interruzioni**. Ogni entry del vettore contiene l'indirizzo di partenza in RAM di una delle porzioni di codice del SO del punto precedente.

Quando un certo vento si verifica, il program counter viene aggiornato con il valore che è indicato nella cella di memoria collegata all'entry point dell'eccezione. L'ultima istruzione di ogni procedura di gestione di un evento sarà sempre una istruzione di "return from event" (**ra**).

1.2 Struttura della Memoria

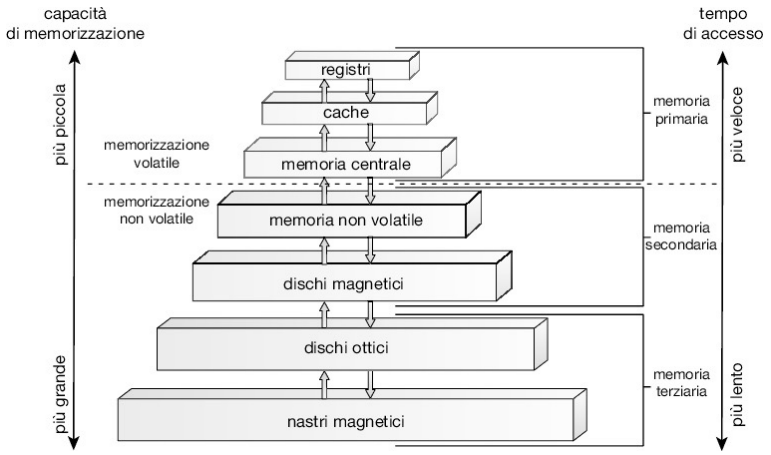
Nel contesto del **SO**, ci sono due principali tipi di memoria:

- **Memoria Principale** (*RAM*): Memoria primaria in cui risiedono programmi e dati durante l'esecuzione.
- **Memoria Secondaria**: Memoria di massa, come **hard disk** o **memorie a stato solido**, utilizzata per la conservazione permanente dei dati.

1.3 Gerarchia delle Memorie

Nella figura: *Velocità implica complessità maggiore, costo maggiore e capacità minore.*

- **Caching**: Ogni livello di memoria fa da cache per il livello successivo. Esempio: la **RAM** fa da cache per l'**hard disk**, la **CACHE** per la **RAM**, e i **registri della CPU** per la **CACHE**.



Domanda 1.1

Sarebbe bello avere 500GB di registri di CPU o Hard Disk veloci quando i registi della CPU?

Per una informazione la si deve copiare in una memoria più veloce (più costosa). La **RAM** fa da cache per l'HDD. La **CACHE** fa da cache per la RAM I **REGISTRI** fanno da cache per la CACHE.

Due tecnlogie di memoria RAM

- SRAM: per la cache e i registri della CPU
  - Creato con i FLIP-FLOP, questo porta un costo maggiore ma una maggiora efficienza rispetto alla DRAM.
- DRAM: per la memoria principale/centrale
  - Creato con i Condensatori, che tendono a perdere il loro stato. Questo obbliga a refresharli costantemente portando un dispendio maggiore di energie ma un minore costo di produzione.

1.4 Struttura di I/O

Un generico computer è composto da una CPU e da un insieme di dispositivi di I/O connessi fra loro da un bus comune. Ogni dispositivo di I/O è controllato da un apposito componente hardware detto **controller** Il controller è a sua volta un piccolo processore, con alcuni registri e una memoria interna, detto **buffer**. il SO interagisce con il controller attraverso un software apposito noto come **driver del dispositivo**.

Esempio 1.4.1 (Driver)
Il driver del dispositivo, carica nei registri dei controller opportuni valori che specificano le operazioni da compiere.
Il controller esamina i registri e intraprende l'operazione corrispondente.
Il controller trasferisce i dati dal dispositivo al proprio buffer.
Il controller invia un interrupt al SO indicando che i dati sono pronti per essere prelevati.

Questo modo di gestire l'IO con grandi quantità di dati è molto **inefficiente**. Una soluzione utile è avere un canale di comunicazione diretto tra il dispositivo e la RAm, in modo da non "disturbare" troppo il SO. Tale canale è detto **Direct Memory Access** (DMA). Il SO, tramite il driver del disco, istruisce opportunamente il controller del disco, con un comando (scritto nei registri del controller) del tipo:

Osservazioni 1.4.1
Trasferisci il blocco numero 1000 del disco in RAM a partire dalla locazione di RAM di indirizzo F2AF

Il controller trasferisce direttamente il blocco in RAM usando il DMA, e ad operazione conclusa avverte il SO mediante un interrupt opportuno.

1.5 Multitasking e Time-Sharing

Quando lanciamo un programma, il SO cerca il codice del programma sull'hard disk, lo copia in RAM, e *"fa partire il programma"* Noi utenti del SO non dobbiamo preoccuparci di sapere dov'è memorizzato il programma sull'hard disk, né dove verrà caricato in RAM per poter essere eseguito.

Dunque, il SO rende **facile** l'uso del computer. Ma il SO ha anche il compito di assicurare un uso **efficiente** delle risorse del computer, in primo luogo la CPU stessa.

Osservazioni 1.5.1
Consideriamo un programma in esecuzione: a volte deve fermarsi temporaneamente per compiere una operazione di I/O (esempio: leggere dall'hard disk dei dati da elaborare).
Fino a che l'operazione non è completata, il programma non può proseguire la computazione, e non usa la CPU.
Invece di lasciare la CPU inattiva, perché non usarla per far eseguire il codice di un altro.

Questo è il principio della multiprogrammazione (multitasking), implementato da tutti i moderni SO: il SO mantiene in memoria principale il codice e i dati di più programmi che devono essere eseguiti. (Detti anche job)

Domanda 1.2

Alcune applicazioni degli utenti però sono per loro natura interattive, come fa ad esserci una interazione continua tra il programma e l'utente che lo usa?

Oltre a questo, i sistemi di calcolo son multi-utente cioè permettono di essere connessi al sistema e di usare "contemporaneamente" il sistema stesso.

1.5.1 Time-Sharing

È meglio allora **distribuire** il tempo di CPU fra i diversi utenti (i loro programmi in "esecuzione") frequentemente (ad esempio ogni 1/10 di secondo) così da dare una impressione di **simultaneità** (che però è solo apparente).

Questo è il **time-sharing**, che estende il concetto di **multiprogrammazione**, ed è implementato in tutti i moderni sistemi operativi.

1.6 Compiti del sistema operativo

E' necessario tenere traccia di tutti i programmi **attivi** nel sistema, che stanno usando o vogliono usare la CPU, e gestire in modo appropriato il passaggio della CPU da un programma all'altro, nonché **lanciare** nuovi programmi e **gestire** la terminazione dei vecchi.

Note:-

Questo è il problema della gestione dei processi (cap. 3) e dei thread (cap. 4)

.

Quando la CPU è libera, e più programmi voglionousare, a quale programma in RAM assegnare la CPU?

Note:-

Questo è il problema di CPU Scheduling (cap. 5)

I programmi in esecuzione devono interagire fra loro senza danneggiarsi ed evitando situazioni di stallo (ad esempio, il programma A aspetta un dato da B che aspetta un dato da C che aspetta un dato da A)

Note:-

Questi sono i problemi di sincronizzazione (cap. 6/7) e di deadlock (stallo dei processi) (cap. 8)

Come gestire la RAM, in modo da poterci far stare tutti i programmi che devono essere eseguiti? Come tenere traccia di quali aree di memoria sono usate da quali programmi?

Note:-

La soluzione a questi problemi passa attraverso i concetti di gestione della memoria centrare (cap. 9) e di memoria virtuale (cap. 10).

Infine, un generico computer è spesso soprattutto un luogo dove gli utenti **memorizzano** permanentemente, organizzano e recuperano vari tipi di informazioni, all'interno di "contenitori" detti **file**, a loro volta suddivisi in cartelle (o folder, o directory) che sono organizzate in una struttura gerarchica a forma di albero (o grafo aciclico) nota come **File System**.

Note:-

Il SO deve gestire in modo efficiente e sicuro le informazioni memorizzate nella memoria di massa (o secondaria) (cap. 11) deve permettere di organizzare i propri file in modo efficiente, ossia fornire una adeguata interfaccia col file system (cap. 13), deve implementare il file system (cap. 14)

Domanda 1.3

come fa il SO a mantenere sempre il controllo della macchina?

Soprattutto, come fa anche quando non sta girando? Ad esempio, come evitare che un programma utente acceda direttamente ad un dispositivo di I/O usandolo in maniera impropria? Oppure, che succede se un programma, entra in un loop infinito? E' necessario prevedere dei modi per proteggersi dai malfunzionamenti dei programmi utente (voluti, e non)

1.6.1 Duplice modalità di funzionamento

Nei moderni processori le istruzioni macchina possono essere eseguite in due modalità diverse:

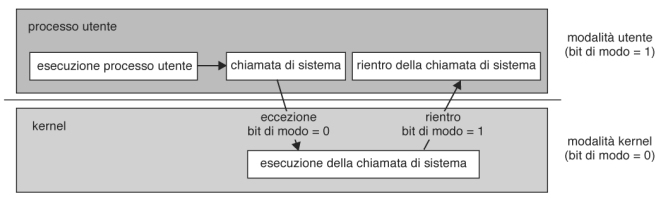
1. normale (modalità utente)
2. di sistema (modalità priviligiata, o kernel / monitor / supervisor mode)

La CPU è dotata di un "bit di modalità" di sistema (0) o utente (1), che permette di stabilire se l'istruzione corrente è in esecuzione per conto del SO o di un utente normale.

Osservazioni 1.6.1

Le istruzione macchina **sensibili**, nel senso che se usate male possono danneggiare il funzionamento del sistema nel suo complesso, possono essere eseguite solo in modalità di sistema, e quindi solo dal SO, altrimenti se nel codice di un programma normale in esecuzione è contenuta una istruzione delicata, quando questa istruzione entra nella CPU viene generata una **trap**.

I programmi utente hanno a disposizione le **system call** (chiamate di sistema) per compiere operazioni che richiedono l'esecuzione di istruzioni privilegiate.
Una system call si usa in un programma come una normale subroutine, ma in realtà provoca una **eccezione**, e il controllo passa al codice del SO di gestione di quella eccezione.
Ovviamente, quando il controllo passa al SO, il bit di modalità viene settato in modalità di **sistema** in modo automatico, via **hardware**.



Si dice di solito che il processo utente sta eseguendo in **kernel mode**

1.6.2 Timer

Domanda 1.4

Che succede se un programma utente, una volta ricevuto il controllo dalla CPU, si mette ad eseguire il seguente codice: $for(;;)i++;$?

Per evitare questo tipo di problemi, nella CPU è disponibile un **Timer**, che viene inizializzato con la quantità di tempo che si vuole concedere **consecutivamente** al programma in esecuzione.
Qualsiasi cosa faccia il programma in esecuzione, dopo 1/10 di secondo il Timer invia un **interrupt** alla CPU, e il controllo viene restituito al sistema operativo. Il SO **verifica** che tutto stia procedendo regolarmente, riinizializza il Timer e decide quale programma mandare in esecuzione, questa è l'essenza del **time-sharing**

Osservazioni 1.6.2

Ovviamente, le istruzioni macchina che gestiscono il timer, sono istruzioni privilegiate.
Altrimenti un programa utente potrebbe modifcare semplicemente i valori :D

1.6.3 Protezione della memoria

Domanda 1.5

Cosa succede se un programma in esecuzione scrive i dati di un altro programma in "esecuzione"?

E' necessario proteggere la memoria primaria da accessi ad aree riservate.

Due possibili soluzioni

Una possibile soluzione: in due registri appositi della CPU (base e limite) il SO carica gli indirizzi di inizio e fine dell'area di RAM assegnata ad un programma.
Ogni indirizzo I generato dal programma in esecuzione viene **confrontato** con i valori contenuti nei registri base e limite.

$$\text{Se } I < \text{base} \,||\, I > \text{limite} \implies \text{TRAP!}$$

I controlli vengono fatti in parallelo a livello hardware, altrimenti richiederebbero troppo tempo.

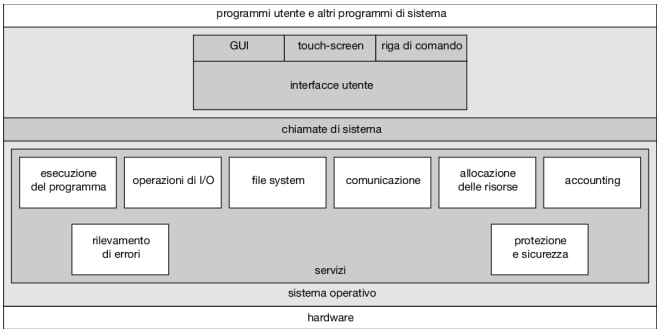
Un'altra variante: simile: in due registri appositi della CPU il SO carica rispettivamente l'indirizzo di inizio (base) e la dimensione (offset) dell'area di RAM assegnata ad un programma.

$$\text{Se } I < \text{base} \,||\, I > \text{base} + \text{offset} \implies \text{TRAP!}$$

2

Strutture dei Sistemi Operativi

Un sistema operativo mette a disposizione degli utenti (e dei loro programmi) molti servizi



Alcuni di questi servizi sono completamente invisibili agli utenti, altri sono parzialmente visibili, e altri sono direttamente usati dagli utenti. Ma il *grado di visibilità* dipende anche dal tipo di utente (Root, user, group)

Esempio 2.0.1 (Esempi di visibilità)

- Interfaccia col sistema operativo (terminale) (visibili)
- Chiamate di sistema (quasi sempre visibili)
- Gestione di processi (praticamente invisibili

2.1 Interfaccia del Sistema Operativo

L'interfaccia è lo strumento con il quale gli utenti interagiscono con il So, e ne sfruttano i servizi offerti.
Può essere un **interprete di comandi**, o un'**interfaccia grafica** con finestre e menù, ma di solito è possibile usare una combinazione di entrambi

2.1.1 Interprete dei comandi

Normalmente non fa parte del **kernel** SO; ma è un programma (o collezione di essi) fornito insieme al SO.
Un esempio d'interprete è la **shell** dell'MS-Dos oppure la **shell** Unix.
Una shell rimane semplicemente in attesa di ciò che l'utente scrive da linea di comando, ed ovviamente, esegue

il comando stesso. Spesso, i comandi che possonoe ssere usati dagli utenti del SO sono dei semplici **eseguibili**. L'interprete si occupa di trovare sull'hard disk e lanciare il codice dell'eseguibile passando eventuali argomenti specificati.

Esempio 2.1.1 (Comando shell unix)

1. L'utente scrive *rm myfile*
2. l'interprete cerca un file eseguibile di nome "rm" e lo lancia, passandogli come parameteo "myfile"

Note:-

Un comando utile può essere *ps* che ti permette di vedere i processi attaccati alla tua shell

2.2 Interfaccia grafica

I moderni SO offrono anche una interfaccia grafica (GUI) per gli utenti, spesso più facile da imparare ed usare. **Unix** offre varie interfaccie grafiche, sia proprietarie che open-source, come **KDE** e **GNOME**, e ogni utente del SO può scegliersi la sua

2.3 Programmi/servizi di Sistema

Non fanno parte del kernel del SO, ma vengono forniti insieme al SO, e rendono più facile, comodo e conveniente l'uso del Sistema.
Gli interpreti dei comandi e le interfacce grafiche sono gli esempi più evidenti di programmi di sistema.
Oltre a questi: editor, compilatori, browser, task manager etc etc.

2.4 Chiamate di sistema (Syscall)

Da ora in poi, chiameremo un programma in "esecuzione" come **processo** Le system call costituiscono la vera e propria interfaccia tra i processi degli utenti e il Sistema Operativo.
Ad esempio, in Unix assumono la forma di procedure che possono essere inserite direttamente in programmi scritti con linguaggi ad alto livello (C, C++, ...)
Sembra di usare una **subroutine**, ma l'esecuzione della system call trasferisce il controllo al SO, e in particolare alla porzione di codice del SO che implementa la particolare System Call invocata.
Ad esempio, in un programma C, per scrivere dentro ad un file:

```
fd = open("nomefile", O_WRONLY);
i = write(...)
close(fd)
```

Open, write e close sono delle syscall

2.4.1 Chiamate di sistema: le "API"

Application Programming Interface Le API non sono ltro che uno strato intermedio tra le applicazione sviluppate dai programmatori e le syscall, per rendere più **facile** l'uso e migliorare la **portabilità** tra versioni.

Esempio 2.4.1 (Chiamate)

Ad esempio, la libreria C dell'ambiente Unix è una semplice forma di API. In questa libreria esiste la funzione per aprire un file:
fopen, *fprintf* e *fclose*

2.5 Gestione dei processi

In un dato istante, all'interno di un SO sono attivi più processi (anche se uno solo è in esecuzione, in un dato istante). Si parla allora di **Processi Concorrent**, perchè si contendono l'uso delle risorse hardware della macchina.

1. La CPU
2. Lo spazio in memoria primaria e secondaria
3. I dispositivi di input e output

Il SO ha la responsabilità di fare in modo che ogni processo abbia la sua parte di risorse, senza danneggiare e venire danneggiato dal altri processi.

Il SO quindi deve gestire tutti gli aspetti riguardo la vita dei processi.

- Creazione e cancellazione dei processi
- Sospensione e riavvio dei processi
- Sincronizzazione tra i processi
- Comunicazione tra processi

Per eseguire un programma deve essere caricato in memoria principale.
In un sistema time-sharing, più processi possono essere contemporaneamnte attivi: il loro codie e i loro dati sono caricati in qualche area della RAM. Quindi il SO deve:

- Tenere traccia di quali parti della RAM sono utilizzati e da quale processo
- Distruibire la RAM tra i processi
- Gestire la RAM in base alla necessità e ai cambiamenti

2.6 Gestione dei file e del filesystem

Quasi ogni informazione presente in un sistema è contenuta in un file: una raccolta di informazioni denotata da un nome (e di solito da altre proprietà).

I file sono organizzati in una struttura **gerarchica** detta File System, mediante le cartelle (o directory, o folder) Il SO e' responsabile della: – Creazione e cancellazione
- Fornitura di strumenti per gestire i file e dir
- Memorizzazione efficiente del file system in memoria secondaria.

I file sono memorizzati permanentemente in memoria secondaria, di solito su un hard disk.
Il SO deve:
– decidere dove e come memorizzare i file su disco, ed essere in grado di ritrovarli velocemente.
– Trovare spazio libero velocemente quando un file è creato o aumenta di dimensione, e recuperare spazio alla rimozione di un file.
– Gestire efficientemente accessi concorrenti ai file dai vari processi attivi.

2.7 Macchine Virutali

Un moderno SO trasforma una macchina reale in una sorta di macchina virtuale (MV).

3
Gestione dei processi

3.1 Processi

Il processo è l'unità di lavoro del sistema operativo, perché ciò che fa un qualsiasi SO è innanzi tutto amministrare la vita dei processi che girano sul computer gestito da quel SO. Il sistema operativo è responsabile della creazione e cancellazione dei processi degli utenti, gestisce lo scheduling dei processi, fornisce dei meccanismi di sincronizzazione e comunicazione fra i processi.

3.1.1 Concetto di processo

- Un **processo** è più di un semplice programma in esecuzione, infatti, ha una struttura in memoria primaria, suddivisa in più parti assegnategli dal sistema operativo (vedi fig. 3.1).
- Le principali componenti della struttura di un processo sono:
  - **Codice** da eseguire (il "testo")
  - **Dati**
  - **Stack** (per le chiamate alle procedure/metodi e il passaggio dei parametri)
  - **Heap** (memoria dinamica)
- La somma di queste componenti forma l'immagine del processo:

codice + dati + stack + heap = immagine del processo

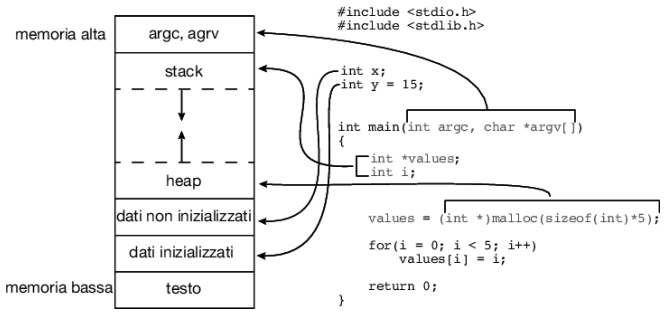


Figure 3.1: Concetto di processo

È anche corretto osservare che attraverso un programma si possono definire più processi, infatti:

- Lo stesso programma può contenere codice per generare più processi
- Più processi possono condividere lo stesso codice

Tuttavia, la distinzione fondamentale tra processo e programma è che un processo **è un'entità attiva**, mentre un programma è **un'entità statica**.

Domanda 3.1

Lo stesso programma lanciato due volte può dare origine a due processi diversi (perché?)

Attenzione: processo, task, job sono **sinonimi**.

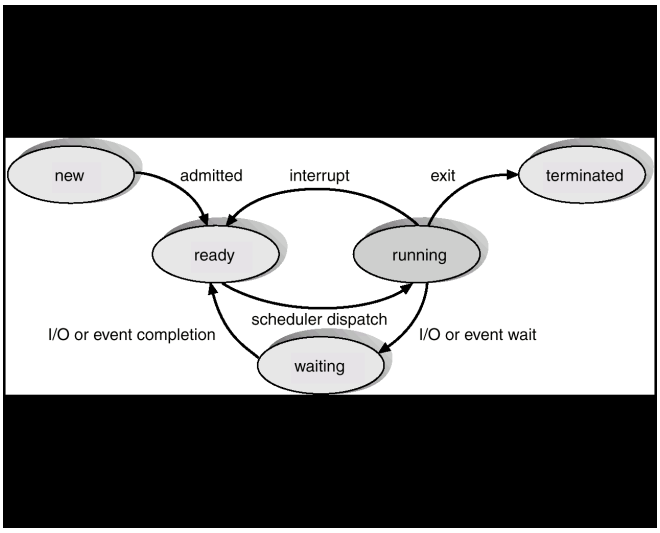
Un programma si **trasforma** in un processo quando viene lanciato, con il doppio click o da riga di comando. Un processo può anche **nascere** a partire da un altro processo, quando quest'ultimo esegue una opportuna system call (fork, spawn, etc)

Definizione 3.1.1: Processo

In realtà, non sono due meccanismi distinti: un processo nasce sempre a partire da un altro processo, e sempre sotto il controllo e con l'intervento del SO (con un'unica eccezione, all'accensione del sistema).

3.1.2 Stato del processo

Da quanto nasce a quando termina, un processo passa la sua esistenza muovendosi tra un insieme di stati, e in ogni stante ogni processo si trova in un ben determinato stato.
Lo stato di un processo evolve a causa del codice eseguito e dell'azione del SO sui processi presenti nel sistema in un dato istante, secondo quanto illustrato dal diagramma di transizione degli stati di un processo.



Gli stati

Gli stati in cui può trovarsi un processo sono:

Definizione 3.1.2: Stati del processo

- **New**: Il processo è appena stato creato
- **Ready (to Run)**: Il processo è pronto per entrare in esecuzione
- **Running**: La CPU sta eseguendo il codice del processo
- **Waiting**: Il provesso ha lasciato la CPU e attende il completamento di un evento
- **Terminated**: Il processo è terminato, il SO sta recuperando le strutture dati e le aree di memoria liberate

Il diagramma di transizione degli stati di un processo sintetizza una serie di possibili varianti del modo in cui un sistema operativo (SO) può amministrare la vita dei processi di un computer.

- Infatti, nel caso reale lo sviluppatore del SO dovrà decidere quali scelte implementative fare quando (ad esempio):
  - Mentre il processo P_x è *running*, un processo entra nello stato *Ready to Run*
  - Mentre il processo P_x è *running*, un processo più importante di P_x entra nello stato *Ready to Run*
  - Mentre il processo P_x è nello stato *Ready to Run*, un processo più importante di P_x entra nello stato *Ready to Run*

Domanda 3.2

Che significato ha eliminare l'arco "interrupt"?

Di avere un sistema non time-sharing

3.1.3 Processo Control Block (PCB)

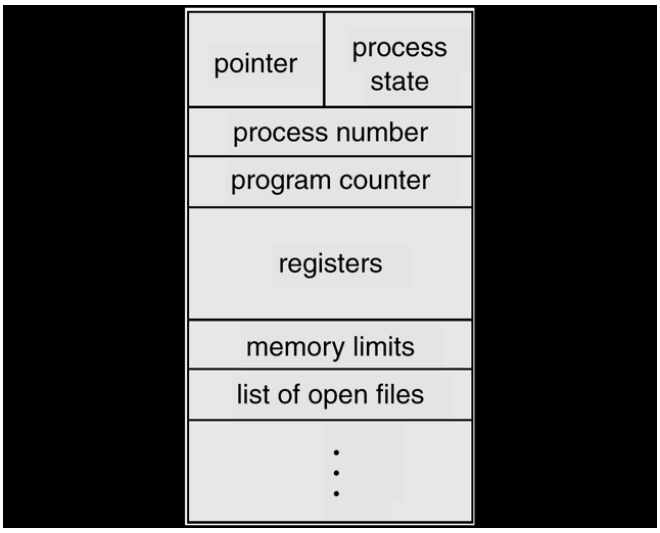
Per ogni processo, il sistema operativo (SO) mantiene una struttura dati chiamata *Process Control Block* (PCB), che contiene le informazioni necessarie per amministrare la vita di quel processo, tra cui:

- Il numero del processo (o *Process ID*)(PID)
- Lo stato del processo (*ready*, *waiting*,...)
- Il contenuto dei registri della CPU salvati nel momento in cui il processo è stato sospeso (valori significativi solo quando il processo non è *running*)
- Gli indirizzi in RAM delle aree dati e codice del processo
- I file e gli altri dispositivi di I/O correntemente in uso dal processo
- Le informazioni per lo *scheduling* della CPU (ad esempio, quanta CPU ha usato fino a quel momento il processo)

3.2 Scheduling dei processi

Conosciamo già i seguenti due concetti:

- **Multiprogrammazione**: avere sempre un processo *running* $\Rightarrow$ massima utilizzazione della CPU.
- **Time Sharing**: distribuire l'uso della CPU fra i processi a intervalli prefissati. Così più utenti possono usare "allo stesso tempo" la macchina, e i loro processi procedono in "parallelo" (notate sempre le virgolette).



Definizione 3.2.1: Scheduling

Per implementare questi due concetti, il sistema operativo deve decidere periodicamente quale sarà il prossimo processo a cui assegnare la CPU. Questa operazione è detta *Scheduling*.

In un sistema time sharing single-core, attraverso lo scheduling, ogni processo "crede" di avere a disposizione una macchina "tutta per se"... Ci pensa il SO a farglielo credere, **commutando** la CPU fra i processi (ma succede la stessa cosa in un sistema ad n-core se ci sono più di n processi attivi contemporaneamente)

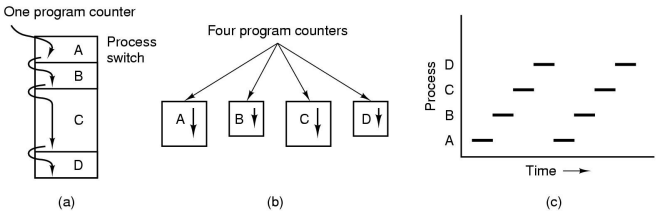


Figure 3.2: a) Ciò che succede in realtà
b) ciò che vede ogni singolo processo
c) Il risultato finale

3.2.1 Il cambio di contesto (context switch)

Per commutare la CPU tra due processi, il sistema operativo deve:

1. Riprendere il controllo della CPU (ad esempio attraverso il meccanismo del *Timer* visto nel capitolo 1).
2. Con l'aiuto dell'hardware della CPU, salvare lo stato corrente della computazione del processo che lascia la CPU, ossia copiare il valore del *Program Counter* (PC) e degli altri registri nel suo *Process Control Block* (PCB).
3. Scrivere nel PC e nei registri della CPU i valori relativi contenuti nel PCB del processo utente scelto per entrare in esecuzione.

   Questa operazione prende il nome di: **cambio di contesto**, o *context switch*.
   Notate che, tecnicamente, anche il punto 1 è già di per sé un *context switch*.

- Il *context switch* richiede tempo, perché il contesto di un processo è composto da molte informazioni (alcune le vedremo quando parleremo della gestione della memoria).

- Durante questa frazione di tempo, la CPU non è utilizzata da alcun processo utente.
- In generale, il *context switch* può costare da qualche centinaio di nanosecondi a qualche microsecondo.
- Questo tempo "sprecato" rappresenta un *overhead* (sovraccarico) per il sistema e influisce sulle sue prestazioni.

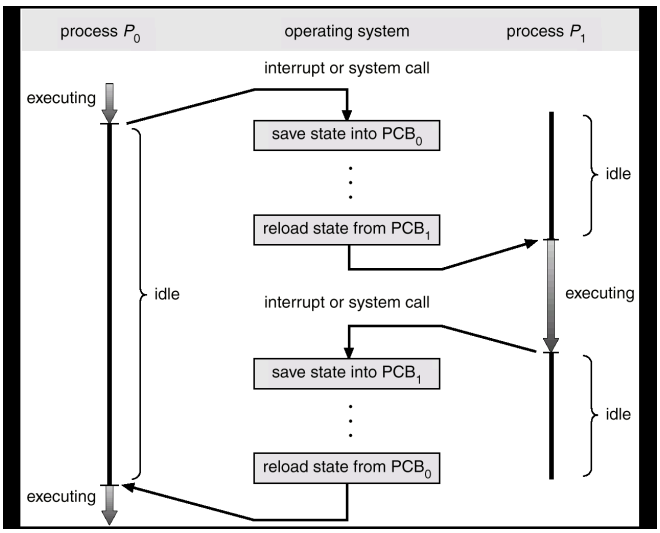
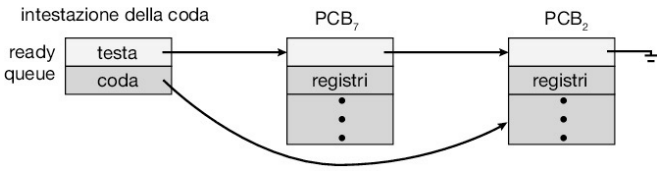


Figure 3.3: Fasi dello scheduling tra un processo e un altro

3.2.2 Code di scheduling

Per **amministrare** la vita di ciascun processo, il SO gestisce varie **code** di processi. Ogni processo "si trova" in una di queste code, a seconda di cosa sta facendo. Una coda di processi non è altro che una lista di PCB, mantenuta in una delle aree di memoria primaria che il SO riserva a se stesso.
La coda dei processi più importante è la coda **ready**, o **ready queue (RQ)**: l'insieme dei processi **ready to run**. Quando un processo rilascia la CPU, ma non termina e non torna nella *ready queue*, vuol dire che si è messo



in **attesa** di "qualcosa", e il SO lo "parcheggia" in una tra le possibili code, che possiamo dividere in due grandi categorie:

- **Device queues**: code dei processi in attesa per l'uso di un dispositivo di I/O. Una coda per ciascun dispositivo.

  Esempio 3.2.1 (Esempi)
  - Una coda d'attesa per il primo hard disk
  - Una coda per l'ssd
  - Una coda per la stampante, etc..

- **Code di waiting**: code di processi in attesa che si verifichi un certo evento. Una coda per ciascun evento (ci torneremo nella sezione 6.6).

Dunque, durante la loro vita, i processi si spostano (meglio: il SO sposta i corrispondenti PCB) tra le varie code. Quindi lo stato **waiting** nel diagramma di transizione degli stati di un processo **corrisponde a più code di attesa**

Possiamo riformulare il diagramma di transizione degli stati di un processo come un **diagramma di accodamento** in cui i processi si muovono fra le varie code

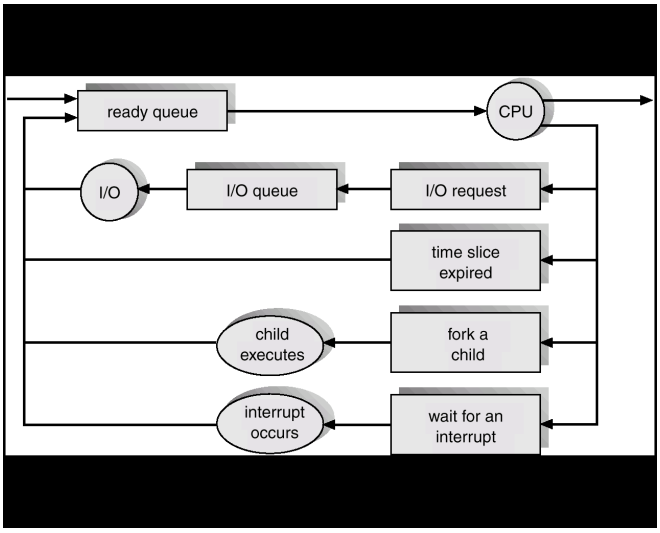


Figure 3.4: SX: new, DX: Terminated

3.2.3 CPU Scheduler

Un componente del Sistema Operativo detto *CPU Scheduler* sceglie uno dei processi nella coda *ready* e lo manda in esecuzione.

- Il *CPU scheduler* si attiva ogni 50/100 millisecondi, ed è responsabile della realizzazione del *time sharing*.
- Per limitare l'*overhead*, deve essere molto veloce.
- Il *CPU scheduler* è anche chiamato *Short Term Scheduler*.

3.3 Operazione sui processi

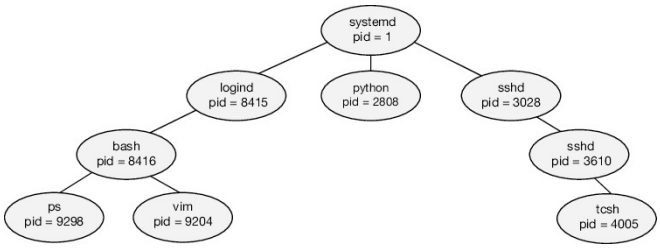
La creazione di un processo è di gran lunga l'operazione più importante all'interno di qualsiasi sistema operativo. Ogni SO possiede almeno una *System Call* per la creazione di processi, e ogni processo è creato a partire da un altro processo usando la system call relativa (eccetto il processo che nasce all'accensione del sistema).
Il processo "creatore" è detto *processo padre* (o *parent*).
Il processo creato è detto *processo figlio* (o *child*).

> **Osservazioni 3.3.1**
>
> Poiché ogni processo può a sua volta creare altri processi, nel sistema si forma un "albero di processi".

3.3.1 Creazione di un processo

Quando nasce un nuovo processo, il SO:

- gli assegna un identificatore del processo unico, un numero intero detto **pid** (process-id). È il modo con cui il SO conosce e si riferisce a quel processo.
- recupera dall'hard disk il codice da eseguire e lo carica in RAM (a meno che il codice non sia già in RAM).



- alloca un nuovo *PCB* e lo inizializza con le informazioni relative al nuovo processo.
- inserisce il *PCB* in coda *ready*.

Domanda 3.3

Che cosa fa il processo padre quando ha generato un processo figlio?

- Prosegue la sua esecuzione in modo concorrente all'esecuzione del processo figio, oppure:
- Si ferma, in attesa del completamento dell'esecuzione del processo figlio

Domanda 3.4

Quale codice esegue il processo figlio?

- al processo figlio viene data una copia del codice e dei dati in uso al processo padre, oppure:
- al processo figlio viene dato un nuovo programma, con eventualmente nuovi dati.

3.3.2 Creazione di un processo in Unix

```
int main() {
    /* fig. 3.8 modificata */
    pid_t pid, childpid;

    pid = fork(); /* genera un nuovo processo */
    printf("questa la stampano padre e figlio");

    if (pid == 0) {
        /* processo figlio */
        printf("processo figlio");
        execlp("/bin/ls", "ls", NULL);
    } else {
        /* processo padre */
        printf("sono il padre, aspetto il figlio");
        childpid = wait(NULL);
        printf("il processo figlio   terminato");
        exit(0);
    }
}
```

3.3.3 Passi dell'SO all'invocazione delle fork

1. Alloca un nuovo *PCB* per il processo figlio e gli assegna un nuovo *PID*; cerca un'area libera in RAM e vi copia le strutture dati e il codice del processo *parent* (si veda più avanti): queste copie verranno usate dal processo figlio.
2. Inizializza il *PC* del figlio con l'indirizzo della prima istruzione successiva alla *fork*.

3. **Nella cella di memoria associata alla variabile che riceve il risultato della *fork* del processo figlio scrive 0.**
4. **Nella cella di memoria associata alla variabile che riceve il risultato della *fork* del processo *parent* scrive il *PID* del figlio.**
5. Mette i processi *parent* e figlio in coda *ready*.

Osservazioni 3.3.2

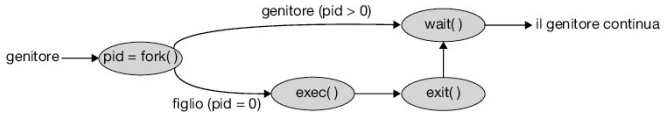
pid == 0 Lo ha solo il processo figlio.
pid = id-child lo ha solo il processo padre.
Così sono in grado di distinguere se sto operando con il figlio o con il padre.

Significato delle altre sys

Execlp: Riceve in input un puntatore ad un file contenente codice eseguibile. Il processo che la invoca prosegue eseguendo il codice specificato, senza più ritornare alla porzione di codice che viene dopo la execlp.
Wait: Invocata da un processo parent, lo sospende fino alla terminazione del processo figlio. La wait restituisce il PID del figlio appena terminato. **Exit:** provoca la terminazione istantanea del processo che la invoca.

Domanda 3.5

Come cambia lo schema se il processo parent non esegue la wait?



3.3.4 Altro esempio

```
int main() {
    /* un altro esempio */
    int a, b, c = 57;
    a = fork(); // genera un nuovo processo
    printf("questa la stampano padre e figlio");

    if (a == 0) {
        /* processo figlio */
        c = 64; // ***
        printf("c = %d", c);
    } else {
        /* processo padre */
        printf("c = %d", c);
        b = wait(NULL);
        printf("b = %d", b);
    }
}
```

3.3.5 Osservazioni

- Il codice viene condiviso tra padre e figlio, evitando duplicazione e spreco di memoria.
- Lo spazio dati viene duplicato:

  - Le modifiche di variabili non sono condivise tra padre e figlio.
  - Le nuove variabili dichiarate dopo la `fork` non sono visibili all'altro processo.

- Un padre può chiamare `fork` più volte, e usare il PID dei figli per tracciarli.
- `fork` restituisce 0 al figlio per distinguerlo dal padre.
- Se `fork` restituisse un valore maggiore di 0 al figlio, non si potrebbe distinguere facilmente tra padre e figlio, complicando la gestione delle operazioni diversificate (come illustrato in fig. 3.8).

3.3.6 Terminazione di un processo

Un processo termina dopo l'esecuzione dell'ultima istruzione del suo codice. Esiste una system call chiamata `exit()` per terminare un processo.
I dati di output, come il `pid`, possono essere inviati al processo padre in attesa della terminazione del figlio. Il sistema operativo **rimuove** le risorse allocate al processo terminato, recuperando la RAM e chiudendo eventuali file aperti.

- Un processo può uccidere esplicitamente un altro processo appartenente allo stesso utente tramite la system call `kill` (in Unix) o `TerminateProcess` (in Win32).
- In alcuni casi, il sistema operativo può decidere di terminare un processo utente, ad esempio se:
  - il processo utilizza troppe risorse.
  - il suo processo padre è morto (in questo caso può avvenire una terminazione a cascata, che non avviene però in Unix o Windows).

3.4 Comunicazione tra processi

Processi indipendenti e cooperanti

I processi attivi in un sistema possono essere classificati come:

- **Indipendenti**: quando non si influenzano esplicitamente durante l'esecuzione.
- **Cooperanti**: quando si influenzano a vicenda per:
  - Scambiarsi informazioni.
  - Collaborare su un'elaborazione suddivisa per efficienza o modularità.

I processi cooperanti necessitano di meccanismi di comunicazione e sincronizzazione.

3.5 Esempio: il problema Produttore-Consumatore

Problema del produttore-consumatore

Un classico problema di processi cooperanti è il *problema del produttore-consumatore*:

- Un **processo produttore** produce informazioni che vengono consumate da un **processo consumatore**.
- Le informazioni sono collocate in un buffer di dimensione limitata.
- Un esempio pratico è un **processo compilatore** (produttore) che genera codice assembler.
- Il **processo assemblatore** (consumatore) traduce il codice assembler in linguaggio macchina.
- L'assemblatore potrebbe poi diventare un produttore per un modulo che carica in RAM il codice.

```
#define SIZE 10

typedef struct {
    // Definizione della struttura dell'item
    ...
} item;

// Buffer condiviso
item buffer[SIZE]; (shared array)

// Variabili condivise
int in = 0, out = 0;
```

Buffer circolare di `SIZE` elementi con due puntatori `in` e `out`:

- `in`: indica la prossima posizione libera nel buffer.
- `out`: indica la prossima posizione piena da consumare.
- **Condizione di buffer vuoto**: `in == out`.
- **Condizione di buffer pieno**: `(in + 1) % SIZE == out`.

Nota: la soluzione utilizza solo `SIZE-1` elementi per evitare conflitti tra la condizione di buffer pieno e vuoto.

3.5.1 Inter-Processo Comunication (IPC)

Domanda 3.6

Come fanno due processi a scambiarsi le informazioni necessarie alla cooperazione?

- Il Sistema Operativo (SO) fornisce dei meccanismi di **Inter-Process Communication (IPC)**.
- Sono disponibili opportune **system call** che permettono a due (o più) processi di:
  - **scambiarsi messaggi** oppure
  - **usare la stessa area di memoria condivisa**, in cui possono scrivere e leggere.

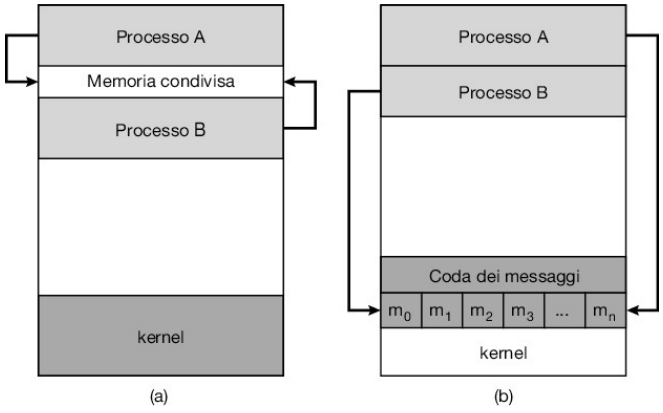


Figure 3.5: a) Memoria condivisa
b) Scambio di messaggi

In entrambi i casi, il SO mette a disposizione delle opportune **system call**. Ad esempio, per lo scambio di messaggi, saranno disponibili delle system call del tipo:

- `line = msgget();`
- `send(message, line, process-id);`
- `receive(message, line, process-id);`

(Nota:) I parametri sono solo indicativi, ogni specifica implementazione avrà il proprio insieme di argomenti.

Saranno necessarie alcune **scelte implementative**.
Nel caso dei **messaggi** (si parla spesso di *code di messaggi*):

- Una coda può essere usata da più di due processi?
- Quanti messaggi può ospitare al massimo una coda?
- Cosa deve fare un processo ricevente se non ci sono messaggi, o un processo trasmittente se la linea è piena?
- Si possono trasmettere messaggi di lunghezza variabile?

Nel caso della **memoria condivisa**:

- Può avere dimensione variabile?
- Quali processi hanno diritto di usarla?
- Cosa succede se la memoria condivisa viene rimossa?

4 Threads

Note:-

Questo capitolo va studiato solo dopo il capitolo 9 sulla memoria centrale. Ai fini del programma del corso e dell'esame, si fa riferimento a queste slide, il cui contenuto è molto semplificato rispetto al capitolo 4 del libro di testo.

Consideriamo due processi che devono lavorare sugli stessi dati. Come possono farlo, se ogni processo ha la propria area dati (ossia, gli spazi di indirizzamento dei due processi sono separati)?

- I due processi possono richiedere al sistema operativo un'area di memoria condivisa, oppure scambiarsi i dati usando messaggi.
- I dati possono essere mantenuti in un file, al quale i due processi accedono a turno.

Sarebbe comodo poter avere processi in grado di lavorare sugli stessi dati senza usare meccanismi espliciti di condivisione/comunicazione, e senza l'utilizzo di file, che risiedono su supporti relativamente lenti. Ad esempio, in un editor di testo:

- Un processo gestisce l'input e i comandi di formattazione dell'utente;
- Un altro processo esegue il controllo automatico degli errori.

In questo caso, i due processi dovrebbero poter lavorare sullo stesso testo, la cui copia corrente è mantenuta in memoria principale. Tuttavia, poiché ogni processo ha un diverso spazio di indirizzamento, come possono lavorare sulla stessa copia dei dati?

Inoltre, durante il *context switch* tra processi, occorre disattivare le aree dati e di codice del processo uscente e attivare quelle del processo entrante.

- Le **cache fisiche della CPU** contengono ancora i dati del processo uscente, quindi il processo entrante inizialmente genera molti *miss cache*.
- Se due (o più) processi potessero condividere dati e codice, il context switch tra di loro sarebbe molto meno oneroso.

Da queste considerazioni nasce il concetto di **thread**: un gruppo di *peer thread* è un insieme di "processi" che condividono lo spazio di indirizzamento (codice e dati).

Definizione 4.0.1: Terminologia

Un processo P (per come è stato studiato finora) è caratterizzato da un unico **Thread** di computazione: una sequenza di istruzioni eseguite, che ovviamente può cambiare da un'esecuzione all'altra in base, ad esempio, ai dati di input. (Fig. 4.1a).

Nessun altro processo ha accesso allo spazio di indirizzamento logico di P, quindi nessun processo utente, oltre a P, può accedere alle aree dati di P. Altri processi possono eseguire lo stesso codice di P, ma in uno spazio di indirizzamento logico separato e quindi con dati diversi. (Fig. 4.1a).

4.1 Processo Multi-Thread

Un **processo Multi-Thread** (o **Multi-Threaded**) è invece composto da più thread di computazione, detti *peer thread*. Un processo multi-threaded è anche detto *task* (Fig. 4.1).

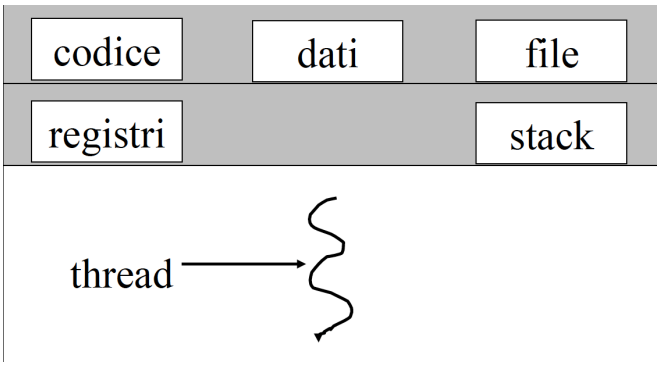


Figure 4.1: thread$_e$*xample*

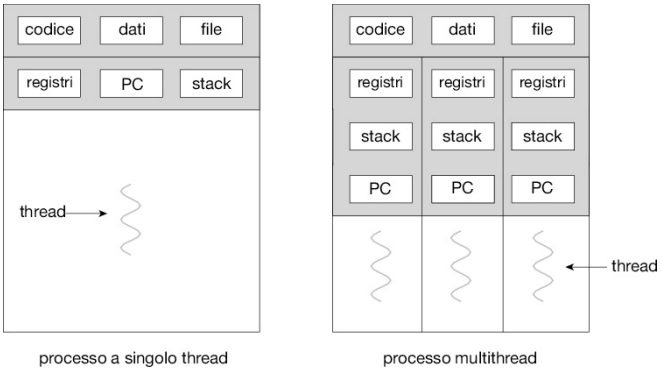


Figure 4.2: MultiThreadExample.png

Definizione 4.1.1: Processo Multi-Thread

Un processo **Multi-Thread** (o **Multi-Threaded**) è composto da più thread di computazione, detti *peer thread*. Un processo multi-threaded è anche detto *task* (Fig. 4.1).

Ad ogni *peer thread* viene assegnata l'esecuzione di codice, che solitamente è diverso da quello degli altri peer thread. Di conseguenza:

- Ogni thread ha uno stato di computazione autonomo, costituito da:
  - **Program Counter** e registri della CPU
  - uno **stack** indipendente
- Tuttavia, un insieme di *peer thread* condivide il codice in esecuzione e, soprattutto, le **aree dati**.

Il codice deve specificare quale *peer thread* esegue ciascuna parte del codice, similmente a come, in un programma che utilizza la `fork`, si può definire la porzione di codice eseguita dal processo padre e quella eseguita dal processo figlio.

Osservazioni 4.1.1

Il *context switch* avviene anche tra ciascun *peer thread* di un processo multi-threaded, per permettere a ciascuno di continuare l'esecuzione del proprio codice assegnato. (Per ora si considera un'architettura **single core**.)

- Il *context switch* tra *peer thread* richiede solo il salvataggio e il ripristino del Program Counter, dei registri della CPU e dello stack, che sono distinti per ogni thread.
- Il codice e i dati, cioè lo **spazio di indirizzamento logico**, rimangono invariati tra *peer thread*, per cui non è necessario cambiare la tabella delle pagine del processo multi-threaded durante il *context switch*.

Il *context switch* tra processi è molto più oneroso per il SO rispetto a quello tra peer thread, poiché nel primo caso devono essere cambiate molte più informazioni.

- A causa della gestione delle cache, il *context switch* tra processi genera inizialmente più *cache miss* rispetto a quello tra peer thread.
- Per questa ragione, i processi normali (quelli studiati finora) sono spesso chiamati **heavy-weight process** (HWP), mentre i *peer thread* sono definiti **light-weight process** (LWP).

All'interno di un *task*, nuovi peer thread possono essere creati tramite apposite *system call*, e a ciascun thread può essere assegnato codice specifico da eseguire, in modo simile a quanto avviene con `fork` ed `exec`. (In **Linux** un nuovo thread si crea con la system call `clone`, mentre in **Windows** con `CreateThread`.)

Note:-

Altre *system call* permettono ai peer thread di sincronizzarsi fra loro, analogamente a come i processi possono sincronizzarsi tramite semafori. La sincronizzazione è fondamentale per garantire un accesso ordinato ai dati condivisi tra i thread. Inoltre, molti linguaggi moderni, come Java, offrono primitive apposite per la programmazione multi-threaded.

In tutti i sistemi operativi moderni, la gestione dello **scheduling** dei thread avviene a livello di *kernel*. Il SO mantiene strutture dati per gestire sia i processi normali che tutti i peer thread di un *task* multi-threaded.

- Quando un thread si blocca volontariamente o termina il suo quanto di tempo, è il SO a gestire l'assegnazione della CPU, decidendo se assegnarla:
  - ad un altro *peer-thread* dello stesso task,
  - ad uno dei *peer-thread* di un altro task,
  - oppure ad un processo normale.

Note:-

In Solaris, la creazione di un nuovo **LWP** (light-weight process) richiede circa 30 volte meno tempo rispetto alla creazione di un **HWP** (heavy-weight process). Inoltre, il *context switch* tra peer thread è cinque volte più rapido rispetto al *context switch* tra processi.

- **Condivisione di dati e risorse**: più thread possono accedere e operare su dati condivisi in modo efficiente, anche se devono essere sincronizzati adeguatamente per evitare condizioni di gara o accessi non sicuri ai dati.
- **Architetture multi-core**: i thread sono particolarmente idonei per essere eseguiti su processori multi-core e, ancor di più, su architetture multithreaded.

In un processore *single-core*, tutti i *peer thread* di un task si alternano in esecuzione esattamente come un insieme di processi (Figura 4.3). Come già osservato, il *context switch* tra vari *peer thread* è meno oneroso rispetto al *context switch* tra processi normali. Tuttavia, un *context switch* con un altro processo rimane oneroso e può causare un degrado nelle prestazioni a causa dei miss di cache generati inizialmente dal processo entrante.

- **Architetture multi-core e task multi-threaded**: le architetture multi-core sono particolarmente adatte alla gestione di task multi-threaded. Supponiamo un sistema **dual-core** in cui sono attivi due task multi-threaded: se si assegna un task a ciascun core, i *context switch* tra thread saranno limitati ai soli *peer thread* del medesimo task, ottimizzando l'efficienza.

- **Bilanciamento del carico**: in pratica, il numero totale di task multi-threaded (e quindi di *peer thread*) è spesso superiore al numero di core disponibili. Il sistema operativo distribuisce quindi i *peer thread* di uno stesso task su core diversi per bilanciare il carico di lavoro in modo ottimale (Figura 4.4 mostra un task con 4 *peer thread* distribuito su due core).

Corollario 4.1.1 Opportunità di esecuzione parallela nei processori moderni

I processori moderni offrono un'opportunità ulteriore per l'esecuzione dei thread. In un processore multi-core, ciascun core è in grado di eseguire fino a 4 o 5 istruzioni in parallelo del programma in esecuzione; questa tecnica è nota come **multiple issue**.

- Per eseguire più istruzioni in parallelo, ogni core deve disporre di più unità funzionali, come le ALU (Arithmetic Logic Units) e le unità di calcolo in virgola mobile. Tale struttura è detta *superscalare*, poiché consente a ciascun core di eseguire in parallelo fino a 4 o 5 istruzioni, migliorando l'efficienza complessiva del sistema.

4.2 CPU/Core multi-threaded

I processori moderni offrono un'importante opportunità per l'esecuzione dei thread grazie alla capacità di eseguire in parallelo più istruzioni per ciclo di clock. Questa tecnica, chiamata **multiple issue**, consente a ciascun core di avviare fino a 4 o 5 istruzioni contemporaneamente. Ciò è possibile grazie alla presenza di più unità funzionali in ogni core, come ALU (Arithmetic Logic Units) e unità per calcoli in virgola mobile, configurando il core con un'architettura **superscalare**.

4.2.1 Limitazioni delle Architetture Superscalari

Nonostante la potenzialità di eseguire più istruzioni in parallelo, spesso non si riesce a sfruttare appieno questa capacità. La causa principale risiede nelle **dipendenze tra istruzioni**:

- Quando un'istruzione B necessita del risultato di un'istruzione A, è obbligatorio eseguire prima A e poi B, impedendo il parallelismo.

- Di conseguenza, molte unità funzionali rimangono inutilizzate (es. solo 2 delle 4 ALU disponibili vengono utilizzate).

Tuttavia, le istruzioni appartenenti a **peer thread diversi** sono generalmente indipendenti, poiché ciascun thread esegue una porzione distinta di codice, anche se accede al medesimo spazio di indirizzamento.

4.2.2 Simultaneous Multi-Threading (SMT)

Per sfruttare al meglio questa indipendenza tra thread, ogni core moderno supporta il **Simultaneous Multi-Threading** (SMT). Questa tecnica consente al core di:

- Eseguire in parallelo istruzioni provenienti da **peer thread distinti**.

- Aumentare l'utilizzo delle unità funzionali, migliorando così la produttività del core.

Note:-

(Per questa immagine) Ad ogni ciclo di clock il core può eseguire istruzioni di peer-thread diversi, ma con un massimo di 2 peer-thread per core.

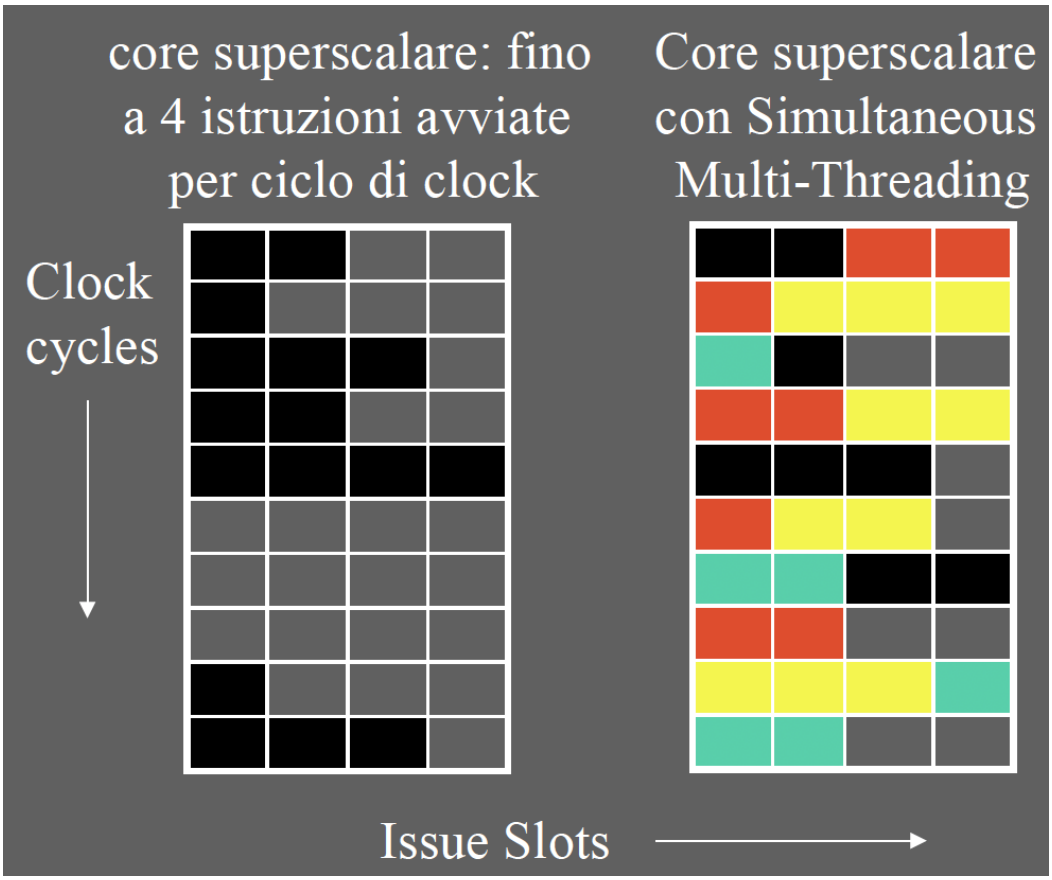


Figure 4.3: Esecuzione parallela con SMT: istruzioni di diversi thread (colori diversi) avviate simultaneamente.

4.2.3 Definizioni: Dual-Threaded vs Dual-Core

L'introduzione del **Simultaneous Multi-Threading** può generare confusione nella terminologia:

- Un **core dual-threaded** è un singolo core capace di eseguire contemporaneamente istruzioni appartenenti a due o più **peer thread**.
- Un **dual-core** è un processore con due core fisici distinti, ciascuno capace di eseguire istruzioni di processi distinti in parallelo.

Note:-

Ogni core di un dual-core può essere a sua volta multi-threaded, combinando le capacità di esecuzione parallela di thread e processi.

4.3 Ma l'SMT è sempre vantaggioso?

Si: ecco lo speed-up ottenuto su uno dei core di una CPU i7 nel passare da un solo thread a 2, per diversi benchmarks.
Il dato "Energy efficiency" ci dice se l'introduzione dell'SMT è vantaggiosa dal punto di vista dell'energia consumata. Un valore superiore a 1.0 significa che l'SMT riduce il tempo di esecuzione più di quanto aumenti i consumi.

Domanda 4.1: difficile

Nessuna CPU implementa una forma di Simultaneous Multi-Processing, in cui cioè istruzioni appartenenti a processi diversi vengono eseguite in parallelo all'interno dello stesso core. Perché?

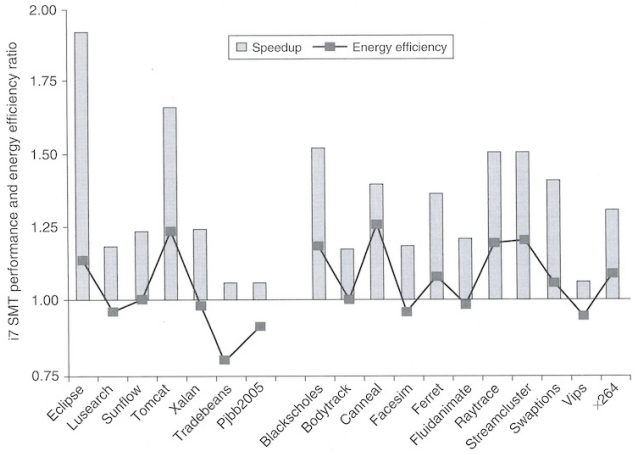


Figure 4.4: smt$_b$*enchmark*

Note:-

Due processi non condividono lo stesso spazio di indirizzamento, quindi non condiviono la stessa tabella delle pagine .
Invece i peer-thread di un task multi-threaded condividono lo spazio di indirizzamento, quindi possono essere eseguiti in parallelo all'interno dello stesso core.

5 Scheduling della CPU

5.1 Scheduling

5.1.1 Fasi di eleborazion e di I/O

Durante la vita di un processo, si alternano fasi di uso della **CPU** (CPU burst) e fasi di attesa per il completamento di operazioni di **I/O** (I/O burst).

Possiamo distinguere due categorie di processi:

- **Processi CPU-bound:** usano intensamente la CPU e interagiscono poco con i dispositivi di I/O (ad esempio un compilatore).
- **Processi I/O-bound:** usano poco la CPU ma fanno ampio uso dei dispositivi di I/O (ad esempio un editor o un browser).

5.1.2 Lo Scheduler della CPU

Consideriamo la situazione in cui un processo utente abbandona la **CPU**. Il Sistema Operativo (SO) si "sveglia" e deve decidere a quale, fra i processi in **Coda di Ready** (processi *ready to run*), assegnare la CPU. Questa operazione è detta **Scheduling della CPU**, e viene gestita dal modulo del SO detto **scheduler**.

Quando interviene lo scheduler per scegliere il successivo processo a cui assegnare la CPU? Possiamo considerare quattro situazioni, che ci porteranno a definire i concetti di **scheduling con** e **senza diritto di prelazione**:

1. Il processo che sta usando la CPU passa volontariamente dallo stato di running allo stato di waiting.
2. Il processo che sa usando la CPU termina.
   - In questi **due casi**, lo scheduler deve prendere un processo dalla coda di readt e mandarlo in eseucuzione
   - **Note:-**
     Un sistema operativo che intervenga nei casi 1 e 2 è sufficiente per implementare il multi-tasking
   - **Domanda 5.1**
     Che succede se mandiamo in eseucuzione un programma che contiene una istruzione del tipo `while(true) printf("who's carr?");`
   - Il **SO** deve poter intervenire in modo da evitare che un processo si simpossessi della CPU, quindi...

3. Il processo che sta usando la CPU viene obbligato a passare dallo stato di running allo stato di ready
   - Il passaggio non avviene mai **volontariamente**, il processo non vorrebbe lasciare la CPU a favore di qualcun'altro.
   - Nei sistemi time sharing il SO non perde **mai** completamente il controllo del sistema
   - Il SO mantiene il controllo della CPU attraverso un timer hardware, allo scadere del timer il controllo della CPU verrà restituito al SO, che sceglierà un processo dalla RQ un processo da mandare in esecuzione.
4. Un processo Px entra in coda di ready arrivando da una coda di wait oppure perchè è appena stato lanciato.

Domanda 5.2

Perchè il SO interviene in questo caso?

- **Primo:** i processi non si spostano autonomamente da una coda all'altra. È il Sistema Operativo (SO) che gestisce i loro **PCB** (Process Control Block). Ad esempio, quando il SO si accorge del completamento di un'operazione di I/O per cui il processo P_x era in attesa, interviene per spostare (il PCB di) P_x dalla **coda di wait** alla **coda di ready**.
- **Secondo:** se il processo P_x risulta più "importante" rispetto al processo attualmente in esecuzione, il SO può decidere di togliere quest'ultimo dalla **CPU** e mandare in esecuzione P_x.

Quando un sistema interviene solo nei casi 1 e 2 si parla di: **Scheduling senza (diritto di) prelazione**.
Quando un sistema interviene anche nei casi 3 e 4 si parla di: **Scheduling con (diritto di) prelazione**
Chiaramente, lo **scheduling preemptive** è più sicuro per gli utenti, ma la sua implementazione richiede un sistema operativo e un'architettura hardware più sofisticati (ad esempio, un **timer dedicato**).
I moderni **sistemi operativi general purpose** usano tutti una qualche variante di **scheduling preemptive**. Tuttavia, per applicazioni specifiche può essere sufficiente uno **scheduling non-preemptive**, permettendo l'uso di sistemi operativi più semplici e leggeri.
Lo **scheduling preemptive** può portare a situazioni che necessitano di essere gestite con attenzione. Ad esempio, consideriamo un processo che deve compiere un'operazione di **I/O** e chiama la relativa system call. Il controllo viene trasferito al sistema operativo, che inizia l'operazione per conto del processo utente. Nel frattempo, scade il timer e il controllo viene passato a un'altra porzione del codice del SO.
Di conseguenza, una operazione delicata (altrimenti non sarebbe stata gestita dal SO) viene interrotta a metà, e le **strutture dati** potrebbero trovarsi in uno stato inconsistente poiché la system call di **I/O** non ha finito di aggiornarle.

Domanda 5.3

Cosa succede se ora la CPU viene data a un altro processo utente che tenta di usare lo stesso dispositivo di **I/O** che il processo precedente stava utilizzando? Quale semplice soluzione può essere adottata in tali casi?

Note:-

Vogliamo che il processo riesca a completare la richiesta al controller dell'I/O, niente di più

. Mentre **Unix** è stato sviluppato fin dall'inizio come sistema di tipo **preemptive**, nei sistemi **Microsoft** la preemption è stata introdotta solo con **Windows 95**. Questo è dovuto al fatto che i sistemi operativi della famiglia **MS-Dos** sono nati come sistemi **mono-utente**, per i quali era sufficiente un sistema operativo più semplice. Inoltre, i primi sistemi per PC giravano su CPU semplici ed economiche, non dotate del supporto hardware necessario per implementare un sistema operativo **preemptive**.

5.1.3 Il Dispatcher

Quando lo scheduler ha scelto il processo a cui assegnare la CPU, interviene un altro modulo del SO, il *dispatcher*, che:

- Effettua l'operazione di *context switch*.
- Effettua il passaggio del sistema in *user mode*.
- Posiziona il *PC* della CPU alla corretta locazione del programma da far ripartire.

Si definisce **Dispatch latency** il tempo impiegato per effettuare la commutazione da un processo ad un altro.

5.2 Criteri di Scheduling

Come abbiamo visto, lo scheduler della CPU interviene per assicurare il corretto funzionamento del sistema. Tuttavia, quando lo scheduler deve mandare in esecuzione un processo, quale criterio usa per scegliere tra tutti i processi presenti nella coda di ready?

Si possono prendere in considerazione diversi obiettivi:

- Massimizzare l'**utilizzo** della CPU nell'unità di tempo, anche se questo dipende dal carico. - Massimizzare il **throughput**, ossia la produttività del sistema, che si misura come il numero di processi completati in media in una certa unità di tempo. - Minimizzare il **tempo di risposta**, cioè il tempo che intercorre da quando si avvia un processo a quando questo inizia effettivamente ad eseguire. Questo aspetto è particolarmente importante per i sistemi interattivi.

- Minimizzare il *Turnaround time*: ossia il tempo medio di completamento di un processo, che va da quando entra per la prima volta nella *ready queue* a quando termina. - Minimizzare il *Waiting time*: ossia la somma del tempo trascorso dal processo in *ready queue*, ovvero quando il processo è pronto per eseguire il suo codice ma la CPU è occupata da un altro processo.

Domanda 5.4

Che relazione c'è tra waiting time e turnaround time?

Note:-

Turnaround time - waiting time = tempo di esecuzione

5.3 Algoritmi di Scheudling

- **First Come, First Served (FCFS)**: Scheduling per ordine di arrivo.
- **Shortest Job First (SJF)**: Scheduling per brevità.
- **Priority Scheduling**: Scheduling per priorità.
- **Round Robin (RR)**: Scheduling circolare.
- **Multilevel Queue**: Scheduling a code multiple.
- **Multilevel Feedback Queue**: Scheduling a code multiple con retroazione.

Nota Bene: Nel seguito, considereremo processi con un unico *burst* di CPU, senza *burst* di I/O e con una durata espressa in generiche unità di tempo. Questo semplifica la comprensione degli algoritmi senza perdita di generalità.

5.3.1 First Come First Served (FCFS)

L'algoritmo *First Come, First Served (FCFS)* è facile da implementare: gestisce la *ready queue* (RQ) in modo FIFO (*First In, First Out*).

- Il *PCB* di un processo che entra nella *RQ* viene inserito in fondo alla coda.
- Quando la CPU si libera, viene assegnata al processo il cui *PCB* si trova in testa alla coda FIFO.

FCFS è un algoritmo non *preemptive*, per cui non è adatto per i sistemi *time-sharing*. Inoltre, con FCFS, il tempo di attesa per il completamento di un processo può risultare spesso molto lungo.

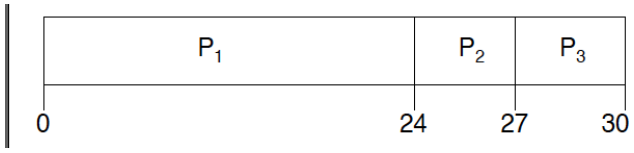
Esempio

Consideriamo tre processi che arrivano assieme al tempo t=0, e che entrano in CPU nell'ordine P1, P2, P3. Come abbiamo già detto, i tre processi eseguono per un unico burst di CPU, e poi terminano.

Process	Burst Time
P_1	24
P_2	3
P_3	3

Table 5.1: Process and Burst Time

Usiamo un diagramma di Gnatt per rappresentare questa situazione Tempi di attesa $P_1 = 0; P_2 = 24; P_3 =$



17
Tempo medio di attesa $(0 + 24 + 27)/3 = 17$
Si dice che si è verificato **effetto convoglio**; i job più corti si sono dovuti accodare a quello lungo. Se invece

supponiamo che l'ordine di arrivo sia: P2, P3, P1 Tempi di attesa $P_1 = 6; P_2 = 0; P_3 = 3$
Tempo medio di attesa $(6 + 0 + 3)/3 = 3 \longleftrightarrow$ Molto meglio del caso precedente!

Osservazioni

Dunque, l'algoritmo *FCFS* sembra comportarsi male nei confronti dei processi brevi.
Inoltre, *FCFS* è pessimo per i sistemi *time-sharing* poiché non garantisce un tempo di risposta ragionevole.
Ancora peggio, *FCFS* non è adatto ai sistemi *real-time* perché non è *preemptive*.

Dall'esempio visto, sembra che le prestazioni migliorino facendo eseguire prima i processi più corti, indipendentemente dall'ordine di arrivo nella *ready queue*. Tuttavia, questo apre la porta a nuovi problemi, che andremo a considerare.

5.3.2 Shortest Job First (SJF)

Si esamina la durata del prossimo *burst* di CPU di ciascun processo in *RQ* e si assegna la CPU al processo con il *burst* di durata minima.
Il nome esatto di questo algoritmo è *Shortest Next CPU Burst*.
Può essere usato in modalità *pre-emptive* e *non pre-emptive*.
Nel caso *preemptive*, se arriva in *ready queue* un processo il cui *burst time* è inferiore al tempo rimanente del processo attualmente in esecuzione, quest'ultimo viene interrotto e la CPU passa al nuovo processo. Questo schema è noto come *Shortest-Remaining-Time-First* (SRTF).

Esempio

Non-preemptive

Esempio:

preemptive

Esempio:

Process	Arrival Time	Burst Time
P_1	0	7
P_2	2	4
P_3	4	1
P_4	5	4

Table 5.2: Process, Arrival Time, and Burst Time

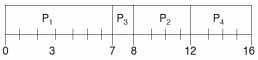


Figure 5.1: Average waiting time $(0 + 6 + 3 + 7)/4 = 4$

5.3.3 Osservazioni

Si può dimostrare che l'algoritmo *Shortest Job First* (SJF) è ottimale: spostando un processo breve prima di uno di lunga durata (anche se quest'ultimo è arrivato prima) si migliora l'attesa del processo breve più di quanto si peggiori quella del processo lungo. Di conseguenza, il tempo medio di attesa diminuisce, così come il *turnaround time*.

SJF è ottimale: nessun altro algoritmo di *scheduling* può produrre un tempo di attesa medio e un *turnaround time* medio migliori. Tuttavia, c'è un problema...
Purtroppo, la durata del prossimo burst di CPU di un processo non è nota, il che rende l'algoritmo *Shortest Job First* (SJF) non implementabile nella sua forma pura. SJF può al massimo essere approssimato utilizzando medie pesate per stimare la durata del prossimo burst di CPU di un processo, basandosi sulla durata dei burst di CPU precedenti.

Note:-
Da rivededere!!!!

Lo *scheduling* viene quindi eseguito sulla base di queste stime, fatte per tutti i processi nella *Ready Queue* in un dato momento.
In sostanza, il *First Come, First Served* (FCFS) è il peggiore degli algoritmi ragionevoli: funziona, ma spesso fornisce tempi medi di attesa e di *turnaround* pessimi.

Al contrario, lo *Shortest Job First* (SJF) è il migliore algoritmo possibile, ma non è implementabile nella pratica, e possiamo solo usarlo per fare simulazioni con processi i cui burst di CPU siano noti a priori.

FCFS e SJF rappresentano i due estremi di uno spettro di possibili algoritmi di *scheduling*. Un algoritmo di *scheduling* sarà tanto migliore quanto più le sue prestazioni si allontanano da quelle di FCFS e si avvicinano a quelle di SJF.

5.3.4 Scheduling a Priorità

SJF è un tipo scheduling a priorità, la durata del prossimo burst time è la priorità corrente di ogni processo. FCFS è uno scheduling a priorità, viene data più alta ai primi che arrivano In generale, il calcolo della priorità dei processi può essere:

- **Interna al sistema**: calcolata dal SO sulla base del comportamento di ogni processo (ad esempio, in base alle risorse usate fino a quel momento da un processo).

- **Esterna al sistema**: assegnata con criteri esterni al SO (ad esempio, una priorità che cambia in base a quale utente ha lanciato il processo).

Lo *scheduling* a priorità può essere implementato sia in modalità **preemptive** che **non preemptive**.

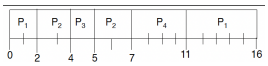


Figure 5.2: Average waiting time $(9 + 1 + 0 + 2/4 = 3$

Starvation e aging

Domanda 5.5: Problema

Che succede se un processo in RQ ha sempre una priorità peggiore di qualche altro processo in RQ?

Il processo potrebbe non essere mai scelto dallo scheduler. Questo fenomeno è noto come **starvation** (muore di fame...).

Per risolvere il problema della *starvation*, si usa un meccanismo chiamato **aging**: il SO aumenta progressivamente la priorità di un processo P_x man mano che P_x passa tempo nella Ready Queue (RQ). In questo modo, prima o poi, P_x avrà una priorità maggiore rispetto agli altri processi e verrà scelto dallo scheduler.

Domanda 5.6

Gli algoritmi *FCFS*, *SJF preemptive* e *non preemptive* possono provocare starvation?

Note:-

Per SJF, arrivano sempre processi con burst piccolissimi e quindi un processo più grande aspetterà (Sia per preemptive che non)

5.3.5 Scheduling Round Robin (RR)

Ogni processo ha a disposizione una certa quantità di tempo di CPU, chiamata **quanto di tempo** (valori ragionevoli vanno da 10 a 100 millisecondi). Per ora, assumiamo un unico quanto di tempo prefissato assegnato a tutti i processi.

Se entro questo arco di tempo il processo non lascia volontariamente la CPU, viene interrotto e rimesso nella Ready Queue (RQ). La RQ è vista come una coda circolare, e si verifica una sorta di *"girotondo"* di processi.

L'implementazione dello scheduling **round robin** è concettualmente molto semplice:

- Lo scheduler sceglie il primo processo in RQ (ad esempio secondo un criterio FCFS).
- Lancia un timer inizializzato al quanto di tempo.
- Passa la CPU al processo scelto.

Se il processo ha un CPU burst minore del quanto di tempo, il processo rilascierà la CPU volontariamente prima dello scadere del tempo assegnatogli. Se invece il CPU burst del processo è maggiore del quanto di tempo, allora:

- Il timer scade e invia un interrupt.
- Il SO riprende il controllo della CPU.
- Togliere la CPU al processo in esecuzione e metterlo in fondo alla RQ.
- Prendere il primo processo in RQ e ripetere tutto.

Osservazioni

Se ci sono n processi in coda ready e il quanto di tempo è q, allora ogni processo riceve $\frac{1}{n}$ del tempo della CPU e nessun processo aspetta per più di $(n-1)q$ unità di tempo.

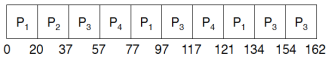
Il **Round Robin** è l'algoritmo di scheduling naturale per implementare il time sharing ed è quindi particolarmente adatto per i sistemi interattivi: nel caso peggiore, un utente non aspetta mai più di $(n-1)q$ unità di tempo prima che il suo processo venga servito.

Come vedremo negli esempi di casi reali, il SO adotta poi ulteriori misure per migliorare il tempo di risposta dei processi interattivi.

5.3.6 Esempio

Process	Burst Time
P_1	53
P_2	17
P_3	68
P_4	24

Table 5.3: Process and Burst Time



Tipicamente sia ha un *turnaround* medio maggiore di SJF, ma un migliore **tempo di risposta** Le prestazioni del **Round Robin** dipendono molto dal valore del quanto di tempo q scelto:

- q tendente a infinito rende **RR** uguale a **FCFS**.
- q tendente a zero produce un maggior effetto di "parallelismo virtuale" tra i processi.
- Tuttavia, questo aumenta il numero di context switch e, di conseguenza, l'overhead.

5.3.7 Scheduling a Code Multiple

I processi possono essere suddivisi in classi differenti:

- **foreground**: processi interattivi (es. un editor)
- **background**: processi che non interagiscono con l'utente
- **batch**: processi la cui esecuzione può essere differita

La risorsa di esecuzione (RQ) può essere partizionata in più code:

- I processi vengono inseriti in una coda basata sulle loro proprietà
- Ogni coda viene gestita con lo scheduling appropriato

Ogni coda ha quindi la sua politica di scheduling, ad esempio:

- *foreground*: **RR**
- *background e batch*: *FCFS*

Domanda 5.7

Ma come si sceglie fra le code?

- **Scheduling a priorità fissa**: servire prima tutti i processi nella coda foreground e poi quelli in background e batch. Possibilità di **starvation**.
- **Time slice**: ogni coda ha una certa quantità di tempo di CPU, ad esempio: 80% alla coda foreground e 20% alla coda background e batch

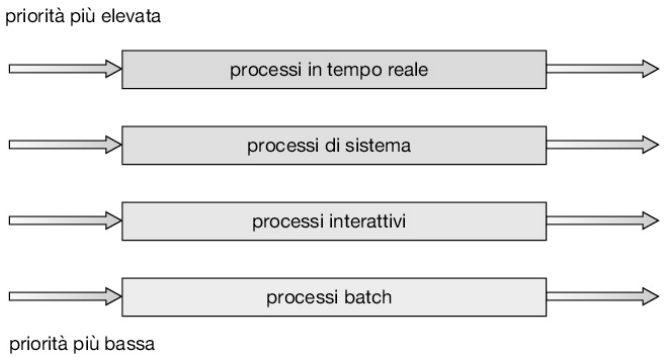


Figure 5.3: Partizionamento dei processi in più code

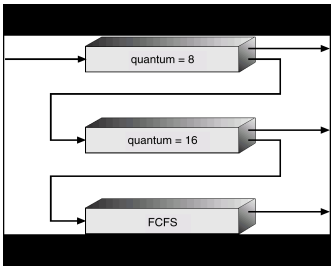


Figure 5.4: Esempio di MFQS

5.3.8 Scheduling a Code Multilivello con retroazione (MFQS)

Il tipo più generale di algoritmo di scheduling è lo **scheduling a code multilivello con retroazione** (MFQS), utilizzato dai sistemi operativi moderni.

- L'assegnamento di un processo a una coda non è fisso: i processi possono essere spostati dal SO per adattarsi alla lunghezza del loro *CPU burst*.
- Ogni coda è gestita con lo scheduling più adatto ai processi in essa contenuti.

Esempio (fig. 5.4):

- Le prime due code sono gestite con *Round Robin (RR)*, mentre la terza con *First-Come, First-Served (FCFS)*.
- Quando un processo nasce, è inserito nella prima coda ($q = 8$). Se non finisce il *CPU burst* entro il quanto, viene retrocesso alla coda successiva.
- È definita una priorità tra le code, che sono gestite con *preemption*.

La politica MFQS è caratterizzata da:

- numero di code
- algoritmo di scheduling per ogni coda
- quando declassare o promuovere un processo
- in che coda inserire un processo quando arriva (dall'esterno o da un *I/O burst*)

MFQS è il tipo di scheduling più generale e complesso da configurare.

5.4 Scheduling per sistemi multi-core

Sono ormai comuni le architetture con CPU a 2, 4, 8 core, Sono, in sostanza, dei piccoli sistemi multiprocessore in cui sullo stesso chip sono presenti due o più core che vedono la stessa memoria principale e condividono un livello di cache. La presenza di più "unità di esecuzione" dei processi, permette naturalmente di aumentare le prestazioni della macchina, posto che il SO sia in grado di sfruttare a pieno ciascun core.

I sistemi operativi moderni prevedono la **multielaborazione simmetrica** (SMP), in cui uno scheduler gira su ciascun core.

- I processi "ready to run" possono essere inseriti in una coda comune oppure in una coda separata per ogni core.
- Lo scheduler di ciascun core sceglie un processo dalla propria coda e lo manda in esecuzione.

Un aspetto chiave nei sistemi multi-core è il **bilanciamento del carico**, ossia la distribuzione omogenea dei processi tra i core.

- Con una coda comune, il bilanciamento è automatico: un core inattivo prende un processo dalla coda comune.
- Con code separate per ogni core, è necessario un meccanismo per spostare i processi dai core sovraccarichi a quelli scarichi. **Questo è il processo preferito dai moderni SO**
- Ad esempio, Linux SMP attiva il bilanciamento del carico ogni 200 ms o quando una coda si svuota.

Spostare un processo tra core può causare rallentamenti dovuti alla **cache**, poiché il processo potrebbe non trovare i dati nelle cache private del nuovo core. Non trovandoli è costretto a spendere più tempo per recuperarli, se va bene, dalla cache L3, che condivide con gli altri core. Per evitare questo problema, specifiche *system call* permettono di vincolare un processo a un certo core.

5.4.1 Esempio

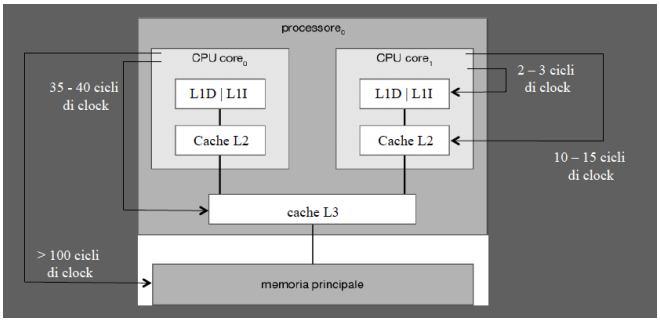


Figure 5.5: Scheduling example

Qui vediamo il costo in cicli di clock necessari per accedere ad un dato/istruzione in un certo livello di cache. I valori si riferiscono ad un processore Intel core i7, ma valgono per la maggior parte dei processori moderni 5.5.

5.5 Esempi di sistemi operativi

Solaris utilizza uno **scheduling a priorità con code multiple a retroazione**, suddividendo i processi in 4 classi:

1. real time (priorità maggiore)
2. sistema
3. interattiva

4. time sharing (priorità minore)

- Un processo usa la CPU fino a quando non termina, va in *wait*, esaurisce il quanto di tempo o è preemptato.
- I processi delle classi *time sharing* e *interattiva* hanno criteri di scheduling simili con 60 livelli di priorità.
- La priorità di un processo e il quanto di tempo assegnato sono inversamente proporzionali. Se un processo esaurisce il quanto, la sua priorità viene abbassata; al contrario, se si sospende prima, la priorità aumenta.

Il comportamento del processo stabilisce se rientra nella classe *time sharing* (priorità 0-49) o *interattiva* (priorità 50-59)

Priorità corrente	Quanto di tempo (millisecondi)	Nuova priorità (quanto esaurito)	Nuova priorità (quanto non esaurito)
0	200	0	50
20	120	10	52
30	80	25	53
59	20	49	59

Table 5.4: Tabella delle priorità e tempi

La **priorità corrente** di un processo determina il **quanto di tempo** che gli viene assegnato. Priorità e tempo assegnato sono inversamente proporzionali.

- **Quanto esaurito**: se un processo ha esaurito tutto il quanto di tempo senza sospendersi, la sua nuova priorità sarà più bassa, e in futuro riceverà un quanto di tempo più lungo.
- **Quanto non esaurito**: se un processo si sospende prima di consumare tutto il quanto, la sua nuova priorità sarà più alta, e in futuro riceverà un quanto di tempo più breve.
- I processi *real time* e di *sistema* hanno priorità fissa, superiore a quella delle classi *time sharing* e *interattiva*.
- Lo scheduler assegna la CPU al processo con la priorità globale più alta e, in caso di parità, utilizza il *Round Robin* (RR).
- L'algoritmo è **preemptive**: un processo attivo può essere interrotto da uno con priorità globale più alta.

5.5.1 Lo scheduling in Windows

Lo scheduling in Windows è basato su **priorità con retroazione e prelazione**, utilizzando 32 livelli di priorità:

- I processi *real-time* hanno priorità da 16 a 31.
- I processi non real-time hanno priorità da 1 a 15, con 0 riservato.

Per i processi non real-time:

- Quando un processo nasce, ha una priorità iniziale di 1.
- Lo scheduler assegna la CPU al processo con la priorità più alta, usando *Round Robin* (RR) in caso di parità.
- Se un processo va in *wait* prima di esaurire il quanto di tempo, la sua priorità viene aumentata (fino a 15), a seconda dell'evento in attesa (maggiore incremento per input da tastiera, minore per I/O da disco).
- Se un processo esaurisce il quanto di tempo, la sua priorità viene abbassata, ma mai sotto 1.

Questa strategia favorisce i processi interattivi (mouse e tastiera) per migliorare il **tempo di risposta**. Inoltre, quando un processo passa in *foreground*, il suo quanto di tempo viene moltiplicato per 3, consentendogli di mantenere la CPU per un periodo più lungo.

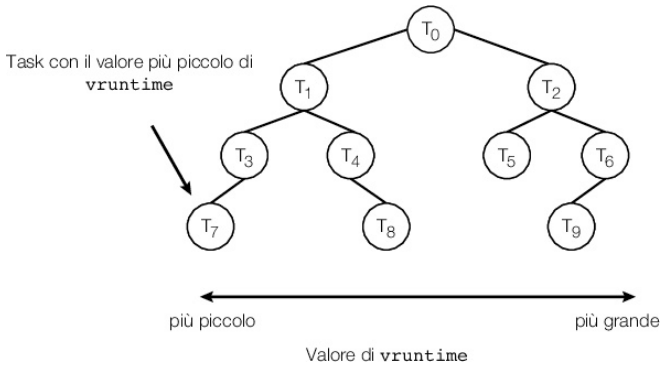


Figure 5.6: vrun time tree

5.5.2 Lo Scheduling in Linux

Dal 2007, Linux utilizza il **Completely Fair Scheduler** (CFS) come algoritmo di scheduling predefinito.

Il CFS distribuisce equamente il tempo di CPU tra i processi *ready to run*, seguendo l'assunzione che se ci sono N processi attivi, ciascun processo dovrebbe ricevere esattamente $\frac{1}{N}$ del tempo di CPU.

Ad ogni *context switch*, il CFS ricalcola per quanto tempo assegnare la CPU a un processo P, in modo che tutti abbiano la stessa quantità di tempo CPU. Siano:

- P.`expected_run_time`: il tempo di CPU spettante a P;
- P.`vruntime`: il tempo di CPU già consumato da P;
- P.`due_cputime`: il tempo di CPU che ancora spetta a P.

Dunque:

$$P.\texttt{vruntime} = P.\texttt{expected\_run\_time} - P.\texttt{due\_cputime}$$

La CPU viene assegnata al processo con il valore più basso di P.`vruntime`, ossia al processo che ha usato meno CPU fino a quel momento.

Nel CFS, i processi *ready to run* non sono organizzati in code di scheduling, ma come nodi in un **red-black tree (R-B tree)**, che consente operazioni di ricerca, inserimento e cancellazione con complessità computazionale $O(\log n)$, dove n è il numero di nodi.

Negli alberi R-B il nodo più a sinistra è sempre quello col valore chiave più basso, e nel CFS i processi sono inseriti nel R-B tree usando come chiave P.vruntime. Dunque, il processo associato al nodo più a sinistra ha il valore P.vruntime più basso, cioè è il processo che ha usato la CPU per meno tempo, e al context switch sarà scelto per entrare in esecuzione.

6 Sincronizzazione dei Processi

6.1 Introduzione

Più processi possono cooperare per compiere un determinato lavoro, e spesso **condividono dei dati**.

- È fondamentale che l'accesso ai dati condivisi da parte dei vari processi non produca dati inconsistenti.
- I processi cooperanti devono quindi **sincronizzarsi** per accedere ai dati condivisi in modo ordinato.
- **Problema**: mentre un processo P sta elaborando dati condivisi, il SO potrebbe toglierlo dalla CPU in qualsiasi momento. Altri processi non devono poter accedere ai dati condivisi finché P non ha completato l'elaborazione.

6.1.1 Esempio: Produttore-Consumatore con n eleemnti

Usiamo una variabile ***condivisa*** **counter** inizializzata a 0 che indica il numero di elementi nel buffer.

I due programmi sono corretti se considerati separatamente, ma possono non funzionare quando vengono eseguiti insieme.

- Il problema risiede nell'uso della variabile condivisa `counter`.
- Che succede se il produttore esegue `counter++` mentre *contemporaneamente* il consumatore esegue `counter-`?
- Se `counter` all'inizio vale 5, dopo `counter++` e `counter-` può valere 4, 5 o 6!
- N.B.: diciamo che possono non funzionare, e non che non funzionano, perché la condizione problematica potrebbe non verificarsi sempre.

Il problema si verifica perchè `counter++, counter-` non sono **operazioni atomiche** Le operazioni sui dati condivisi possono portare a risultati imprevisti. Consideriamo le istruzioni per il **produttore** e il **consumatore** relative alla variabile condivisa `counter`: **Produttore:**

```
load(registro1, counter);  % Carica il valore di counter in registro1
add(registro1, 1);         % Incrementa il valore nel registro di 1
store(registro1, counter);  % Salva il valore incrementato in counter
```

Consumatore:

```
load(registro1, counter);  % Carica il valore di counter in registro1
sub(registro1, 1);         % Decrementa il valore nel registro di 1
store(registro1, counter);  % Salva il valore decrementato in counter
```

Se il produttore e il consumatore accedono a `counter` in modo non sincronizzato, il valore finale di `counter` può risultare errato e instabile.

Quando i processi devono accedere e modificare dati condivisi, è fondamentale che si **sincronizzino** affinché ciascuno possa completare le proprie operazioni sui dati prima che un altro processo possa accedervi.

- Questo approccio assicura l'integrità dei dati e previene condizioni di competizione.
- Da notare che il problema non si presenta se tutti i processi coinvolti nell'accesso a un insieme di dati condivisi devono solo **leggere** quei dati.

6.2 Sezioni critiche

Siano dati n processi $P_1, \ldots, P_n$ che usano variabili condivise.

Ogni processo ha una porzione di codice, detta **sezione critica**, in cui manipola le variabili condivise (o anche solo un loro sottoinsieme).
Quando un processo P_i è dentro alla propria sezione critica, nessun altro processo P_j può eseguire il codice della propria sezione critica, poiché userebbe le stesse variabili condivise (o anche solo un loro sottoinsieme).
L'esecuzione delle sezioni critiche di $P_1, \ldots, P_n$ deve quindi essere **mutualmente esclusiva**.
Mentre un processo P_i sta eseguendo codice nella propria sezione critica, potrebbe essere tolto dalla CPU dal sistema operativo a causa del normale avvicendamento tra processi.
Fino a che P_i non ha terminato di eseguire il codice della sua sezione critica, **nessun altro processo P_j che deve manipolare le stesse variabili condivise potrà eseguire il codice della propria sezione critica**.

Osservazioni 6.2.1

È importante notare che P_j può comunque eseguire del codice, quando entra in esecuzione, ma non il codice della propria sezione critica.

Definizione 6.2.1: Sezione critica

Sezione critica: porzione di codice che deve essereeseguito senza intrecciarsi (nell'avvicendamento in CPU) col codice delle sezioni critiche di altri processi che usano le stesse variabili condivise

6.2.1 Problema della Sezione Critica

Per garantire l'accesso sicuro alle variabili condivise, è necessario stabilire un **protocollo di comportamento** per i processi.

- Un processo deve **"chiedere il permesso"** per entrare nella sezione critica, utilizzando una opportuna porzione di codice detta **entry section**.
- Un processo che esce dalla sua sezione critica deve **"segnalarlo"** agli altri processi, usando una opportuna porzione di codice detta **exit section**.

Un generico processo Pi contiene una sezine critica che avrà la seguente struttura

```
altro codice
    entry section
    sezione critica
    exit section
altro codice
```

Siano dati n processi $P_1, \ldots, P_n$ che usano delle variabili condivise. Una soluzione corretta al problema della sezione critica per $P_1, \ldots, P_n$ deve soddisfare i seguenti tre requisiti:

1. **Mutua esclusione:** Se un processo P_i è entrato nella propria sezione critica ma non ne è ancora uscito (attenzione, P_i non è necessariamente il processo in esecuzione, cioè quello che sta usando la CPU), nessun altro processo P_j può entrare nella propria sezione critica.

2. **Progresso:** Se un processo lascia la propria sezione critica, deve permettere ad un altro processo P_j di entrare nella propria (di P_j) sezione critica. Se la sezione critica è vuota e più processi vogliono entrare, uno tra questi deve essere scelto in un tempo finito (*in altre parole, esiste un processo che entrerà in sezione critica in un tempo finito)*

   - **Osservazioni 6.2.2**

     Qquesta condizione garantisce che l'insieme dei processi P1,... Pn (o anche solo un loro sottoinsieme) non finisca in una condizione di deadlock: tutti fermi in attesa di riuscire ad entrare nella loro sezione critica

3. **Attesa limitata:** se un processo P_i ha già eseguito la sua entry section (ossia ha già chiesto di entrare nella sua sezione critica), esiste un limite al numero di volte in cui altri processi possono entrare nelle loro sezioni critiche prima che tocchi a P_i *(in altre parole,* ***qualsiasi*** *processo deve riuscire ad entrare in sezione critica in un tempo finito)*

   - **Osservazioni 6.2.3**

     Quest'ultima condizione assicura che il processo Pi non subisca una forma di **starvation**: non riesce a proseguire la sua computazione perché viene sempre sopravanzato da altri processi.

Una qualsiasi soluzione corretta al problema della sezione critica deve permettere ai processi di portare avanti la loro computazione **indipendentemente** dalla velocità relativa a cui essi procedono (ossia da quanto frequentemente riescono ad usare la CPU), purchè questa sia maggiore di zero.
Notate: dire che la soluzione deve essere indipendente dalla velocità relativa a cui procedono i processi significa, più tecnicamente, che:

- la soluzione non deve dipendere dal tipo di *scheduling* della CPU adottato dal SO (ossia dall'ordine e dalla frequenza con cui i processi vengono eseguiti);
- purché, chiaramente, si usi un algoritmo di *scheduling* ragionevole, come quelli che abbiamo visto.

Il problema della sezione critica è particolarmente delicato quando sono coinvolte strutture dati del sistema operativo.

- Ad esempio, se due processi utente eseguono una `open` sullo stesso file, vi saranno due accessi concorrenti alla stessa struttura dati del SO: la tabella dei file aperti nel sistema.
- È importante che questa tabella (come tutte le strutture dati del SO) non venga lasciata in uno stato inconsistente a causa dell'accesso concorrente dei due processi.

Il progettista del SO deve decidere come vanno gestite le sezioni critiche del sistema operativo, e le due scelte possibili sono di sviluppare un *kernel* con o senza diritto di prelazione.

- In un *kernel* **con diritto di prelazione**, un processo in *kernel mode* può essere interrotto da un altro processo (ad esempio, perché è scaduto il quanto di tempo).
- In un *kernel* **senza diritto di prelazione**, un processo in *kernel mode* non può essere interrotto da un altro processo. (*Secondo voi questo potrebbe essere rischioso?*)

Un *kernel* senza diritto di prelazione è più facile da implementare: basta disattivare gli interrupt quando un processo è in *kernel mode*.

- Non c'è più bisogno di preoccuparsi dell'accesso concorrente alle sezioni critiche del *kernel*: un solo processo alla volta può accedere alle strutture dati del *kernel*, perché un solo processo alla volta può essere in *kernel mode*.

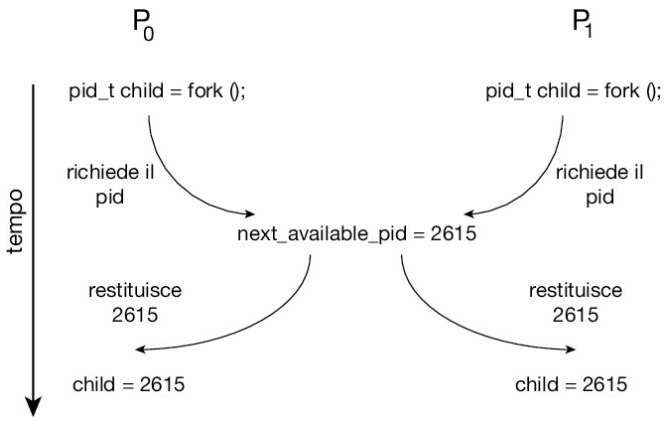


Figure 6.1: il caso di due processi che eseguono "insieme" una fork: senza opportune precauzioni, potrebbero nascere nel sistema due nuovi processi con lo stesso PID!

- Se il codice del SO è scritto correttamente, la disabilitazione degli interrupt sarà temporanea e di breve durata, e tutto continuerà a funzionare normalmente.

Del resto, i *kernel* con diritto di prelazione sono più adatti per le applicazioni *real time*, in cui la disabilitazione degli interrupt (che potrebbero essere allarmi da gestire immediatamente) non è accettabile.

- In generale, i *kernel* con diritto di prelazione hanno un tempo di risposta inferiore, per ovvie ragioni.

Note:-

- Windows 2000 e XP erano *kernel* senza diritto di prelazione, mentre i loro successori sono stati tutti progettati con diritto di prelazione.
- La maggior parte delle versioni recenti di Solaris, Unix e Linux sono *kernel* con diritto di prelazione.

Osservazioni 6.2.4

Un *kernel* senza diritto di prelazione disabilita gli interrupt per il tempo necessario al codice del SO (ad esempio di una *system call*) per accedere in modo mutuamente esclusivo a una qualche struttura dati del SO.

Domanda 6.1: Domanda d'esame

Perché una tale soluzione non è adatta per proteggere le strutture dati condivise da due processi utente, ossia per implementare le sezioni critiche dei processi utente?

Note:-

Tendenzialmente non si vuole che i processi utenti possano accedere alla modalità kernel.
Magari il codice utente non riabilita più gli interrupt? Siamo rovinati :()

6.2.2 Sincronizzazione via Hardware

Soluzioni semplici ed eleganti al problema della sezione critica (ma con un grave difetto, come vedremo) possono essere ottenute usando speciali istruzioni macchina, presenti in tutte le moderne CPU (i nomi di queste istruzioni negli *instruction set* di diversi processori possono ovviamente variare; l'importante è ciò che fanno):

- `TestAndSet(var1)`: testa e modifica il valore di una cella di memoria;
- `Swap(var1, var2)`: scambia il valore di due celle di memoria.

Importante: sono istruzioni macchina, e quindi atomiche, ovvero non possono essere interrotte a metà da un *context switch*.
Vediamo ad esempio la `TestAndSet`, che potrebbe essere implementata così:

```
boolean TestAndSet(boolean *lockvariable) {
    boolean tempvariable = *lockvariable;
    *lockvariable = true;
    return tempvariable;
}
```

Ossia:

- salva il valore di `*lockvariable` in `tempvariable`;
- setta a `true` `*lockvariable`;
- restituisce il vecchio valore di `*lockvariable`.

Ed ecco come si può **realizzare la mutua esclusione** usando la `TestAndSet`:

```
Shared data: boolean lock = false;
Processo Pi:
do {
    while (TestAndSet(&lock));
    // sezione critica (qui dentro lock = true)
    lock = false;
    // sezione non critica
} while (true);
```

Il semplice algoritmo appena visto è un esempio di soluzione al problema della sezione critica basato sull'uso di una variabile condivisa detta *lock*.

- Si dice che la sezione critica è controllata dal *lock*, e solo il processo che acquisisce il *lock* può entrare in sezione critica.

La struttura generale di queste soluzioni è quindi del tipo:

```
do {
    acquisisci il lock
    sezione critica
    restituisci il lock
    sezione non critica
} while (true);
```

Attesa limitata. NON E' GARANTITA! Infatti P1 potrebbe uscire dalla sezione critica ("lock=false") e, sempre all'interno dello stesso quanto di tempo, tornare immediatamente a tentare di acquisire il lock, riuscendoci, e la situazione può ripetersi all'infinito!

Domanda 6.2

Perché un meccanismo di aging non funzionerebbe, in questo caso?

Note:-

Non funziona perchè P2 entra in CPU, ma spreca tutto il suo quanto di tempo nel while(TestAndSet(block), quindi riperde la priorità

Ecco la soluzione corretta per n processi:

```
shared data boolean attesa[n], lock; // entrambi inizializzati a false
boolean chiave;
do {
    attesa[i] = true; // Pi annuncia di voler entrare in SC
    chiave = true;
    while (attesa[i] && chiave)
        chiave = TestAndSet(&lock);
```

```
    attesa[i] = false;

    // sezione critica

    j = (i + 1) mod n;
    while ((j != i) && !attesa[j])
        j = (j + 1) mod n;
    if (j == i)
        lock = false;
    else
        attesa[j] = false;

    // sezione non critica
} while (true);
```

La soluzione che abbiamo visto, basata sull'uso di speciali istruzioni macchina, ha un problema di fondo:

```
while (TestAndSet(&lock));
```

Il processo che attende il proprio turno per entrare in una sezione critica occupata consuma CPU inutilmente. Tecnicamente, si dice che sta facendo *busy-waiting* (a volte si usa anche l'espressione "attesa attiva").

P1 entra nella sezione critica. Incomincia a lavorare in mutua esclusione. Scade il quanto assegnatogli. La CPU viene data a P2	P2 cerca di entrare in sezione critica, occupata da P1. P2 cicla inutilmente fino alla fine del suo quanto di tempo. La CPU viene ridata a P1	P1 termina il lavoro in sezione critica e ne esce, rilasciando il lock. Al prossimo quanto di tempo P2 potrà acquisire il lock

Figure 6.2: Busy-Waiting

Osservazioni 6.2.5

E se invece di due processi ce ne sono N che competono per acquisire lo stesso lock, un algoritmo di scheduling round robin potrebbe produrre uno spreco del tempo di CPU pari a $N-1$ quanti di temp.

6.3 Semafori

Definizione 6.3.1: Semaforo

Strumento di sincronizzazione che, con l'aiuto del sistema operativo, può essere implementato senza *busy-waiting*.

Definizione di Semaforo S:

- È una variabile intera (per ora assumiamo che sia inizializzata a 1) su cui si può operare solo tramite due operazioni atomiche (che descriviamo così, anche se non implementate in questo modo):
  - `wait(S)`:

    ```
    while (S <= 0) do no-op;
    S = S - 1;
    ```

  - `signal(S)`:

    ```
    S = S + 1;
    ```

Un semaforo può essere visto come un **oggetto** condiviso da tutti i processi che devono usarlo per sincronizzarsi fra loro.

- La variabile intera S viene di solito chiamata *variabile semaforica*, e il suo valore corrente è detto *valore del semaforo*.

- `Wait` e `Signal` sono i metodi con cui si utilizza l'oggetto semaforo (in realtà, è necessaria anche un'operazione per inizializzare il valore della variabile semaforica).

In letteratura, i termini `Wait` e `Signal` sono talvolta sostituiti da termini olandesi, usati originariamente da Dijkstra:

- `P` (*Proberen* = verificare) al posto di `Wait`;
- `V` (*Verhogen* = incrementare) al posto di `Signal`.

Altri termini usati:

- `Down` (per la `Wait`, che decrementa il semaforo);
- `Up` (per la `Signal`, che incrementa il semaforo).

6.3.1 Uso dei semafori

Adesso la soluzione del problema della sezione critica per un gruppo di processi è semplice. La variabile semaforica `mutex` (mutual exclusion) fa da variabile di *lock*:

```
shared variable semaphore mutex = 1;

Generico processo Pi:
do {
    wait(mutex);
    // sezione critica
    signal(mutex);
    // sezione non critica
} while (true);
```

In realtà, i semafori possono essere usati per **qualsiasi** problema di **sincronizzazione** (ne vedremo più avanti alcuni relativamente complessi). Ad esempio, se vogliamo eseguire una generica operazione S_1 fatta dal processo P_1 prima di S_2, fatta dal processo P_2, possiamo usare un semaforo `sync` inizializzato a 0:

- P_1 esegue:

```
S1;
signal(sync);
```

- P_2 esegue:

```
wait(sync);
S2;
```

Osservazioni 6.3.1

Attenzione: il nome scelto per un semaforo non ha nessuna relazione con l'uso che ne verrà fatto.

- Un semaforo è una **variabile** (in realtà, come vedremo tra poco, è una struttura dati) che ovviamente possiamo chiamare come preferiamo.
- Naturalmente, è meglio usare nomi che ricordino **l'uso che faremo** di un semaforo. Quindi, chiameremo `mutex` un semaforo usato per implementare una mutua esclusione, e `sync` un semaforo usato per implementare un meccanismo di sincronizzazione tra due processi.
- Ma non cambierebbe nulla se chiamassimo i due semafori rispettivamente `X` e `Y`, o anche `Pippo` e `Pluto`.
- Ciò che importa è **come li usiamo**.

6.3.2 Implementazione dei semafori

La definizione di `wait` e `signal` che abbiamo dato utilizza il *busy-waiting*, e questo è proprio ciò che vorremmo evitare.

- I semafori implementati attraverso il *busy-waiting* esistono e prendono di solito il nome di *spinlock* (nel senso che il processo "gira" mentre testa la variabile di lock, proprio come nelle tecniche di sincronizzazione via hardware).
- Per evitare il *busy-waiting* dobbiamo farci aiutare dal Sistema Operativo, che mette a disposizione opportune strutture dati e *system call* per l'implementazione delle operazioni di `wait` e `signal`.

Quando un gruppo di processi ha bisogno di un semaforo, lo richiede al Sistema Operativo tramite una *system call*.

- Il SO alloca un nuovo semaforo all'interno di una lista di semafori memorizzata nelle aree dati del kernel.
- Ogni semaforo è implementato usando due campi: `valore` e `lista di attesa`.

```
typedef struct {
    int valore;
    struct processo *waiting_list;
} semaforo;
```

Due *system call* sono disponibili per **implementare** le operazioni di `wait` e `signal`:

- `sleep()`: toglie la CPU al processo che la invoca e manda in esecuzione uno dei processi nella Ready Queue. Il processo che ha chiamato `sleep` non viene rimesso nella Ready Queue (N.B.: a volte `sleep()` è chiamata `block()`).
- `wakeup(P)`: inserisce il processo P nella Ready Queue.

Implementazione della `wait`:

```
wait(semaforo *S) {
    S->valore--;
    if (S->valore < 0) {
        aggiungi questo processo a S->waiting_list;
        sleep();
    }
}
```

La chiamata di `sleep()` provoca un *context switch*, e il processo sospeso non consuma CPU inutilmente, poiché il suo PCB non è più nella Ready Queue, ma nella lista di attesa del semaforo su cui si è sospeso. Si dice anche che il processo si è "**addormentato**" sul semaforo S. `signal(semaforo *S)`:

```
signal(semaforo *S) {
    S->valore++;
    if (S->valore <= 0) { /* c e ' qualcuno in attesa */
        togli un processo P da S->waiting_list;
        wakeup(P);
    }
}
```

Nota: `wakeup(P)` rimette P nella Ready Queue, quindi P è pronto a usare la CPU quando sarà il suo turno. Si dice che P è stato "svegliato".

NOTATE BENE: `wait` e `signal` sono di solito *system call* direttamente messe a disposizione dal Sistema Operativo, anche se a volte con nomi diversi.

- `wait` e `signal` sono esse stesse sezioni critiche. Perché? **(Perchè condividono delle variabili)**
- Sono sezioni critiche molto corte (circa 10 istruzioni macchina), quindi vanno bene implementate con *spinlock* o disabilitazione degli interrupt (che avviene sotto il controllo del SO).

Domanda 6.3

Quale soluzione è migliore per i sistemi monoprocessore e quale per i sistemi multiprocessore?

NOTATE ANCHE: All'inizio, la semantica di `wait` era:

```
wait (S):
    while (S <= 0) do no-op;
    S = S - 1;
```

Domanda 6.4

Ma nell'implementazione tramite `sleep`, vediamo che il valore del semaforo può essere negativo. Come mai?

Note:-

Per conoscere quanti processi in un certo istante sono addormentanti

- Se $S->valore < 0$, il suo valore assoluto ci dice quanti processi sono in attesa (*in wait*) su quel semaforo (si veda il codice della `wait`).

NOTATE ANCORA: Il valore del semaforo può anche essere un intero maggiore di 1. Ad esempio, se una risorsa può essere usata contemporaneamente da un massimo di tre processi:

```
semaphore counter = 3; // counter viene inizializzato a 3
Generico processo Pi:
repeat {
    wait(counter);
    // usa la risorsa
    signal(counter);
    // remainder section
} until false;
```

6.3.3 Riassunto

Riassumendo, attraverso i semafori implementati usando `sleep()` e `wakeup(P)`, i processi utente possono contenere sezioni critiche arbitrariamente lunghe senza:

- Sprecare inutilmente tempo di CPU (come accadrebbe se implementassimo le sezioni critiche con il *busy-waiting*),
- Rischiare di dare il controllo della CPU al processo (come accadrebbe se si usasse la disabilitazione degli interrupt gestita direttamente dai processi utente).

Osservazioni 6.3.2

Le operazioni di `wait` e `signal` (che sono esse stesse sezioni critiche) possono invece essere implementate con *busy-waiting* o disabilitazione degli interrupt, poiché queste operazioni durano poco tempo e avvengono sotto il controllo del Sistema Operativo.

6.4 Definizione di DeadLock

Definizione 6.4.1

Si definisce **deadlock** di un sottoinsieme di processi del sistema $\{P_1, P_2, \ldots, P_n\} \subseteq P$ la situazione in cui ciascuno degli n processi P_i è in attesa del rilascio di una risorsa detenuta da uno degli altri processi del sottoinsieme;

si forma cioè una catena circolare per cui:

$$P_1 \text{ aspetta } P_2 \ldots \text{ aspetta } P_n \text{ aspetta } P_1$$

Anche se non tutti i processi del sistema sono bloccati, la situazione non è desiderabile in quanto può bloccare alcune risorse e, di conseguenza, danneggiare anche i processi non coinvolti nel deadlock.

6.5 Definizione di Starvation

Definizione 6.5.1

Un processo è in **starvation** se non riesce mai a portare avanti la propria computazione.

Questo può accadere per diverse ragioni:

- Non viene mai selezionato dallo *scheduler* per entrare in esecuzione.
- Non riesce mai ad entrare in una sezione critica.
- Non riesce mai a prelevare una risorsa necessaria per proseguire la sua computazione.

Nota: Il *deadlock* implica *starvation*, ma non vale il contrario.

6.6 Deadlock & Starvation: (stallo e attesa indefinita)

I **semafori** sono le primitive di sincronizzazione più semplici e più usate nei moderni Sistemi Operativi, e permettono di risolvere qualsiasi problema di sincronizzazione fra processi.

- Tuttavia, sono primitive di sincronizzazione *non strutturate* e quindi possono essere "rischiosi".
- Usando i semafori, non è difficile scrivere programmi che funzionano male, portando a situazioni di *deadlock* o *starvation*.

Esempio:

```
P_0           P_1
wait(S);      wait(Q);
wait(Q);      wait(S);
...           ...
signal(S);    signal(Q);
signal(Q);    signal(S);
```

Domanda 6.5

Se S e Q sono inizializzati a 1, cosa succede se P_0 e P_1 vengono eseguiti concorrentemente?

- Se dopo la `wait(S)` di P_0, la CPU viene assegnata a P_1, che esegue `wait(Q)`, e successivamente la CPU torna a P_0, P_0 non può più proseguire, causando così un **deadlock**.

Problema dei semafori: le operazioni `wait` e `signal` sono indipendenti e possono essere usate in modo errato. Esistono primitive di sincronizzazione più strutturate (es. *Regioni Critiche Condizionali*, *Monitor*) che possono evitare questi problemi.

Approfondimento: potete leggere la sezione 6.7 del testo per una descrizione del concetto di *Monitor*.

7 Esempi di sincronizzazione

7.1 Produttori-Consumatori con memoria limitata

Utilizziamo un buffer circolare di `SIZE` posizioni in cui i produttori inseriscono i dati e i consumatori li prelevano.

Dati Condivisi e Inizializzazione dei Semafori

```
typedef struct {...} item;
item buffer[SIZE];
semaphore full, empty, mutex;
item nextp, nextc;
int in = 0, out = 0;
full = 0;
empty = SIZE;
mutex = 1;
```

- `full`: conta il numero di posizioni piene del buffer.
- `empty`: conta il numero di posizioni vuote del buffer.
- `mutex`: semaforo binario per garantire l'accesso in mutua esclusione al buffer e alle variabili `in` e `out`.
- `in` e `out`: servono per gestire l'indice del buffer circolare.

7.1.1 Codice del Produttore

Il codice per il produttore è il seguente:

```
while (true) {
    // produce un item in nextp
    wait(empty);
    wait(mutex);
    buffer[in] = nextp; // inserisce nextp nel buffer
    in = (in + 1) % SIZE; // aggiorna l'indice in
    signal(mutex);
    signal(full);
}
```

7.1.2 Codice del Consumatore

Il codice per il consumatore è il seguente:

```
while (true) {
    wait(full);
    wait(mutex);
    nextc = buffer[out]; // preleva un item dal buffer
    out = (out + 1) % SIZE; // aggiorna l'indice out
    signal(mutex);
    signal(empty);
    // consuma l'item in nextc
}
```

7.1.3 Spiegazione

- Usando il semaforo `mutex`, garantiamo che solo un processo per volta acceda in mutua esclusione al buffer e alle variabili condivise `in` e `out`.
- Il semaforo `empty` assicura che i produttori possano inserire dati solo se ci sono posizioni vuote nel buffer.
- Il semaforo `full` garantisce che i consumatori possano prelevare dati solo se nel buffer sono presenti item da consumare.

Note:-

Implementare Produttori-Consumatori esempio slide :D

Domanda 7.1

Come può essere semplificato il codice se possiamo supporre che esista un solo produttore?
Come può essere semplificato il codice se possiamo supporre che esista un solo consumatore?

7.2 Problema dei Lettori-Scrittori

Vogliamo gestire l'accesso concorrente a un file condiviso tra più processi che possono essere lettori o scrittori:

- I lettori richiedono solo l'accesso in lettura e possono accedere al file contemporaneamente ad altri lettori.
- Gli scrittori richiedono l'accesso in scrittura e devono avere accesso esclusivo al file, senza che altri lettori o scrittori possano accedervi contemporaneamente.

Strutture Dati Condivise

Le seguenti strutture dati vengono utilizzate per la sincronizzazione tra lettori e scrittori:

```
semaphore mutex = 1, scrivi = 1;
int numlettori = 0;
```

- `mutex`: semaforo per garantire la mutua esclusione quando si aggiorna la variabile `numlettori`.
- `scrivi`: semaforo che garantisce l'accesso esclusivo al file per gli scrittori.
- `numlettori`: contatore che tiene traccia del numero di lettori attivi.

7.2.1 Codice del Processo Scrittore

Il codice per uno scrittore è il seguente:

```
wait(scrivi);
// esegui la scrittura del file
signal(scrivi);
```

7.2.2 Codice del Processo Lettore

Il codice per un lettore è il seguente:

```
wait(mutex); // mutua esclusione per aggiornare numlettori
numlettori++;
if (numlettori == 1) wait(scrivi); // il primo lettore blocca eventuali scrittori
signal(mutex);

// leggi il file

wait(mutex);
numlettori--;
if (numlettori == 0) signal(scrivi); // l'ultimo lettore sblocca eventuali scrittori
signal(mutex);
```

7.2.3 Spiegazione

- Lettori quando un lettore vuole accedere al file, incrementa `numlettori` sotto mutua esclusione grazie a `mutex`. Se è il primo lettore, blocca l'accesso agli scrittori tramite il semaforo `scrivi`. Quando un lettore termina di leggere, decrementa `numlettori` e, se è l'ultimo lettore, rilascia `scrivi` per permettere agli scrittori di accedere.
- Scrittori quando uno scrittore vuole accedere al file, esegue una `wait(scrivi)` per ottenere l'accesso esclusivo. Dopo aver completato la scrittura, rilascia il semaforo `scrivi` con `signal(scrivi)`.

Domanda 7.2

La soluzione garantisce assenza di deadlock e starvation per lettori e scrittori?
Riuscite a pensare a soluzioni alternative, a partire da quella vista?

Note:-

Questa soluzione è *reader-first*, quindi se arrivano sempre lettori, gli scrittori possono andare in starvation. Esistono anche altre soluzione che possono essere writer-first

7.3 Problema di cinque filosofi

7.3.1 Dati Condivisi

```
semaphore bacchetta[5]; // tutte inizializzate a 1
```

7.3.2 Codice del Filosofo i (Soluzione Errata)

```
do {
    wait(bacchetta[i]);
    wait(bacchetta[(i+1) mod 5]);
    // mangia
    signal(bacchetta[i]);
    signal(bacchetta[(i+1) mod 5]);
    // pensa
} while (true);
```

7.3.3 Problema di Deadlock

Questa soluzione può portare a una situazione di deadlock, in cui tutti i filosofi tengono una bacchetta e aspettano l'altra, bloccandosi a vicenda.

7.3.4 Soluzioni Migliori

Alcune soluzioni possibili per evitare il deadlock includono:

- Consentire a soli 4 filosofi di sedersi a tavola contemporaneamente.
- Prendere le due bacchette solo se entrambe sono disponibili, usando una sezione critica.
- Prelievo asimmetrico delle bacchette, in cui i filosofi prendono le bacchette in un ordine diverso dai loro vicini.

Note:-

Sezione 6.7 Monitori, Capitolo 8 (Approfondire) + Esercizi

Note:-

es. e) Spreca il quanto di tempo; d) Un processo kernel mode, può essere sostituito (scadenza quanto di tempo, scelta della CPU).0

Note:-

Quarto criterio fondamentale della sezione critica: è quello di evitare il busy waiting

8 Stallo dei processi (deadlock)

Note:-
Questo capitolo è **facoltativo**, presente per dare più integrità agli appunti totali

8.1 Definizione

Definizione 8.1.1

Situazione in cui ciascun processo in un insieme di n processi ($n \geqslant 2$) si trova in uno stato di *attesa* per il verificarsi di un evento che solo uno degli aaltri processi dell'insieme può provocare

Il risultato è, chiaramente, una attesa infinita da parte di tutti gli n processi!

8.2 Situazioni simili anche nella realtà

: *When two trains approach each other at a crossing, both shall come to a full stop and neither shall start up again until the other has gone*
I SO di oggi non affrontano il problema, ma questo spetta a gli utenti. In futuro chi lo sa potrà diventare un compito dei SO.

8.3 Problema dei nastri

8.3.1 Dati Condivisi

```
semaphore avail = 2; // il semaforo controlla la disponibilità dei nastri
```

8.3.2 Codice dei Processi P1 e P2

```
// Processo P1
begin
    // codice preliminare
    wait(avail); // P1 prende il primo nastro
    // altre operazioni
    wait(avail); // P1 tenta di prendere il secondo nastro
    // utilizza i nastri
    signal(avail); // P1 rilascia il primo nastro
```

```
    signal(avail); // P1 rilascia il secondo nastro
    // codice finale
end

// Processo P2
begin
    // codice preliminare
    wait(avail); // P2 prende il primo nastro
    // altre operazioni
    wait(avail); // P2 tenta di prendere il secondo nastro
    // utilizza i nastri
    signal(avail); // P2 rilascia il primo nastro
    signal(avail); // P2 rilascia il secondo nastro
    // codice finale
end
```

8.3.3 Problema di Deadlock

Questo scenario può portare a un deadlock: se entrambi i processi eseguono il primo `wait(avail)` e occupano ciascuno un nastro, nessuno dei due sarà in grado di eseguire il secondo `wait(avail)` perché il semaforo `avail` è inizializzato a 2. Di conseguenza, entrambi i processi rimarranno bloccati.

8.4 Un ponte ad una sola corsia

Ciascuna posizione di marcia può essere vista come una **risorsa**, una situazione di deadlock può essere risolta se un auto **torna indietro** $\Longrightarrow$ (libera una risorsa già occupata), si verifica **starvation** se ciascuna auto sul ponte attende che l'altra liberi l'unica corsia di marcia

8.5 Modello del sistema

Un sistema (HW + SO) può essere visto come formato da:

- un insieme finito di tipi di risorse R (cicli di CPU, spazio di memoria, device di I/O),
- Ogni tipo di risorsa è formata da un certo numero di istanze indistinguibili fra loro (ad esempio la RAM può essere divisa in porzioni identiche, ciascuna delle quali può ospitare un processo),
- Un insieme di processi P che hanno bisogno di una o più istanze di alcune delle risorse per portare a termine la computazione.

Definizione di Deadlock

Si definisce **deadlock** di un sottoinsieme di processi del sistema $\{P_1, P_2, \ldots, P_n\} \subseteq P$ la situazione in cui ciascuno degli n processi P_i è in attesa del rilascio di una risorsa detenuta da uno degli altri processi del sottoinsieme; si forma cioè una catena circolare per cui:

$$P_1 \text{ aspetta } P_2 \ldots \text{ aspetta } P_n \text{ aspetta } P_1$$

Anche se non tutti i processi del sistema sono bloccati, la situazione non è desiderabile in quanto può bloccare alcune risorse e danneggiare anche i processi non coinvolti nel deadlock.

8.6 Caratterizzazione dei Deadklock

Il SO può avvalersi di una opportuna rappresentazione detta **grafo** di assegnazione delle risorse che in ogni istante registra quali risorse sono assegnate a quale processo, e quali risorse sta **aspettando** ciascun processo.

8.7 Metodi per prevenire dei Deadlock

1. Prevenire o evitare i deadlock, usando un opportuno protocollo di richiesta e assegnamento delle risorse
2. lasciare che il deadlock si verifichi, ma fornire strumenti per la scoperta e il recupero dello stesso, esplorando il grafo di assegnazione delle risorse alla ricerca di cicli.
   -
     Osservazioni 8.7.1

     Tuttavia, la soluzione 1 genera un eccessivo sottoutilizzo delle risorse, mentre la soluzione 2 non evita il problema e richiede lavoro al SO per eliminare il deadlock, dunque i SO moderni adottano la soluzione 3:
3. **Lasciare agli utenti la prevenzione/gestione dei deadlock**

9 Memoria centrale

9.1 Introduzione

Abbiamo visto che i moderni SO tentano di massimizzare l'uso delle risorse della macchina, e in primo luogo l'utilizzo della CPU.

- Questo si ottiene mediante le due tecniche fondamentali del multi-tasking e del time-sharing, che richiedono di tenere in memoria primaria contemporaneamente più processi attivi.
- Il SO deve decidere come allocare lo spazio di RAM tra i processi attivi, in modo che ciascun processo sia pronto per sfruttare la CPU quando gli viene assegnata.

Supponiamo però che, ad un certo punto, la RAM sia **completamente** occupata da 3 processi utente, P1, P2, P3 (per semplicità assumiamo che a tutti i processi venga assegnata una porzione di RAM della stessa dimensione).

Domanda 9.1

Un nuovo processo P4 viene fatto partire, è immediatamente pronto per usare la CPU, ma non c'è più spazio per caricare il suo codice in RAM, che si può fare?

- Ovviamente si potrebbe aspettare la terminazione di uno dei 3 processi già in RAM, ma supponiamo che uno dei tre processi (diciamo P2) sia temporaneamente in attesa di compiere una lunga operazione di I/O (per cui non userà la CPU a breve).

Il SO potrebbe decidere di spostare temporaneamente P2 sull'hard disk per far posto a P4, che così può concorrere all'uso della CPU.

Definizione 9.1.1: Swapping

- Che cosa viene spostato sull'hard disk? L'immagine di P2: il codice (anche se, come capiremo meglio più avanti, questo si può anche evitare), i dati e lo stack del processo.
- Dopo un po' P1 termina e libera una porzione di RAM. Il SO potrebbe riportare P2 in RAM (ma ora nello spazio che era stato inizialmente assegnato a P1).

Questa tecnica viene chiamata *swapping* (avvicendamento di processi). L'area del disco in cui il SO copia temporaneamente un processo viene detta area di *swap*.

Domanda 9.2

Lo *swapping* è raramente usato nei moderni sistemi operativi perché troppo inefficiente, ma l'esempio mette in luce un problema fondamentale nella gestione della memoria primaria: P2 contiene istruzioni che usano indirizzi di memoria primaria: funziona ancora correttamente quando viene spostato da un'area di RAM ad un'altra?

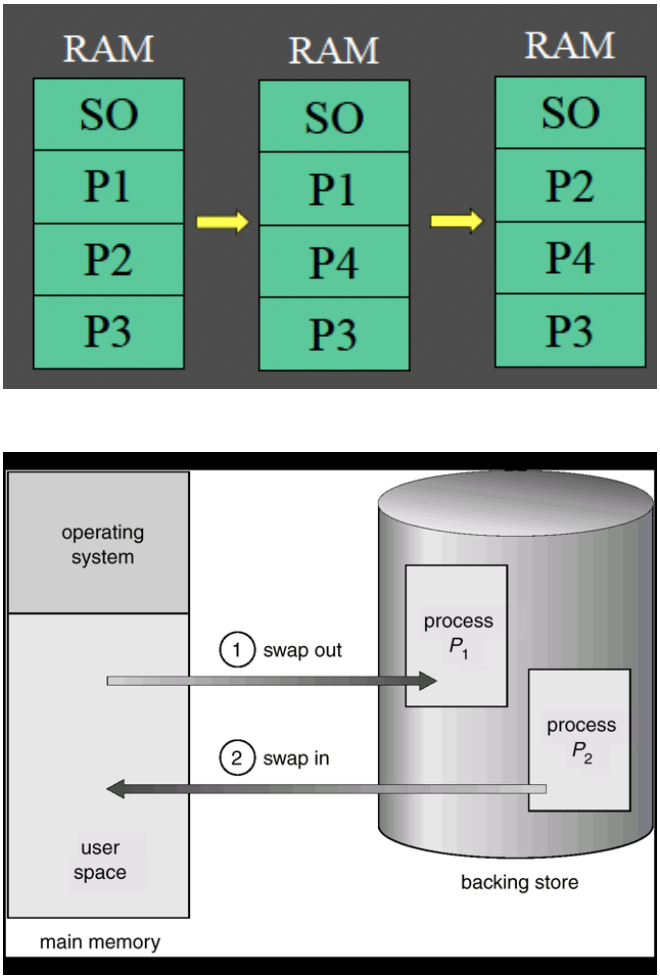


Figure 9.1: Swap mem

Perché un programma possa essere **eseguito**, il suo codice deve trovarsi in **memoria primaria** (ma rivedremo questa affermazione quando parleremo della memoria virtuale) Quindi, quando il SO riceve il **comando** di esecuzione di un programma, deve recuperare il codice del programma dalla memoria secondaria, e decidere in quale porzione della memoria primaria sistemarlo. ossia, a partire da quale indirizzo di RAM.

9.2 Binding (associazione degli indirizzi)

Un programma sorgente usa (tra l'altro) dati (variabili) e istruzioni di controllo del flusso di computazione.

- Quando il programma viene compilato e caricato in Memoria Primaria (MP) per essere eseguito, ad ogni variabile è associato l'**indirizzo** di una locazione di memoria che ne contiene il valore.
- Alle istruzioni di controllo del flusso di esecuzione del programma (ossia i salti condizionati e incondizionati) è associato l'indirizzo di destinazione del salto.
- L'operazione di associazione di variabili e istruzioni agli indirizzi di memoria è detta *binding degli indirizzi*.

In altre parole, ad ogni variabile dichiarata nel programma viene fatto corrispondere l'indirizzo di una cella di memoria di RAM in cui verrà memorizzato il valore di quella variabile.

- L'accesso alla variabile, in lettura e scrittura, corrisponde alla lettura e scrittura della cella di memoria il cui indirizzo è stato "legato" (con l'operazione di binding) alla variabile.
- Le istruzioni di salto, che permettono di implementare costrutti come *if-then-else*, *while*, ecc., sono associate agli indirizzi in RAM dove si trova l'istruzione con cui prosegue l'esecuzione del programma se il salto viene eseguito.

Ad esempio, un'istruzione C come:

```
counter = counter + 1;
```

alla fine diventerà qualcosa del tipo:

```
load(R1, 10456)
Add(R1, #1);
store(R1, 10456)
```

10456 è l'indirizzo della cella di memoria che contiene il valore della variabile *counter*. L'indirizzo 10456 è stato associato alla variabile *counter* durante la fase di binding degli indirizzi.

Analogamente, un'istruzione C come:

```
while (counter <= 100) counter++;
```

alla fine diventerà qualcosa del tipo:

```
100FC jgt(R1, #100, 10110) // jump if greater than
10100 load(R1, 10456)
10104 Add(R1, #1)
10108 store(R1, 10456)
1010C jmp(100FC)
10110 ... ...
```

Rispetto all'indirizzo di istruzione del salto stesso, il *while* della slide precedente potrebbe anche essere tradotto in assembler così:

```
100FC jgt(R1, #100, 00014) // jump if greater than
10100 load(R1, 10456)
10104 Add(R1, #1)
10108 store(R1, 10456)
1010C jmp(100FC)
10110 ... ...
```

Perché un programma sorgente possa essere eseguito deve passare attraverso varie fasi. Il binding degli indirizzi avviene in una di queste fasi:

- compilazione
- caricamento (in RAM)
- esecuzione

9.2.1 Quando?

1. In fase di Compilazione
   - viene generato codice assoluto o statico.
   - Il compilatore deve conoscere l'indirizzo della cella di RAM a partire dal quale verrà caricato il programma, in modo da effettuare il *binding* degli indirizzi (che avviene, appunto, in fase di compilazione).

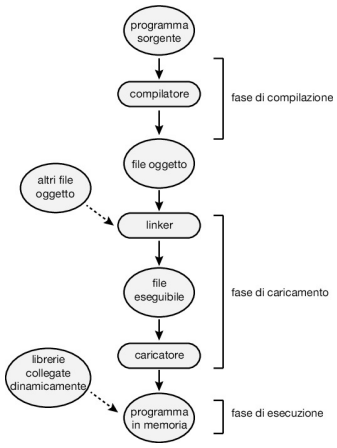


Figure 9.2: Processo di compilazione da programma sorgente

- Se il SO deve scaricare temporaneamente il processo che usa quel codice in Memoria Secondaria (MS), come nell'esempio visto a inizio capitolo, quando lo ricarica in RAM deve rimetterlo esattamente dove si trovava prima. (Oppure?)

2. In fase di caricamento in RAM

   - Viene generato codice staticamente rilocabile.
   - Il compilatore associa ad istruzioni e variabili degli indirizzi relativi rispetto all'inizio del programma, che inizia da un ipotetico indirizzo 0 virtuale.
   - Gli indirizzi assoluti finali vengono generati in fase di caricamento del codice in Memoria Primaria (MP) in base all'indirizzo di MP a partire dal quale è caricato il codice.
   - Il binding degli indirizzi, quindi, avviene in fase di caricamento del programma in RAM: se il processo che usa quel codice viene tolto dalla RAM, si può caricarlo in una posizione diversa solo rieffettuando la fase di caricamento (ma è più efficiente che ricompilare tutto).

3. In fase di esecuzione

   - Viene generato codice dinamicamente rilocabile.
   - Il codice in esecuzione usa sempre e solo indirizzi relativi.
   - La trasformazione di un indirizzo relativo in uno assoluto viene fatta nell'istante in cui viene eseguita l'istruzione che usa quell'indirizzo.
   - È necessario un opportuno supporto hardware per realizzare questo metodo senza perdita di efficienza.
   - Si parla di *binding dinamico* degli indirizzi.
   - In opportuno registro di rilocazione viene usato per trasformare un indirizzo relativo nel corrispondente indirizzo assoluto durante l'esecuzione delle istruzioni.
   - Il registro di rilocazione contiene l'indirizzo di partenza dell'area di RAM in cui è caricato il programma in esecuzione.
   - La memory management Unit (MMU) si occuperà di trasformare gli indirizzi relativi in assoluti, usando il registro di rilocazione, per accedere alle celle di RAM indirizzate dalle istruzioni
   - Lo spostamento del processo da un area all'altra della MP è realizzabile senza problema.
   - Il SO deve solo ricordarsi dell'indirizzo della locazione di MP a partire dalla quale è memorizzato il processo

Note:-

Per spostare i programmi da un'area di RAM ad un'altra ora basta cambiare l'indirizzo scritto nel registro di rilocazione (fig. 9.5 modificata)

9.3 Spazio degli indirizzi (Logici e Fisici)

Consideriamo codice dinamicamente rilocabile (d'ora in poi faremo sempre riferimento a codice dinamicamente rilocabile, se non indicato diversamente). Ogni indirizzo usato nel codice è riferito ad un ipotetico indirizzo 0 (zero): l'indirizzo della prima istruzione di cui è formato il codice.

Gli indirizzi utilizzati in un programma possono essere:

- l'indirizzo di una cella di memoria che contiene una variabile
- l'indirizzo di un'istruzione di salto

Questi indirizzi rientrano nello **spazio di indirizzamento logico o virtuale**, che va da 0 all'ultima cella di memoria occupata. Quando il codice viene caricato in RAM, gli **indirizzi logici** generati dalla CPU vengono trasformati in **indirizzi fisici** attraverso il registro di rilocazione, permettendo di indirizzare correttamente la memoria fisica (RAM).

Lo **spazio di indirizzamento fisico** è l'insieme degli indirizzi fisici che dipende dall'area di memoria dove il sistema operativo ha caricato il programma.

Per i programmi con codice rilocabile dinamicamente esistono due tipi di indirizzi:

- Indirizzi logici, che vanno da 0 a max
- Indirizzi fisici, che vanno da $r + 0$ a $r + max$, dove r è l'indirizzo iniziale della RAM in cui il programma è caricato

Gli indirizzi logici vengono sempre mappati in indirizzi fisici per accedere correttamente alla RAM.

Le espressioni **spazio di indirizzamento logico** e **spazio di indirizzamento fisico** si riferiscono principalmente all'architettura di un sistema, non a singoli programmi.

Consideriamo un computer con un massimo di 64 Kbyte di RAM, ovvero 65536 byte. In questo contesto, possiamo dire che:

1. Il computer può indirizzare 2^{16} byte di RAM.
2. Gli indirizzi dei byte della RAM vanno da 0000 a FFFF in esadecimale (da 0 a $2^{16} - 1$).
3. L'indirizzo di ciascun byte della RAM è rappresentato da 16 bit.

Pertanto, lo **spazio di indirizzamento fisico** di questo computer è scritto su 16 bit e va da 0000 a FFFF, con una dimensione di 64 Kbyte.

Se un compilatore genera codice dinamicamente rilocabile e utilizza 12 bit (**perchè 12?**) per scrivere un indirizzo logico, lo **spazio di indirizzamento logico** di un programma sarà di massimo 2^{12} byte, ovvero 4 Kbyte. Nessun programma potrà superare questo limite, anche se può usare uno spazio logico inferiore.

Quindi, possiamo dire che lo spazio di indirizzamento logico dei programmi di questo computer è scritto su 12 bit, va da 0000 a 0FFF (esadecimale) ed è di 4 Kbyte.

In seguito, quando parleremo di spazi di indirizzamento, ci riferiremo a quelli dell'intera macchina, e non a quelli dei singoli programmi. Tuttavia, è possibile considerare un programma che occupa tutto lo spazio di indirizzamento logico della macchina.

Domanda 9.3

Ha senso che la dimensione dello spazio di indirizzamento fisico sia diversa da quella dello spazio di indirizzamento logico in un sistema reale?

In effetti, è comune che lo **spazio di indirizzamento fisico** e lo **spazio di indirizzamento virtuale** siano diversi. Nei processori moderni a 64 bit, lo spazio di indirizzamento fisico può variare da 2^{40} a 2^{64} byte (da 40 a 64 bit per gli indirizzi fisici). Tuttavia, non si usano sempre 64 bit per gli indirizzi fisici perché sono eccessivi, e un computer raramente ha una quantità di RAM pari al massimo indirizzabile dal processore (ad esempio, 2^{40} byte = 1 Terabyte = 1000 Gigabyte).

Sistemi operativi e applicazioni adottano spazi di indirizzamento virtuali che variano tipicamente da 2^{48} a 2^{64} byte, ovvero da 48 a 64 bit per gli indirizzi virtuali.

In generale, per i computer moderni vale la relazione:

$$|\text{RAM}| \neq |\text{spazio di indirizzamento fisico}| \neq |\text{spazio di indirizzamento virtuale}|$$

E di solito:

$$|\text{RAM}| < |\text{spazio di indirizzamento fisico}| < |\text{spazio di indirizzamento virtuale}|$$

In molti casi, per vincoli architetturali e dimensionali, la quantità effettiva di RAM di un computer è molto inferiore allo spazio di indirizzamento fisico. Pertanto, è spesso vero che:

$$|\text{RAM}|_{\text{effettiva}} \leqslant |\text{RAM}|_{\text{massima}} \ll |\text{spazio fisico}| < |\text{spazio virtuale}|$$

Osservazioni 9.3.1

Ad esempio, nei processori Intel Core i7 lo spazio di indirizzamento fisico è scritto su 52 bit, mentre quello virtuale è su 48 bit, quindi può capitare che:

$$|\text{spazio fisico}| > |\text{spazio virtuale}|$$

Domanda 9.4

Se un sistema ha uno spazio di indirizzamento virtuale di X byte, significa che possiamo scrivere un programma che occupa al massimo X byte, cioè usa indirizzi virtuali da 0 a $X-1$. Tuttavia, un programma può girare su una macchina in cui:

$$|\text{RAM}| < |\text{spazio fisico}| < X?$$

Questo aspetto sarà approfondito nel capitolo sulla **memoria virtuale**.

9.4 Le librerie

Definizione 9.4.1

Una **libreria** è una collezione di subroutine di uso comune messe a disposizione dei programmatori per lo sviluppo software. Ad esempio, la libreria matematica del C fornisce funzioni come `sqrt(x)` per calcolare la radice quadrata.

Le librerie sono utili perché permettono di riutilizzare codice già esistente, evitando ai programmatori di doverlo riscrivere ogni volta. Sebbene "libreria" sia una traduzione impropria di "library", il termine è ormai comunemente accettato.

9.4.1 Tipi di Librerie

Esistono principalmente due tipi di librerie:

1. **Librerie statiche**: le subroutine sono collegate al programma principale durante la fase di compilazione o di caricamento, diventando parte dell'eseguibile. Tuttavia, ciò può portare a duplicazione di codice, sia su disco che in RAM, soprattutto se più programmi usano la stessa libreria. Inoltre, il codice di una libreria statica viene caricato in RAM anche se non viene utilizzato durante l'esecuzione del programma.

2. **Librerie dinamiche**: vengono caricate in RAM solo al momento in cui il programma chiama una subroutine specifica, ossia a **run-time**. Il programma specifica solo il nome della subroutine, e il sistema operativo carica la libreria nello spazio di memoria assegnato al processo. Queste librerie sono anche dette **librerie condivise**, perché possono essere utilizzate da più processi contemporaneamente, evitando la duplicazione di codice in RAM. Inoltre, le versioni aggiornate delle librerie dinamiche possono sostituire le vecchie senza dover ricompilare i programmi che le utilizzano.

9.4.2 Estensioni delle Librerie Dinamiche

In Unix, Linux e Solaris, le librerie dinamiche hanno estensione `.so` (shared object) e si trovano solitamente nella directory `/lib`. In ambiente Windows, le librerie dinamiche hanno estensione `.DLL` (Dynamic Link Library) e si trovano nella cartella `C:\WINDOWS\system32`.

9.5 Tecniche di gestione della memoria primaria

Le principali tecniche di gestione della **Memoria Principale (MP)** vanno dalle più semplici alle più complesse. Alcune di queste tecniche sono ormai obsolete, ma aiutano a comprendere concetti più avanzati. Le tecniche includono:

- **Swapping**
- **Allocazione contigua a partizioni multiple fisse**
- **Allocazione contigua a partizioni multiple variabili**
- **Paginazione**
- **Paginazione a più livelli**

Swapping

Definizione 9.5.1

Lo **swapping** consiste nel salvare in memoria secondaria (hard disk) l'immagine di un processo non in esecuzione (*swap out*) e ricaricarla in MP (*swap in*) prima di assegnarle la CPU.

Questa tecnica permette di attivare più processi di quanti la sola MP possa contenere, utilizzando un'area dell'hard disk chiamata **area di swap**, riservata al sistema operativo. Tuttavia, se il processo viene ricaricato in una diversa area di MP, è necessario utilizzare codice dinamicamente rilocabile.

Problemi dello Swapping

Il principale problema dello swapping è il tempo impiegato per copiare il codice e i dati di un processo tra l'hard disk e la RAM, che è nell'ordine dei millisecondi. Poiché in un millisecondo un singolo core di una moderna CPU può eseguire milioni di istruzioni, l'overhead di tempo risultante dallo swapping è generalmente considerato inaccettabile.

Di conseguenza, lo **swapping di interi processi** è ormai caduto in disuso nei moderni sistemi operativi, salvo rare eccezioni.

L'Idea di Fondo dello Swapping

Nonostante l'obsolescenza dello swapping, l'idea di fondo rimane valida: utilizzare parte della memoria secondaria per estendere la memoria primaria, permettendo l'esecuzione di un numero maggiore di processi rispetto a quanto potrebbe ospitare la sola RAM. Questa idea sarà ripresa nel capitolo sulla **memoria virtuale**.

9.6 Allocazione contigua della Memoria Primaria

In un computer, la **Memoria Principale (MP)** è solitamente divisa in due partizioni:

- una per il **Sistema Operativo (SO)**
- una per i **processi utente**.

Il sistema operativo si posiziona nella stessa area di memoria puntata dal **vettore delle interruzioni**, che è spesso allocato nella parte bassa della memoria.

Protezione della memoria

Nei sistemi operativi più semplici (ad esempio MS-DOS), l'area non assegnata al SO viene occupata da un solo processo. La protezione della MP consiste nella protezione delle aree di memoria del SO.

Registro Limite e Registro di Rilocazione

Il **registro limite** è inizializzato dal SO e garantisce che ogni indirizzo logico usato dal processo utente sia inferiore al valore scritto nel registro. Poiché si utilizza codice dinamicamente rilocabile, il **registro di rilocazione** viene usato per trasformare l'indirizzo logico in indirizzo fisico.

- **Registro di rilocazione**: 100.040
- **Registro limite**: 74.600

Gli indirizzi fisici validi vanno da 100.040 a 174.640 (vedi Fig. 9.3).

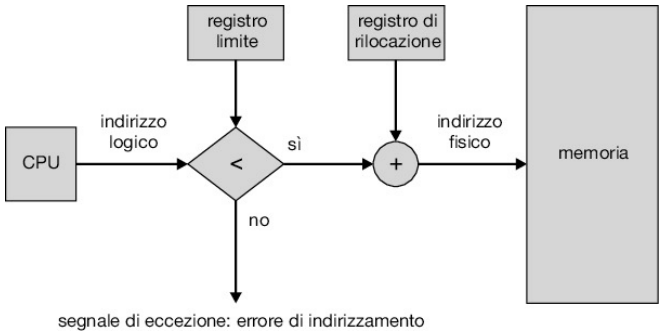


Figure 9.3: Protezione memoria

9.6.1 Allocazione a partizioni multiple fisse

Nell'allocazione a partizioni fisse:

- La memoria è divisa in **partizioni di dimensione fissa**, che non devono necessariamente essere tutte uguali.
- Ogni partizione contiene un **unico processo** dall'inizio alla fine della sua esecuzione.
- Il **grado di multiprogrammazione** è determinato dal numero di partizioni.
- Quando un processo termina, la partizione può essere occupata da un altro processo.

Il meccanismo di **registri limite e di rilocazione** protegge le partizioni da accessi non autorizzati. Durante il *context switch*, il **dispatcher** carica:

- Nel registro di rilocazione, l'indirizzo di partenza della partizione assegnata al processo.
- Nel registro limite, la dimensione della partizione.

Questa tecnica, utilizzata nel **IBM OS/360**, richiede CPU dotate di registri di rilocazione e limite, ma non è più utilizzata nei sistemi operativi moderni per i seguenti svantaggi:

- Il grado di multiprogrammazione è limitato dal numero di partizioni disponibili.
- Si verifica **frammentazione interna**, dove una parte della partizione rimane inutilizzata se il processo è più piccolo della partizione stessa.
- Si può verificare **frammentazione esterna**, quando lo spazio libero disponibile è frammentato in più aree non contigue e quindi non utilizzabile per un nuovo processo di dimensione maggiore.

- L'arrivo di un processo più grande della partizione più grande non può essere gestito.

Frammentazione interna:

- Parte dello spazio di memoria di una partizione viene sprecato se il processo è più piccolo della partizione stessa.

Frammentazione esterna:

- Se la memoria libera è frammentata in più blocchi non contigui, pur avendo spazio sufficiente in totale, non può essere utilizzata per allocare nuovi processi.

L'allocazione a partizioni fisse ha anche altri problemi:

Domanda 9.5

Che succede se arriva un processo più grande della partizione più grande?

Osservazioni 9.6.1

Notate che se si aumenta la dimensione media delle partizioni, aumenta anche la frammentazione interna, e diminuisce il grado di multiprogrammazione

9.7 Allocazione a partizioni multipli variabili

Nell'allocazione a partizioni variabili:

- Ogni processo riceve una quantità di memoria esattamente pari alla sua dimensione.
- Quando un processo termina, lascia un "buco" in RAM che può essere occupato da un altro processo.
- Tuttavia, nel tempo si creano **buchi sparsi e più piccoli**, rendendo difficile l'allocazione di nuovi processi.

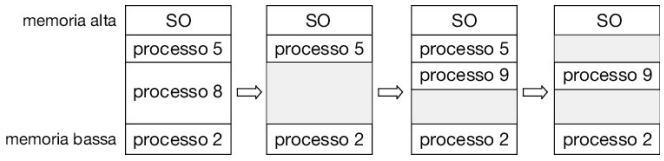


Figure 9.4: Buchi Ram

Il sistema operativo (SO) deve:

- Tenere traccia delle aree di memoria libere e occupate.
- Aggiornare continuamente le informazioni quando un processo nasce o termina.
- Assegnare una partizione sufficientemente grande quando un nuovo processo deve essere caricato.

Strategie di Allocazione

Esistono diverse strategie per scegliere quale partizione assegnare a un processo:

- **First Fit**: seleziona la prima partizione abbastanza grande da ospitare il processo.
- **Best Fit**: seleziona la partizione più piccola che può contenere il processo.
- **Worst Fit**: seleziona la partizione più grande disponibile.

Osservazioni sperimentali:

- La strategia **Worst Fit** tende a funzionare peggio in termini di utilizzo della memoria, poiché lascia spazi grandi frammentati.
- Le strategie **Best Fit** e **First Fit** hanno prestazioni simili, ma si preferisce **First Fit** poiché è più veloce, dato che interrompe la ricerca al primo spazio sufficiente.

9.7.1 La frammentazione

Nel tempo, l'allocazione a partizioni variabili può portare alla formazione di piccoli **buchi non contigui** in RAM:

- Circa **1/3 a 1/2 della memoria principale** (MP) può essere sprecato a causa della **frammentazione esterna**, ossia la presenza di buchi di memoria troppo piccoli per ospitare un processo.
- Esiste anche il problema della **frammentazione interna**, dovuto all'impossibilità di tenere traccia di buchi molto piccoli che vengono quindi aggregati a partizioni adiacenti, causando uno **spreco nascosto**.

Compattazione della Memoria

Una tecnica per recuperare la memoria inutilizzata è la **compattazione**:

- Spostare le partizioni occupate dai processi in modo che siano tutte **contigue**, liberando un unico grande buco libero.
- La compattazione richiede la **rilocazione dinamica** del codice e dei dati dei processi.
- Questo processo è **costoso in termini di tempo** e durante la compattazione il sistema non è utilizzabile.

9.8 Paginazione della memoria

Definizione 9.8.1

L'allocazione contigua della memoria principale presenta quindi diversi problemi. L'alternativa è ammettere che l'area di memoria allocata ad un processo possa essere in realtà suddivisa in tanti pezzi non contigui fra loro Se tutti i "pezzi" hanno la stessa dimensione allora il termine esatto per indicare questa tecnica è: paginazione della memoria (primaria)

9.8.1 Metodo di base

La **Memoria Primaria** (o lo **spazio di indirizzamento fisico**) è suddivisa in unità dette **frame** (o pagine fisiche), con le seguenti caratteristiche:

- I **frame** sono tutti della stessa dimensione, che è sempre una **potenza di due** (ad esempio: 512, 1024, 2048, fino a 8192 byte).
- Lo spazio di indirizzamento fisico del processo è visto come un **unico spazio contiguo** di indirizzi, ma nella realtà è suddiviso in **pagine logiche**, ciascuna di dimensione uguale ai frame fisici.

Per eseguire un processo con **x pagine**:

- Il Sistema Operativo (**SO**) cerca **x frame** liberi in cui caricare le pagine del processo. Questi frame non devono essere adiacenti e le pagine possono essere caricate in un ordine qualsiasi.
- Ogni processo ha una propria **Tabella delle Pagine** (o **Page Table**, **PT**): un array le cui *entry* contengono i numeri dei frame in cui le pagine del processo sono state caricate.
- Il SO tiene traccia dei **frame liberi** nella memoria primaria, utilizzandoli per memorizzare le pagine di nuovi processi.

Struttura della Tabella delle Pagine

- Ogni *entry* della Page Table rappresenta una pagina del processo.
- L'indice di ciascuna entry corrisponde al numero della pagina, mentre il valore dell'entry contiene il numero del frame dove è stata memorizzata la pagina.

Problema degli Indirizzi Virtuali

Gli **indirizzi relativi (o virtuali)** del programma, una volta caricati in RAM, non funzionano più come indirizzi lineari contigui. Ad esempio:

- Un'istruzione come `jmp_if_odd R1,C` deve saltare a un indirizzo specifico (ad es. 0004), ma in RAM non esiste più un punto lineare di partenza.
- La soluzione è **riconsiderare** gli indirizzi virtuali, non più come indirizzi lineari, ma come indirizzi all'interno della Page Table, utilizzando **conversioni da indirizzi logici a indirizzi fisici**.

Indirizzi Logici e Fisici, vedere registrazione 25/10/2024 min Circa 30

Paginazione: metodo di base

Gli indirizzi logici diventano delle coppie di valori, in cui:

- il primo elemento della coppia specifica il numero della pagina all'interno della quale si trova la cella di memoria che vogliamo indirizzare;
- il secondo elemento specifica la posizione (o offset) della cella di memoria che vogliamo indirizzare rispetto ad un ipotetico indirizzo 0 (zero), ovvero l'indirizzo del primo byte della pagina specificata dal primo elemento della coppia.

Quindi un indirizzo logico assume la forma: (pagina, offset).

Note:-

Come vedremo più avanti, sotto opportune condizioni gli indirizzi logici lineari e gli indirizzi logici specificati come coppie di valori coincidono.

Un indirizzo logico viene tradotto in uno fisico secondo il seguente processo:

- Il numero di pagina p viene usato come indice nella *page table* del processo per individuare il frame f in cui è contenuta la pagina.
- Una volta noto il frame f che contiene la pagina p, l'offset d (dove "d" sta per *displacement*) può essere applicato a partire dall'inizio del frame per indirizzare il byte specificato dalla coppia (p, d).

Vediamo più in dettaglio: ogni informazione all'interno del computer, incluso un indirizzo logico, è in definitiva una sequenza di bit. Ad esempio: 001100010101.

Abbiamo deciso di dividere un indirizzo logico in due parti. Se abbiamo a disposizione 12 bit in tutto, dobbiamo scegliere quanti usare per specificare il numero della pagina e quanti per l'offset all'interno della pagina.

Ad esempio, possiamo decidere di utilizzare 4 bit per rappresentare il numero della pagina e i restanti 8 bit per l'offset. In questo caso, l'indirizzo logico 001100010101 rappresenta:

- pagina: 3
- offset: 21

La scelta del numero di bit da utilizzare per scrivere il numero di pagina e l'offset dipende dall'hardware su cui dovrà girare il sistema operativo, il quale impone:

- il numero di bit su cui va scritto un indirizzo logico (ad esempio m bit);
- la dimensione di un frame, e quindi di una pagina (ad esempio 2^n byte).

Di conseguenza, dobbiamo utilizzare n bit per rappresentare l'offset all'interno di una pagina/frame, e il numero di bit usati per rappresentare il numero di pagina sarà pari a $m - n$ bit.

Note:-

A questo punto, la dimensione dello spazio di indirizzamento logico è stabilita come $2^{(m-n)} \times 2^n$ byte.

Esempio 9.8.1 (Esempio di Spazio di Indirizzamento Logico)

Consideriamo una macchina in cui i frame hanno dimensione 2^{12} byte, ovvero 4096 byte. Di conseguenza, anche la dimensione di una pagina nello spazio di indirizzamento logico del sistema sarà di 4096 byte (quindi $n = 12$).
Se la macchina mette a disposizione $m = 22$ bit per rappresentare un indirizzo logico, allora il numero di bit necessari per il numero di pagina sarà $m - n = 10$ bit.
In questo caso, lo spazio di indirizzamento logico della macchina risulterà pari a:

$$(2^{10} \text{ pagine}) \times (2^{12} \text{ byte}) = 4 \text{ megabyte}$$

Struttura di un Indirizzo Logico

- **Numero di Pagina (p)**: $m - n$ bit
- **Offset di Pagina (d)**: n bit

Note:-

Qui possiamo osservare uno dei tanti casi di interazione tra il sistema operativo e l'hardware sottostante. È infatti l'hardware a decidere la dimensione dei frame e il numero di bit su cui rappresentare un indirizzo logico. Il sistema operativo, quindi, deve adeguarsi a queste impostazioni.
Questa configurazione consente una traduzione efficiente degli indirizzi logici in indirizzi fisici direttamente a livello hardware. Alcune parti di un sistema operativo sono infatti progettate in base allo specifico hardware su cui girerà il sistema, al fine di sfruttare al meglio le caratteristiche hardware e minimizzare l'overhead introdotto.

Osservazioni 9.8.1 Scelta della Dimensione dei Frame

Nei processori moderni, è possibile scegliere tra vari valori per la dimensione dei frame, e questa scelta, una volta effettuata, rimane fissa. Ad esempio:

- I processori ARM, utilizzati negli iPhone e negli iPad, permettono di scegliere tra 4KB, 16KB, 64KB, 1MB e 16MB come dimensione di un frame.
- La famiglia dei processori Intel, dal Pentium fino ai Core i9, consente di scegliere tra 4KB e 4MB.
- Gli Intel Itanium-2 offrono una gamma più ampia, con opzioni tra 4KB, 8KB, 64KB, 256KB, 1MB, 4MB, 16MB e 256MB.

Definizione 9.8.2: Traduzione degli Indirizzi Logici in Indirizzi Fisici

Definiamo più precisamente l'operazione di traduzione da indirizzi logici a indirizzi fisici. Un indirizzo logico è formato da due componenti (p, d):

- **Numero di Pagina (p)**: utilizzato come indice per selezionare la entry corrispondente nella *Page Table*, che contiene il numero del frame in cui è caricata la pagina.
- **Offset di Pagina (d)**: utilizzato all'interno del frame individuato al passo precedente per localizzare il byte specificato dall'indirizzo logico.

Paginazione: traduzione degli indirizzi

Possiamo ora riassumere e integrare quanto detto finora, osservando che:

1. Gli indirizzi logici nello spazio di indirizzamento logico possono essere interpretati come valori lineari oppure come coppie (pagina, offset).
2. Ciascuna entry di una tabella delle pagine può contenere il numero del frame o, alternativamente, l'indirizzo di partenza del frame. Tuttavia, per ragioni pratiche, viene scelta la prima opzione: memorizzare il numero del frame.

Note:-

Ecco quindi la doppia natura degli indirizzi logici, che sono contemporaneamente valori lineari e coppie (pagina, offset).

Consideriamo uno spazio di indirizzamento logico di dimensione 2^m byte, e assumiamo che ogni pagina abbia una dimensione pari a 2^n byte. Un indirizzo logico lineare scritto su m bit può quindi essere scomposto in:

- i bit più significativi $(m - n)$, che indicano il numero di pagina p;
- i bit meno significativi n, che rappresentano l'offset d.

Struttura di un Indirizzo Logico

- **Numero di Pagina (p)**: $(m - n)$ bit
- **Offset di Pagina (d)**: n bit

Esempio 9.8.2 (Esempio di Spazio di Indirizzamento Logico)

Supponiamo di avere uno spazio di indirizzamento logico di 16 byte, quindi gli indirizzi logici sono rappresentati con $m = 4$ bit, poiché $2^4 = 16$. Ogni pagina ha dimensione $2^2 = 4$ byte, con $n = 2$.
Di conseguenza, gli indirizzi logici vanno da 0000 a 1111, mentre il programma che occupa questo spazio è costituito dalle istruzioni/dati etichettati come "a, b, c, d, e, f, g, h, i, l, m, n" che coprono tre pagine, dove ciascuna istruzione/dato occupa un byte.

Note:-

In questo esempio, lo spazio di indirizzamento logico del programma è suddiviso in pagine. Gli indirizzi di istruzioni e dati, che costituiscono lo spazio di indirizzamento logico, sono quindi valori consecutivi che vanno da 0000 (prima istruzione) a 1011 (ultima istruzione).

Osserviamo come gli $(m - n)$ bit più significativi di ciascun indirizzo rappresentano il numero di pagina, mentre i restanti n bit meno significativi indicano l'offset all'interno della pagina.

Indirizzo Logico	Contenuto
00 00	a
00 01	b
00 10	c
00 11	d
01 00	e
01 01	f
01 10	g
01 11	h
10 00	i
10 01	l
10 10	m
10 11	n

Ogni indirizzo è quindi composto da:

- I primi due bit, che rappresentano il numero di pagina:
  - 00 per la pagina 0
  - 01 per la pagina 1
  - 10 per la pagina 2
- Gli ultimi due bit, che rappresentano l'offset all'interno della pagina.

Note:-

Supponiamo ora di caricare il programma in RAM. Utilizziamo 5 bit per rappresentare un indirizzo fisico, quindi lo spazio di indirizzamento fisico ha dimensione pari a $2^5 = 32$ byte. Tuttavia, la nostra macchina è dotata di soli 20 byte di RAM, suddivisi in 5 frame, ciascuno con un indirizzo che va da 00000 a 10011.

Esempio 9.8.3 (Esempio di Traduzione Indirizzo Logico in Indirizzo Fisico)

Consideriamo l'indirizzo logico 0010. La sua conversione in un indirizzo fisico procede così:

1. **Numero di Pagina**: il numero di pagina (0) è utilizzato per accedere alla *Page Table*, che ci indica che la pagina 0 è caricata nel frame 4.
2. **Offset**: l'offset di 2 bit (10) viene applicato a partire dall'indirizzo iniziale del frame 4.

Di conseguenza, l'indirizzo fisico risultante è 10010.

Osservazioni 9.8.2 Struttura di un Indirizzo Fisico

Notiamo un aspetto importante: nei nostri indirizzi fisici, i bit più significativi (5 - 2 = 3 bit) indicano il numero di frame corrispondente. Ad esempio, per l'indirizzo 10010, i primi 3 bit 100 identificano il frame 4.

Note:-

Le pagine e i frame sono configurati per avere una dimensione $|P|$ pari a una potenza di 2, ossia $|P| = 2^n$. Grazie a questa proprietà, l'offset all'interno di ogni pagina o frame varia da una configurazione di tutti zeri a una di tutti uno:

$$\text{Offset valido: } 00\ldots00 \text{ a } 11\ldots11.$$

Di conseguenza, l'indirizzo di partenza di ogni pagina o frame richiede che gli n bit meno significativi siano impostati a 0.

Osservazioni 9.8.3 Struttura degli Indirizzi di Partenza

Gli indirizzi logici di inizio pagina nello spazio di indirizzamento di un programma avranno la forma seguente:

$00\ldots00\ 00\ldots00$ inizio della pagina 0
$00\ldots01\ 00\ldots00$ inizio della pagina 1
$00\ldots10\ 00\ldots00$ inizio della pagina 2
$00\ldots11\ 00\ldots00$ inizio della pagina 3

Lo stesso vale per i frame nello spazio di indirizzamento fisico.

Note:-

Se rimuoviamo gli n bit meno significativi da un indirizzo logico di m bit, i rimanenti $m - n$ bit contano semplicemente le pagine (in binario) in cui è suddiviso lo spazio di indirizzamento logico. Ad esempio:

$00\ldots00$ = pagina 0
$00\ldots01$ = pagina 1
$00\ldots10$ = pagina 2
$00\ldots11$ = pagina 3

Questa logica vale anche per i frame nello spazio fisico, utilizzando un numero di bit adeguato per rappresentare i frame.

Esempio 9.8.4 (Ricostruzione dell'Indirizzo Fisico)

Supponiamo di avere un indirizzo logico in cui il numero di pagina e l'offset sono rappresentati come segue:

$$\text{Numero di pagina } = 001010, \quad \text{Frame } = 1110001$$

Offset = 101010

Allora, l'indirizzo logico diventa 001010101010, mentre l'indirizzo fisico corrispondente sarà 1110001101010.

Note:-

La conversione da indirizzo logico a fisico si basa sull'operazione di somma dell'indirizzo base del frame con l'offset, ad esempio:

Indirizzo base del frame: 1110001000000
Offset: 101010

Somma:

1110001000000 + 101010 = 1110001101010

Osservazioni 9.8.4 Efficienza della Traduzione

Se le dimensioni di pagine e frame sono potenze di due:

1. Non è necessario effettuare la somma tra l'indirizzo base del frame e l'offset, risparmiando tempo di calcolo.
2. Non è necessario memorizzare l'indirizzo di base in ogni entry della page table, ma solo il numero del frame, poiché gli n bit meno significativi dell'indirizzo di partenza sono tutti a 0, con un risparmio di spazio.

Generare un indirizzo fisico diventa quindi un'operazione di concatenazione veloce tra il numero di frame e l'offset, eseguibile direttamente dall'hardware.

Note:-

Quando le dimensioni di pagine e frame sono potenze di due, possiamo interpretare l'operazione di calcolo dell'indirizzo fisico come una concatenazione, anziché una somma:

$$1110001000000 + 101010 = 1110001101010 \quad \Rightarrow \quad 1110001 \text{ “attaccato a” } 101010.$$

Usando potenze di due, la costruzione dell'indirizzo fisico è più semplice e può essere gestita velocemente dall'hardware.

Esempio 9.8.5 (Esempio: Dimensione della Page Table)

Consideriamo un sistema in cui:

- l'indirizzo fisico è su 38 bit,
- l'indirizzo logico è su 40 bit,
- una pagina è di 8 Kbyte, quindi 2^{13} byte.

La dimensione della page table più grande possibile per questo sistema si calcola come segue:

1. **Calcolo del numero di entry nella page table**:
$$\frac{2^{40} \text{ byte}}{2^{13} \text{ byte}} = 2^{27} \text{ entry}.$$
2. **Numero di bit necessari per ogni entry**: Ogni entry deve essere in grado di indirizzare un frame nello spazio fisico. Lo spazio di indirizzamento fisico è diviso in:
$$\frac{2^{38} \text{ byte}}{2^{13} \text{ byte}} = 2^{25} \text{ frame}.$$
Il numero minimo di byte per rappresentare un frame è di 4 byte.
3. **Calcolo della dimensione totale della page table**:
$$2^{27} \times 2^2 \text{ byte} = 2^{29} \text{ byte} = 512 \text{ Mbyte}.$$

La paginazione permette di separare lo spazio di indirizzamento logico da quello fisico. Ogni programma "vede" la memoria come uno spazio contiguo che parte sempre dall'indirizzo logico 0, ma in realtà il programma è distribuito in diversi frame fisici, sparsi in memoria assieme ad altri programmi. La paginazione introduce una protezione automatica dello spazio di indirizzamento. Un processo può accedere solo ai frame elencati nella sua page table, poiché ogni page table viene gestita e costruita dal sistema operativo per ciascun processo.

Osservazioni 9.8.5 Vantaggi e Limiti della Paginazione

- **Eliminazione della frammentazione esterna**: Ogni frame libero può essere utilizzato per memorizzare una pagina, riducendo gli sprechi di memoria.
- **Frammentazione interna**: Rimane una media di mezza pagina per processo, poiché l'ultima pagina del processo può non occupare completamente il frame assegnato.
- **Rilocazione dinamica**: La paginazione implementa una forma di rilocazione dinamica, dove ogni pagina è mappata su un diverso valore del registro di rilocazione, ossia l'indirizzo di partenza del frame che contiene la pagina.

9.8.2 Rilocazione Dinamica

Rilocazione Dinamica Tramite Registri di Limite e di Rilocazione

La **rilocazione dinamica** viene gestita inizialmente utilizzando due registri principali:

- **Registro di limite**: Controlla che l'indirizzo logico rientri nello spazio di indirizzamento del processo. Se l'indirizzo eccede questo limite, il sistema lancia un'eccezione di indirizzamento.
- **Registro di rilocazione**: Viene usato per traslare l'indirizzo logico approvato in un indirizzo fisico, aggiungendo il valore del registro di rilocazione all'indirizzo logico (vedi Fig. 9.6).

Questo approccio verifica dunque prima la validità dell'indirizzo logico rispetto allo spazio di indirizzamento del processo, per poi calcolare l'indirizzo fisico finale.

Rilocazione Dinamica di Ogni Singola Pagina Tramite la Page Table

Con la **paginazione**, la gestione della rilocazione dinamica cambia in quanto:

- La suddivisione in pagine elimina la necessità di un controllo tramite il registro limite.
- Ogni indirizzo logico è diviso in **numero di pagina** e **offset**, dove l'offset rappresenta sempre una posizione interna al frame associato, eliminando il rischio di indirizzamenti fuori dai limiti (vedi Fig. 9.8).

In questo caso, il controllo e la mappatura degli indirizzi sono effettuati dalla page table, la quale memorizza l'indirizzo di ciascun frame per ogni pagina del processo attivo.

9.8.3 Dimensione delle Pagine e Gestione della Tabella delle Pagine

Dimensione Storica delle Pagine

Nel tempo, la **dimensione delle pagine** di memoria è aumentata progressivamente per adattarsi alle esigenze di gestione della memoria dei sistemi operativi:

- Le dimensioni comuni oggi includono pagine da **4 KB**, **8 KB**, **16 KB** e possono arrivare fino a **256 MB**.
- **Vantaggi di pagine più grandi**: Permettono di ridurre la lunghezza delle page table, il che riduce lo spazio richiesto in RAM per mantenere la tabella.
- **Svantaggi di pagine più grandi**: Producono una maggiore frammentazione interna, poiché una pagina più grande può avere porzioni non completamente utilizzate da un processo.

Page Table e Frame Table

A ogni processo è associata una **tabella delle pagine (page table)**, che contiene informazioni sui frame assegnati alle pagine del processo. La gestione della memoria fisica richiede inoltre che il sistema operativo mantenga una **frame table**, che descrive lo stato di ogni frame nella memoria fisica:

- Indica quali frame sono liberi e quali sono occupati.
- Registra quale pagina di quale processo occupa ogni frame.

Attivazione della Page Table e Contesto del Processo

Ogni volta che avviene un **context switch** (ovvero, quando la CPU passa da un processo a un altro), il sistema operativo deve attivare la page table del nuovo processo:

- Ciò comporta un **tempo di configurazione** aggiuntivo durante il cambio di contesto, poiché la page table del nuovo processo deve essere caricata e attivata.
- L'attivazione della page table incide sulle performance del sistema, rallentando potenzialmente il cambio di contesto.

Supporto Hardware alla Paginazione

La **paginazione richiede un supporto hardware** per garantire efficienza, poiché ogni accesso alla memoria passa per il meccanismo di traduzione da indirizzi logici a indirizzi fisici. La sfida principale risiede nella gestione della page table:

- Ogni indirizzo logico generato dalla CPU deve passare attraverso una entry della page table, quindi l'**accesso alla page table deve essere rapido**.
- In sistemi con un numero ridotto di pagine per processo, come nel caso del **PDP-11** (che utilizzava **8 registri** per memorizzare la page table con una memoria fisica di 64 KB, divisa in 8 frame da 8 KB), la page table poteva essere memorizzata in registri della CPU.
- Tuttavia, nei computer moderni, le page table contengono migliaia di entry e non è possibile mantenerle tutte all'interno dei registri della CPU.

9.8.4 Supporto Hardware

Siamo costretti a tenere la PT di ogni processo in MP, consumando spazio. Questo rende preferibile l'uso di pagine più grandi. Per accedere a un dato all'indirizzo logico I, dobbiamo prima accedere alla RAM per recuperare il numero del frame e calcolare l'indirizzo fisico, e solo successivamente utilizzarlo per leggere il dato. In questo modo il numero di accessi alla memoria principale raddoppia, il che è inaccettabile, poiché un accesso in RAM può costare oltre 100 cicli di clock, rispetto ai 5 cicli necessari se il dato si trova in un registro o nella cache L1 della CPU. Per migliorare l'efficienza e ridurre i tempi di accesso alla PT, si utilizza una tecnica di caching della PT tramite una memoria associativa della CPU, detta *Translation Look-aside Buffer* (TLB). Questo esempio mostra come i progettisti hardware possano dotare i processori di dispositivi che facilitano e ottimizzano il lavoro del SO. Per facilitare l'accesso rapido alla PT, il sistema operativo ha anche un registro dedicato che contiene l'indirizzo di partenza della PT del processo attivo, detto *Page-Table Base Register* (PTBR). Durante un *context switch*, è sufficiente aggiornare il valore del PTBR per "attivare" la PT del nuovo processo in esecuzione. La memoria associativa è costituita da coppie chiave–valore. Fornendo una chiave come input, essa viene confrontata simultaneamente con tutte le chiavi memorizzate, restituendo il valore associato. Questo dispositivo, benché molto veloce e costoso, ha dimensioni ridotte (tra 64 e 1024 elementi).

La TLB contiene una porzione della PT, costituita da coppie chiave–valore sotto forma di numero di pagina (chiave) e relativo frame (valore). Quando viene generato un indirizzo logico, il numero di pagina viene fornito alla TLB per ottenere il frame associato.

> **Corollario 9.8.1** Hit Ratio nella TLB
>
> Il termine *Hit Ratio* indica la percentuale di successi (hit) nella ricerca del numero di pagina nella TLB.

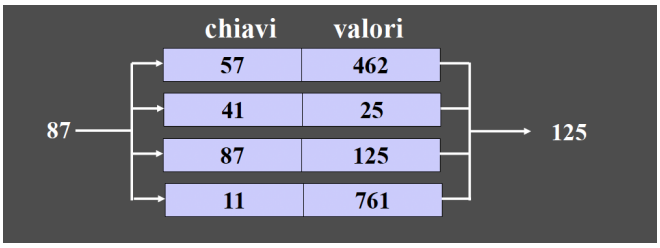


Figure 9.5: chiave-valorePTB

Un maggiore hit ratio implica una minore degradazione delle prestazioni.

Esempio 9.8.6 (Calcolo delle prestazioni con e senza TLB)

Supponiamo: 1) un TLB perfetto, ossia accedere al TLB non costa tempo; 2) 10 nanosec per accedere alla RAM una volta tradotto l'indirizzo logico in fisico; 3) *hit ratio* pari all'80%. Il tempo medio di accesso in RAM sarà:

$$10\,\text{nsec} \times 0,80 + (10 + 10)\,\text{nsec} \times 0,20 = 12\,\text{nsec}$$

Il che comporta una degradazione delle prestazioni del 20%.
Se il TLB avesse invece un *hit ratio* del 99%, il tempo medio di accesso alla memoria sarebbe:

$$10\,\text{nsec} \times 0{,}99 + (10 + 10)\,\text{nsec} \times 0{,}01 = 10{,}1\,\text{nsec}$$

Questo porterebbe a una degradazione delle prestazioni di appena l'1%.
Se, invece, consideriamo che il TLB abbia un tempo di accesso maggiore di zero, diciamo all'incirca 1 nsec (cioè un decimo del tempo di accesso in RAM), con un *hit ratio* del 99% il tempo medio di accesso alla memoria sarebbe:

$$(10 + 1)\,\text{nsec} \times 0{,}99 + (10 + 10)\,\text{nsec} \times 0{,}01 = 11{,}09\,\text{nsec}$$

In questo caso la degradazione delle prestazioni sarebbe di quasi l'11%.
Quando si verifica un *miss* nella TLB, la coppia pagina-frame mancante viene recuperata attraverso la PT in RAM e copiata nella TLB. Questo permette che i successivi riferimenti alla stessa pagina usino la copia memorizzata nella TLB, riducendo il tempo di accesso.
Se il TLB è pieno, una delle sue *entry* deve essere sovrascritta, solitamente scegliendo l'elemento meno recentemente utilizzato (Least Recently Used, LRU) o selezionandone uno casuale.

Note:-

Al *context switch* il TLB deve essere svuotato; verrà successivamente ripopolato con le coppie pagina-frame del nuovo processo in esecuzione.

9.8.5 Pagine condivise

Quando due processi eseguono lo stesso codice, mantenere in memoria principale (MP) due copie identiche del codice non solo è inutile, ma spreca anche spazio in RAM. La paginazione facilita la condivisione del codice, poiché permette di memorizzare una sola copia di codice, condivisa tra i processi. Questo è possibile perché il codice, essendo *non modificabile* durante l'esecuzione, viene definito come *codice puro* o *rientrante*.

Note:-

Una pagina condivisa può essere utilizzata per contenere codice di librerie dinamiche che diversi processi possono usare contemporaneamente, riducendo ulteriormente lo spazio richiesto in MP per il codice eseguito da più processi.

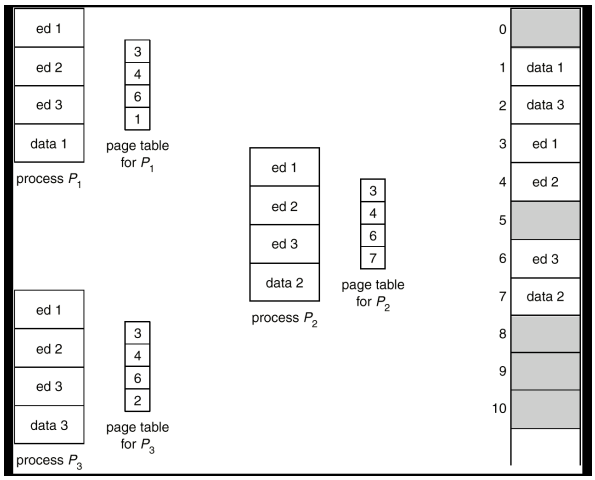


Figure 9.6: shared$_c ode_p aginatedEnv.png$

9.9 Paginazione a più livelli

Nei moderni calcolatori, lo spazio di indirizzi logici può ormai raggiungere anche 2^{64} byte. Per un sistema con 2^{32} byte di spazio logico e pagine da 4 Kbyte (2^{12} byte), la *Page Table* (PT) può avere fino a un milione (2^{20}) di entry. Se ogni entry occupa 4 byte, la PT del processo occuperà quindi 4 Mbyte, richiedendo ben 1024 frame per essere completamente contenuta in memoria principale (MP).

Domanda 9.6

Considerando questi numeri, è possibile ipotizzare la dimensione massima dello spazio di indirizzamento fisico del sistema?

Assumendo uno spazio logico di 2^{32} byte, pagine da 4 Kbyte (2^{12} byte), e una dimensione di 4 byte per ogni entry della PT, si ha che ogni entry della PT deve contenere il numero di un frame del sistema. Con 32 bit a disposizione, possiamo numerare fino a 2^{32} frame (dal frame 0 al frame $2^{32} - 1$). Dunque, lo spazio di indirizzamento fisico del sistema può contenere al massimo 2^{32} frame e avere una dimensione massima di $2^{32} \times 2^{12} = 2^{44}$ byte.

Note:-

L'uso della paginazione consente di evitare la necessità di allocare grandi aree contigue di memoria principale, ma può succedere che la PT del processo attivo sia comunque molto grande, creando problemi di allocazione.

Una possibile soluzione consiste nell'implementare una *paginazione a due livelli*, che suddivide la PT in pagine memorizzate in frame non adiacenti in MP. In questo caso, la PT vera e propria (ora chiamata *PT interna*) richiede una *PT esterna*, che indica in quali frame sono memorizzate le pagine della PT interna.

Esempio 9.9.1 (Esempio di paginazione a due livelli)

Consideriamo una macchina con 32 bit di spazio di indirizzamento logico e fisico, e con pagine/frame da 4 Kbyte. In questo caso, un indirizzo logico sarà composto da:

- 20 bit per il numero della pagina
- 12 bit per l'offset all'interno della pagina.

Il numero di pagina p, espresso su 20 bit, sarà quindi ulteriormente suddiviso in:

- 10 bit più significativi (p_1): entry della PT esterna, che punta al frame F_1 contenente una porzione della PT interna.
- 10 bit intermedi (p_2): offset nel frame F_1.

Pertanto, l'indirizzo logico sarà strutturato come segue:

$$p = p_1\ p_2\ d$$

dove:

- p_1: 10 bit per la pagina esterna
- p_2: 10 bit per la pagina interna
- d: 12 bit per l'offset all'interno della pagina.

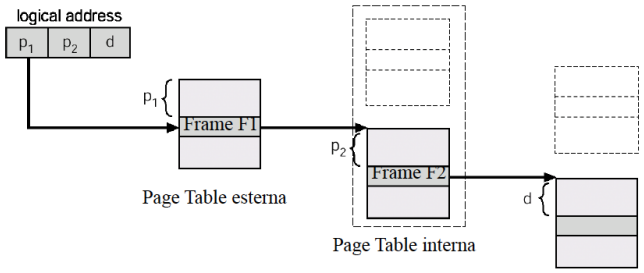


Figure 9.7: traduzione$_i$*ndirizzi*$_p$*aginazione2Livelli.png*

Esempio 9.9.2 (Esempio di calcolo dell'indirizzo fisico)

Per comprendere il funzionamento della paginazione a due livelli in termini numerici, consideriamo un sistema in cui ogni entry delle PT occupa 4 byte (anche se tecnicamente sarebbero sufficienti 20 bit, poiché il numero massimo di frame è limitato). Prendiamo in esame la PT più grande, denominata *PT interna*, che contiene 2^{20} entry e quindi ha una dimensione di 4×2^{20} byte, richiedendo esattamente $2^{10} = 1024$ frame.
La PT interna è quindi memorizzata in 1024 frame non contigui. Per tracciare la loro allocazione, il sistema operativo costruisce una *PT esterna*, che occupa esattamente un frame (poiché ciascuna delle sue 1024 entry occupa 4 byte).
I passi per tradurre un indirizzo logico V di 32 bit in un indirizzo fisico sono i seguenti:

- I 10 bit più significativi di V, indicati come p_1, servono per individuare una delle 1024 entry nella PT esterna, che funziona come un array di 1024 entry da 4 byte ciascuna.
- In questa entry, viene recuperato il numero del frame F_1 che contiene una delle pagine della PT interna.
- Utilizzando i 10 bit intermedi di V (p_2), accediamo a una delle 1024 entry del frame F_1. Questa entry contiene il numero del frame F_2 che memorizza la pagina dell'indirizzo logico V (definito ora da p_1p_2).
- Infine, aggiungendo l'offset d a F_2, otteniamo l'indirizzo fisico finale.

9.10 Paginazione a due livelli

La **paginazione a due livelli** è stata utilizzata, ad esempio, nei vecchi processori Pentium (lo vedremo successivamente come esempio pratico). Altre architetture, come la **VAX** di Digital Equipment Corporation (DEC), implementavano una *PT esterna* composta da soli 4 elementi.

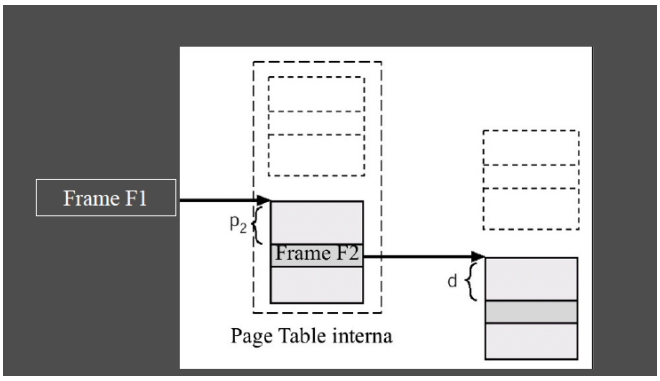


Figure 9.8: ex$_p$*aginazione2Livelli.png*

Domanda 9.7: Gestione dello spazio logico su 64 bit

Che cosa accade con uno spazio di indirizzamento logico di 64 bit?
Dimostrazione: Con pagine da 4 Kbyte e 4 byte per entry nelle PT, la PT esterna può arrivare a occupare 2^{44} byte. (Provate a verificare questo valore!)

Consideriamo ora la **PT interna** più grande con un indirizzamento logico a 64 bit:

- Tale PT richiede $\frac{2^{64}}{2^{12}} = 2^{52}$ entry, occupando dunque $2^{52} \times 2^2 = 2^{54}$ byte.
- Questo corrisponde a $\frac{2^{54}}{2^{12}} = 2^{42}$ frame.
- La PT esterna necessiterà quindi di 2^{42} entry, ognuna da 4 byte, raggiungendo una dimensione di $2^{42} \times 2^2 = 2^{44}$ byte.

Note:-

Con un indirizzamento logico a 64 bit, anche la PT esterna diventa così grande da necessitare una paginazione ulteriore. Questo problema non si limita alle architetture a 64 bit; anche alcune architetture a 32 bit implementavano paginazioni su più livelli:

- *SPARC* (SUN Microsystems) a 32 bit: utilizzava una paginazione a 3 livelli.
- CPU a 32 bit *Motorola 68030*: implementava uno schema di paginazione a 4 livelli.

Osservazioni 9.10.1 Overhead della paginazione su architetture a 64 bit

In sistemi a 64 bit, nemmeno 4 livelli di paginazione risultano sufficienti. Per esempio, l'architettura *UltraSPARC* richiede fino a 7 livelli di paginazione. Se una pagina non è disponibile nel *TLB*, la traduzione da indirizzo logico a fisico può richiedere l'attraversamento di 7 livelli di pagine in RAM, causando un notevole overhead.

9.10.1 Page Table Invertita (IPT)

Una soluzione alternativa adottata in alcune architetture a 64 bit è la **Tabella delle Pagine Invertita (IPT)**, la cui gestione presenta caratteristiche diverse rispetto alla paginazione tradizionale.

- Una *IPT* descrive l'occupazione dei frame nella memoria fisica. Al contrario delle *PT*, esiste una sola *IPT* per tutto il sistema (invece di una per ciascun processo), riducendo così lo spreco di memoria.
- La dimensione dell'*IPT* dipende esclusivamente dalla dimensione della memoria primaria: ogni entry rappresenta un frame specifico, e il numero di entry totali equivale al numero di frame.
- L'indice di ogni entry dell'*IPT* corrisponde al numero di un frame in memoria principale.

Ogni entry dell'*IPT* è costituita da una coppia di valori:

$$\langle \text{process-id}, \text{page-number} \rangle$$

dove:

- *process-id* identifica il processo proprietario della pagina.
- *page-number* indica il numero della pagina contenuta nel frame rappresentato da quella entry.

Ogni indirizzo logico generato dalla CPU è quindi una tripla:

$$\langle \text{process-id}, \text{page-number}, \text{offset} \rangle$$

Per generare l'indirizzo fisico, si cerca nella *IPT* la coppia $\langle \text{process-id}, \text{page-number} \rangle$. Se viene trovata nella i-esima entry, l'indirizzo fisico sarà $\langle i, \text{offset} \rangle$.

Osservazioni 9.10.2 Vantaggi e svantaggi dell'IPT

Utilizzare una *IPT* permette di risparmiare spazio, ma può aumentare il tempo di traduzione degli indirizzi logici in fisici, in quanto:

- Per ottenere l'indirizzo fisico, è necessario scorrere la *IPT* alla ricerca della entry contenente la coppia $\langle \text{process-id}, \text{page-number} \rangle$, il che può richiedere centinaia o migliaia di accessi alla memoria principale (MP) se la *IPT* è memorizzata in RAM.
- Tuttavia, l'uso di *memorie associative* per contenere tutta o parte dell'*IPT* consente la traduzione della maggior parte degli indirizzi senza un significativo impatto sulle prestazioni.

9.11 Il supporto alla paginazione nei vecchi processori Intel

La paginazione può, in teoria, essere implementata senza supporto hardware, ma un aiuto dall'hardware è fondamentale se si vogliono evitare significative degradazioni delle prestazioni. Oltre al supporto essenziale fornito dal *Translation Lookaside Buffer (TLB)*, tutti i processori moderni offrono una gamma di facilitazioni per una gestione efficiente dei riferimenti in memoria.

Un esempio classico è rappresentato dalla famiglia dei vecchi processori Intel Pentium (ad esempio, Pentium 3 e 4).

A scelta del sistema operativo che gira sul processore, è possibile utilizzare pagine da 4 Kbyte o da 4 Mbyte. Nel caso di pagine da 4 Kbyte, il processore adotta uno schema di paginazione a due livelli. La traduzione degli indirizzi da logici a fisici avviene nel modo consueto attraverso l'unità di paginazione:

9.12 Conclusioni

// TODO: :D

10 Memoria virtuale

10.1 Introduzione

I metodi di gestione della Memoria Primaria (MP) cercano di mantenere in RAM il maggior numero possibile di processi per aumentare il livello di multiprogrammazione. Tuttavia, per una data quantità di RAM disponibile, il numero di processi che possono risiedere in MP dipende dalla loro dimensione.

Definizione 10.1.1: Memoria Virtuale (MV)

La *Memoria Virtuale (MV)* è un insieme di tecniche che permette l'esecuzione di processi in cui codice e/o dati non sono completamente caricati in Memoria Primaria. La MV funziona poiché i programmi non necessitano di essere interamente presenti in MP per poter essere eseguiti.

Esempio 10.1.1 (Esempi di utilizzo della Memoria Virtuale)

- Il codice per la gestione delle condizioni di errore potrebbe non essere mai usato durante l'esecuzione di un programma. - Array, liste e tabelle sono spesso dichiarate di dimensioni superiori a quanto effettivamente richiesto. - Alcune opzioni di programma sono raramente utilizzate. - Le librerie dinamiche vengono caricate in RAM solo se e quando effettivamente richieste.

L'idea alla base della Memoria Virtuale è la seguente:

- Carichiamo in MP solo le parti di un programma che devono effettivamente essere eseguite e solo quando è necessario.
- Carichiamo in MP solo la porzione di strutture dati che sono utilizzate in una determinata fase di esecuzione.

Osservazioni 10.1.1 Vantaggi della Memoria Virtuale

La Memoria Virtuale permette di eseguire programmi che superano la dimensione della MP. Formalmente:

- È possibile eseguire un processo che utilizza uno spazio di indirizzamento logico superiore allo spazio fisico disponibile.
- È possibile avere in esecuzione contemporaneamente più processi che, sommati, occupano più spazio della MP disponibile.
- Ne consegue un aumento della multiprogrammazione e quindi del *throughput* della CPU.
- I programmi possono iniziare l'esecuzione più velocemente, poiché non è necessario caricarli interamente in memoria primaria.

Naturalmente, vi sono anche degli *inconvenienti* legati all'uso della Memoria Virtuale:

- Si genera un aumento del traffico tra la RAM e l'hard disk.
- L'esecuzione di un singolo programma potrebbe richiedere più tempo rispetto a uno scenario senza MV.
- In situazioni particolari, le prestazioni complessive del sistema possono degradare drasticamente, fenomeno noto come *thrashing*.

10.2 Paginazione su richiesta (Demand Paging)

L'idea di base della Memoria Virtuale è quella di *portare una pagina in MP solo nel momento del primo indirizzamento di una locazione* (un dato, un'istruzione) appartenente alla pagina stessa. Quando la CPU esegue un'istruzione che indirizza una locazione di RAM in una pagina diversa da quella contenente l'istruzione in esecuzione, e la pagina non è in MP, si dice che il processo ha generato un *page fault*. In questo caso, il Sistema Operativo (SO) deve:

1. sospendere il processo,
2. portare in memoria la pagina mancante,
3. una volta disponibile, riprendere l'esecuzione del processo dal punto in cui era stato interrotto.

Passaggi per la gestione del page fault

Più in dettaglio, quando manca la pagina riferita:

- Il processo viene tolto dalla CPU e messo in uno stato di *waiting for page*.
- Un modulo del SO detto *pager* inizia il caricamento della pagina mancante dalla Memoria Secondaria (MS) in un frame libero della Memoria Primaria (MP).
- Nel frattempo, la CPU viene assegnata a un altro processo.
- Quando la pagina è caricata in MP, il processo corrispondente viene rimesso in coda di *Ready*: riprenderà l'esecuzione dall'istruzione che aveva causato il problema quando sarà scelto dallo scheduler.

Note:-

Vedere Code di Scheduling, nel diagramma di accodamento del capitolo 3, il caso wait for an interrpt lo possiamo anche associare al page fault. **??**

Come viene rilevata l'assenza di una pagina in MP

La CPU determina la presenza di una pagina in RAM attraverso un *bit di validità* associato a ogni entry della Page Table (PT). Questo bit indica se la pagina associata è effettivamente in MP:

- Se si tenta di accedere a una pagina non in MP, il suo bit di validità sarà impostato a 0, e viene generata una *trap* detta *page fault*, attivando il meccanismo descritto.
- Quando la pagina viene caricata in MP, il bit di validità viene impostato a 1 e la PT aggiornata; a questo punto, il processo può riprendere dall'istruzione che aveva causato il page fault.

Pure Demand Paging

Un processo può persino iniziare senza alcuna delle sue pagine in MP. Al primo indirizzamento da parte del Program Counter (PC), che è inizializzato dal SO, si genera un *page fault* perché il PC punta a una pagina non in MP del processo. Questo approccio è detto *Pure Demand Paging*.

Note:-

In alternativa, il SO può caricare in MP almeno la pagina contenente la prima istruzione da eseguire.

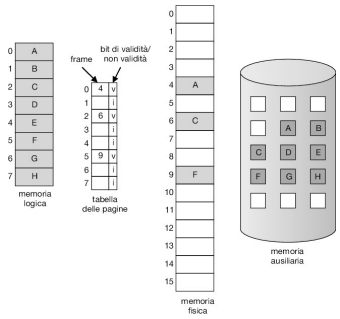


Figure 10.1: table$_o f_p ages_n otAllInRam.png$

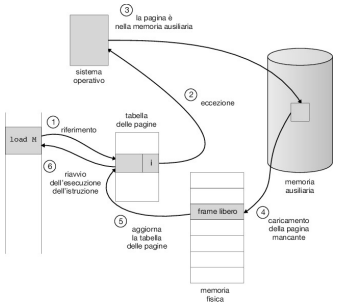


Figure 10.2: page$_f ault_g estion.png$

10.3 Demand Paging

Note:-

Un processo può essere avviato senza che alcuna delle sue pagine sia inizialmente in Memoria Primaria (MP). Alla prima istruzione indirizzata dal Program Counter (PC), inizializzato dal Sistema Operativo (SO), si genera un *page fault*, poiché il PC punta a un indirizzo di una pagina del processo non presente in MP. Questo schema è chiamato *Pure Demand Paging*.

In alternativa, il SO può caricare preventivamente in MP almeno la pagina contenente la prima istruzione da eseguire.

Supporto Hardware per la Memoria Virtuale

Note:-

Per implementare la memoria virtuale è necessario un supporto hardware specifico:

- La tabella delle pagine deve includere un *bit di validità* che l'hardware può testare per generare il *page fault*.
- Le istruzioni devono essere ri-eseguibili in caso di *page fault* oppure, alternativamente, l'hardware della CPU deve controllare la presenza in MP di tutti gli operandi prima di eseguire l'istruzione.

Note:-

Sebbene la paginazione possa essere aggiunta a qualsiasi sistema, la *paginazione su richiesta* e, più in generale, la *memoria virtuale*, richiedono un supporto hardware specifico.

Tempo di Accesso Effettivo

Supponiamo di dover leggere un dato in MP:

- ma = tempo di accesso in MP se il dato è presente (es. 100-200 nanosecondi)

- p = probabilità di un page fault
- eat (effective access time) =

$$\text{eat} = [(1-p) \times ma] + [p \times \text{tempo di gestione del page fault}]$$

Osservazioni 10.3.1 Passi per la gestione di un page fault

L'elenco completo dei passi necessari a gestire un page fault è descritto nel testo, ma le tre operazioni principali sono:

1. **Gestione del page fault**: richiede da 1 a 100 sec.
2. **Recupero della pagina mancante dalla Memoria Secondaria (MS)**: circa 8 millisecondi. Questo valore può variare a seconda del sistema, ma, se la memoria secondaria è su Hard Disk, l'ordine di grandezza è comunque di qualche millisecondo.
3. **Riavvio del processo**: richiede da 1 a 100 sec.

Note:-

I tempi per i punti 1 e 3 sono trascurabili rispetto al punto 2.

Calcolo del Tempo di Accesso Effettivo

Assumendo $ma = 200$ nanosecondi, otteniamo (valori espressi in nanosecondi):

$$\text{eat} = (1-p) \times 200 + p \times 8.000.000 = 200 + p \times 7.999.800$$

Esempio 10.3.1 (Calcolo con probabilità di page fault)

Se consideriamo $p = 0.001$ (un page fault ogni 1000 accessi), allora:

$$\text{eat} = 200 + 0.001 \times 7.999.800 = 8.199,8 \text{ (circa 8,2 microsecondi)}$$

Note:-

In questo caso, l'esecuzione rallenta di oltre 40 volte!

Domanda 10.1: Degrado massimo del 10%

Per avere un degrado massimo del 10%, la probabilità p deve rispettare la seguente disuguaglianza:

$$\text{eat} = 220 > 200 + 8 \times 10^6 \times p$$

$$20 > 8 \times 10^6 \times p$$

$$p < 2,5 \times 10^{-6}$$

Questo significa che non ci dovrebbero essere più di un page fault ogni 400.000 accessi in memoria.

Domanda 10.2: 400.000 accessi: sono molti o pochi?

Se un processo esegue circa un milione di istruzioni, quanti accessi in RAM genera approssimativamente?

Note:-

1 milione di istruzioni = 1 milione di accessi in RAM (ogni istruzione richiede almeno un accesso in RAM). Alcune istruzioni possono richiedere più accessi in RAM, ma 1 milione è un buon punto di partenza.

Note:-

Il numero di page fault deve essere estremamente basso per evitare un aumento inaccettabile del tempo medio di esecuzione dei processi. Se il numero di page fault è elevato, il *throughput* del sistema peggiora invece di migliorare.

Note:-

È possibile intervenire anche sul tempo di gestione del page fault. Ad esempio:

- Usare pagine di grandi dimensioni, che possono ridurre il numero medio di page fault. (Perché?);
- Ottimizzare l'accesso alla Memoria Secondaria, anche se questa strategia presenta limiti, poiché implica comunque l'uso di hard disk. Si veda anche il capitolo 11, che tratta delle memorie a stato solido.

10.3.1 L'area di swap

Definizione 10.3.1: Memoria Virtuale e Area di Swap

Per funzionare, la memoria virtuale necessita di una porzione dedicata dell'Hard Disk, detta *area di swap*.

Note:-

Al momento dell'installazione del sistema operativo (SO), viene riservata una porzione del disco come area di swap ad uso esclusivo del SO. Questa area è gestita con meccanismi più semplici ed efficienti rispetto a quelli del file system:

- Le pagine dei processi non vengono salvate all'interno di file, evitando così l'uso dei file descriptor;
- Spesso, vengono utilizzati blocchi più grandi, con allocazione contigua, per migliorare le prestazioni di accesso (questo concetto sarà più chiaro nella parte sulla gestione della memoria secondaria).

Osservazioni 10.3.2 Utilizzo dell'Area di Swap

Un modo semplice di usare l'area di swap consiste nel copiare l'eseguibile intero di un processo nell'area di swap all'avvio del processo:

- Il tempo di avvio del processo aumenta;
- Sono necessarie aree di swap di grandi dimensioni;
- Tuttavia, la gestione dei page fault è più veloce, poiché il recupero delle pagine è più rapido una volta che queste sono già nell'area di swap, senza passare attraverso il file system.

Osservazioni 10.3.3 Uso Alternativo dell'Area di Swap

In alternativa, le pagine dell'eseguibile o di eventuali file di dati possono essere lette direttamente dal file system:

- Utile quando occorre limitare le dimensioni dell'area di swap;
- L'avvio dei processi è più veloce;
- L'esecuzione può risultare più lenta rispetto all'uso della swap.

Notiamo che l'area di swap sembra meno utile se viene utilizzata solo per ospitare gli eseguibili e gli eventuali dati in input prima di avviare i processi. Funzione Principale dell'Area di Swap L'area di swap viene utilizzata soprattutto per:

- Liberare spazio in memoria primaria (MP) per ospitare pagine mancanti, caricate in RAM in risposta a un page fault.

Note:-

In effetti, se ci fosse sempre un frame libero, una pagina verrebbe caricata in MP solo la prima volta che viene indirizzata. Tuttavia, l'idea principale della memoria virtuale è di:

- Permettere l'esecuzione di un processo più grande della memoria primaria disponibile.
- Consentire l'esecuzione simultanea di processi che, nel complesso, richiedono più spazio di quello disponibile in RAM.

Si consideri la situazione in cui due processi occupano più spazio di quello disponibile in memoria principale (cfr. Fig. 10.9). Dunque, se si verifica un page fault e tutti i frame della RAM sono occupati, occorre liberarne

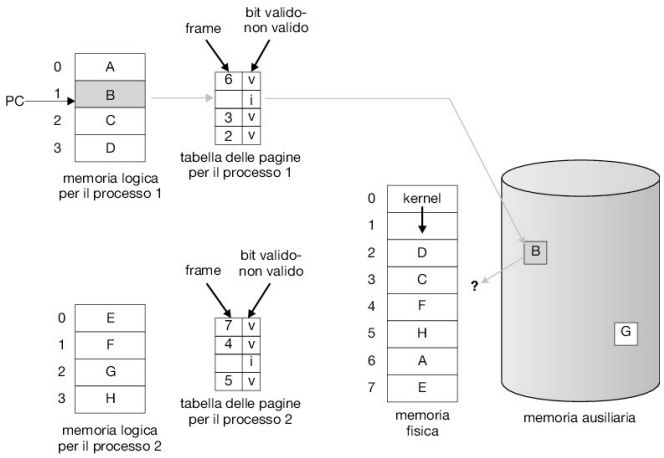


Figure 10.3: Swap 10.9

uno rimuovendo la agina che ospita, che prende il nome di **pagina vittima**.
Se la pagina vittima contiene dati modificati o fa parte dello stack o della heap di un processo, la pagina va salvata nell'area di swap, in modo che possa essere recuperata quando il processo a cui appartiene vi farà riferimento.
Le pagine di codice non devono essere salvate, tanto ce n'è comunque una copia nel file system, ma se erano state copiate inizialmente nell'area di swap potranno poi essere recuperate più velocemente se riferite di nuovo.

10.4 Sostituzione delle pagine

Domanda 10.3: Cosa succede se si verifica un page fault e non c'è alcun frame libero in RAM?

In questo caso, il sistema operativo esegue i seguenti passi:

- Seleziona una pagina "vittima" da rimuovere.
- Salva la pagina vittima nell'area di swap (se necessario).
- Carica la pagina mancante nel frame liberato.

Note:-

Questa procedura è simile al concetto di *swapping* di processi interi. Tuttavia, con la memoria virtuale, il sistema sposta tra RAM e hard disk solo frammenti di processo, ossia una pagina alla volta.

Note:-

Se la pagina vittima non è stata modificata da quando è stata caricata in RAM, si può evitare di salvarla nuovamente su disco, poiché ne esiste già una copia in memoria secondaria.

Se la pagina vittima è stata modificata, il tempo di gestione del page fault raddoppia poiché è necessario sia salvarla che caricare la nuova pagina. Un *dirty bit* associato a ciascuna entry della tabella delle pagine (PT)

può semplificare il processo: il *dirty bit* viene settato a 1 dall'hardware della CPU la prima volta che la pagina viene modificata in RAM. In questo modo, solo le pagine vittima con il *dirty bit* a 1 devono essere salvate in memoria secondaria.

Osservazioni 10.4.1 Modifica e salvataggio delle pagine in memoria virtuale

Solo le pagine di dati (stack e heap) possono essere modificate e, quindi, avere il *dirty bit* a 1. Le pagine di codice, essendo accedute solo in lettura, non necessitano di essere salvate nell'area di swap se scelte come pagine vittima. Inoltre, se il codice era stato inizialmente copiato nell'area di swap, riportare le pagine di codice in RAM è più rapido rispetto al caricamento dall'eseguibile nel file system.

Note:-

La gestione della sostituzione delle pagine è fondamentale per consentire l'esecuzione di programmi più grandi della memoria primaria disponibile. Tuttavia, ciò comporta due problemi rilevanti:

- **Scelta della pagina da sostituire**: Quale pagina scegliere come vittima?
- **Allocazione dei frame**: Quanti frame assegnare a ciascun processo? (aspetto che verrà discusso in seguito)

Il metodo utilizzato per risolvere questi problemi influisce notevolmente sulle prestazioni di esecuzione dei processi.

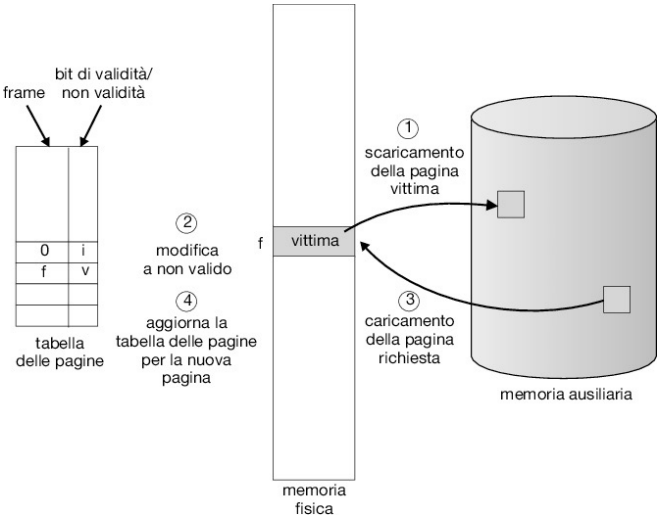


Figure 10.4: sostituzione$_d i_p agina.png$

10.4.1 Algoritmi di sostituzione delle pagine

Se una pagina vittima appena rimossa viene nuovamente indirizzata dal processo a cui appartiene, si verifica un *page fault* e la pagina deve essere ricaricata in memoria principale. In questo caso, parte del lavoro viene sprecato. Al contrario, se la pagina vittima scelta non verrà mai più utilizzata, non sarà necessario ricaricarla in memoria, ottimizzando l'uso della memoria primaria.

Osservazioni 10.4.2 Obiettivo degli algoritmi di sostituzione delle pagine

Un buon algoritmo di sostituzione minimizza il numero di *page fault*. In letteratura, questi algoritmi sono spesso chiamati *algoritmi di rimpiazzamento delle pagine*.

Domanda 10.4: Come possiamo valutare l'efficacia di diversi algoritmi di sostituzione delle pagine?

Si può valutare l'efficacia tramite sequenze di riferimenti in memoria principale:

- **generate casualmente**, oppure
- **generate dall'esecuzione di programmi reali**.

Non interessa l'indirizzo esatto dell'istruzione, ma solo il numero della pagina indirizzata, ignorando quindi l'offset.

Esempio 10.4.1 (Esempio di sequenza di riferimento)

Supponiamo che la sequenza di riferimento sia:

$$10, 7, 4, 5, 6, 1, 10, 4, \ldots$$

Durante l'esecuzione, la CPU ha generato una sequenza di indirizzi logici che indirizzano le pagine in quest'ordine.

Domanda 10.5: Quanti *page fault* genera questa sequenza di riferimento?

Supponiamo di avere a disposizione un solo frame in memoria. Con questa ipotesi, la sequenza provoca 8 *page fault*.
Consideriamo invece la seguente sequenza:

$$10, 7, 7, 7, 4, 5, 5, 5, 5, 6, 1, 10, 10, 10, 10, 4$$

Anche questa sequenza causa 8 *page fault*, poiché i riferimenti consecutivi alla stessa pagina non generano nuovi *page fault* dopo che la pagina è stata caricata in memoria primaria.

Note:-

Pertanto, le sequenze:

$$10, 7, 4, 5, 6, 1, 10, 4$$

e

$$10, 7, 7, 7, 4, 5, 5, 5, 5, 6, 1, 10, 10, 10, 10, 4$$

sono equivalenti per quanto riguarda la valutazione della bontà di un algoritmo di sostituzione.

Note:-

Il numero di *page fault* generati da una sequenza per un dato algoritmo di sostituzione dipende anche dal numero di frame disponibili. Intuitivamente, all'aumentare dei frame disponibili, il numero di *page fault* tende a ridursi — ma come vedremo, ciò non accade sempre!

10.4.2 Sostituzione delle pagine secondo l'ordine d'arrivo (FIFO)

Nel seguito, assumeremo che a ogni processo venga assegnato un numero prestabilito di frame e che la scelta della pagina vittima avvenga esclusivamente tra le pagine del processo stesso. Questa politica prende il nome di sostituzione **locale** delle pagine.
Assumiamo inoltre uno schema di *paginazione su richiesta puro*: quando un processo inizia, nessuna delle sue pagine è in RAM, e quindi il primo riferimento a una pagina qualsiasi genera un *page fault*.

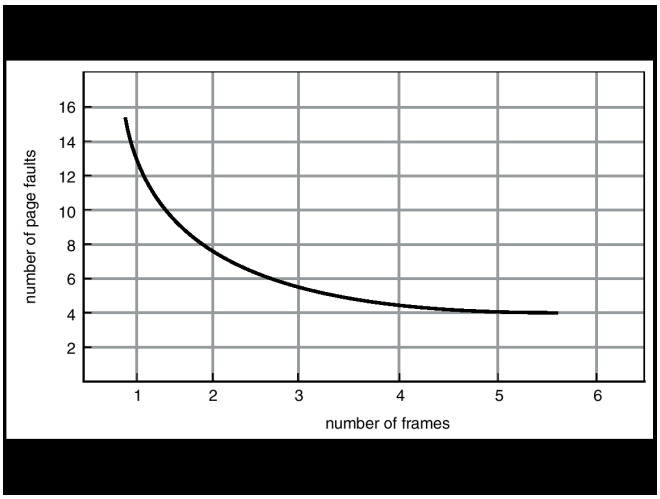


Figure 10.5: page$_f$*aultGraph.png*

Definizione 10.4.1: Algoritmo FIFO

L'algoritmo **FIFO** (First In, First Out) sceglie come pagina vittima quella presente da più tempo in memoria principale. Questo approccio è semplice da implementare ma non garantisce sempre buone prestazioni:

- Se la pagina vittima contiene codice di inizializzazione usato solo all'inizio, allora la rimozione va bene, poiché la pagina non verrà più utilizzata.
- Tuttavia, se la pagina contiene una variabile usata per tutta l'esecuzione del codice o una procedura richiamata frequentemente, la rimozione potrebbe rivelarsi inefficiente.

Esempio 10.4.2 (Esempio di FIFO)

Consideriamo la sequenza di riferimenti:

$$7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1$$

Con 3 frame disponibili, questa sequenza produce 15 *page fault* (vedi Fig. 10.12).

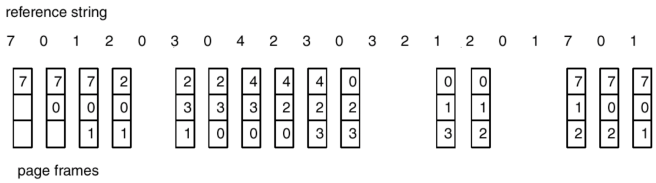


Figure 10.6: 10.12

Note:-

L'algoritmo FIFO è affetto dalla cosiddetta *Anomalia di Belady*: in alcuni casi, aumentando il numero di frame, il numero di *page fault* può paradossalmente aumentare!

Esempio 10.4.3 (Anomalia di Belady)

Consideriamo la seguente stringa di riferimento:

$$1, 2, 3, 4, 1, 2, 5, 1, 2, 3, 4, 5$$

In alcuni casi, aumentando il numero di frame, il numero di *page fault* prodotti può aumentare (vedi Fig. 10.13). È importante notare che questo fenomeno si verifica solo per alcune specifiche sequenze di riferimento.

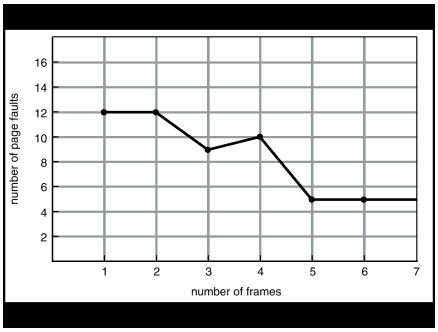


Figure 10.7: 10.13

Definizione 10.4.2: Algoritmo OPT (o MIN)

L'algoritmo **OPT**, o **MIN**, seleziona come vittima la pagina che verrà utilizzata più avanti nel tempo rispetto a tutte le altre pagine attualmente in memoria. Questo algoritmo garantisce il numero minimo di *page fault* per un dato numero di frame, evitando l'anomalia di Belady.
Tuttavia, **OPT non è implementabile** in un sistema reale, poiché richiederebbe una conoscenza anticipata dell'uso delle pagine. Viene usato solo come termine di paragone per valutare le prestazioni di altri algoritmi.

Esempio 10.4.4 (Esempio di OPT)
Utilizzando l'algoritmo OPT con 3 frame, la sequenza:

$$7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1$$

produce 9 *page fault* (vedi Fig. 10.14).

10.4.3 Algoritmo LRU (Least Recently Used)

Note:-
L'algoritmo OPT è ideale perché sceglie come pagina vittima quella che non verrà più utilizzata nel futuro, ma ovviamente non è implementabile. Un tentativo di approssimazione di questo comportamento è l'algoritmo **LRU** (Least Recently Used), che seleziona come vittima la pagina che non è stata usata da più tempo.

Definizione 10.4.3: Algoritmo LRU

L'algoritmo **LRU** non soffre dell'anomalia di Belady e si avvicina molto a OPT, ma ha il difetto di essere difficile da implementare in modo efficiente. In pratica, LRU guarda al passato (anziché al futuro) per determinare quale pagina rimuovere. Questo approccio funziona bene nella maggior parte dei casi, ma è più complesso da implementare.

Esempio 10.4.5 (Esempio di LRU)
Consideriamo la seguente sequenza di riferimenti:

$$7, 0, 1, 2, 0, 3, 0, 4, 2, 3, 0, 3, 2, 1, 2, 0, 1, 7, 0, 1$$

Con 3 frame disponibili, l'algoritmo LRU produce 12 *page fault* (vedi Fig. 10.15).

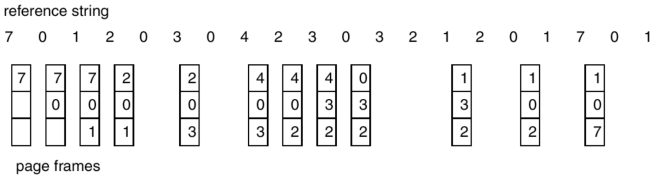


Figure 10.8: 10.15

L'implementazione dell'algoritmo LRU richiederebbe un supporto hardware da parte della CPU che, purtroppo, non è disponibile nelle architetture moderne. Tuttavia, esistono approcci che utilizzano un supporto hardware più semplice per approssimare LRU in modo accettabile.

Definizione 10.4.4: Reference Bit

Molti processori forniscono un *reference bit*, un bit associato a ciascuna pagina nella Page Table di un processo. Quando un processo inizia, tutti i reference bit delle sue pagine sono inizializzati a 0 dal sistema.

- Quando una pagina viene indirizzata (sia in lettura che in scrittura), l'hardware imposta a 1 il reference bit di quella pagina.
- In questo modo, possiamo sapere quali pagine sono state utilizzate di recente, anche se non sappiamo l'ordine esatto di accesso.

Algoritmo LRU Seconda Chance

Partendo da un algoritmo FIFO, in caso di *page fault* il sistema operativo esamina la pagina che è entrata per prima in RAM. Se il reference bit di questa pagina è 0, essa viene scelta come pagina vittima. Se invece il reference bit è 1, la pagina riceve una "seconda chance": il suo reference bit viene azzerato e viene trattata come se fosse appena entrata in memoria.

Definizione 10.4.5: Algoritmo della Seconda Chance

L'algoritmo della seconda chance funziona come segue:

- Se il reference bit di una pagina è 0, la pagina viene selezionata come vittima.
- Se il reference bit è 1, il bit viene azzerato e la pagina viene spostata in fondo alla coda FIFO.

Se in una sequenza di sostituzioni tutte le pagine hanno il reference bit impostato a 1, l'algoritmo riprende a esaminare la prima pagina della coda (quella entrata per prima in RAM) e la seleziona come vittima, tornando di fatto a un algoritmo FIFO.

Se una pagina viene riferita frequentemente, il suo reference bit rimarrà a 1 per la maggior parte del tempo, riducendo la probabilità che venga selezionata come vittima. Questo algoritmo è una buona approssimazione di LRU ed è decisamente più efficiente di una gestione completa di LRU in hardware.
Se "next victim" non viene riferita prima di una seconda chiamata dell'algoritmo, il suo reference bit resta a 0, quindi non è stata riferita di recente, e diviene quindi una buona candidata alla sostituzione.

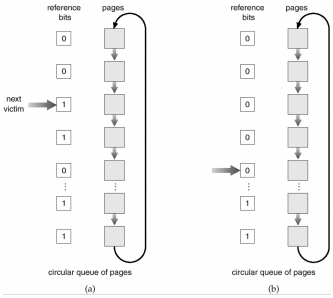


Figure 10.9: 10.17

10.4.4 Algoritmo Seconda Chance con Dirty Bit

Quando l'hardware fornisce sia il *reference bit* che il *dirty bit*, le pagine possono essere classificate in quattro gruppi, ognuno dei quali ha una priorità diversa per essere sostituito:

Definizione 10.4.6: Classificazione delle Pagine

Le pagine sono raggruppate in base a due bit: il *reference bit* e il *dirty bit*.

- **(0, 0)**: La pagina non è stata utilizzata di recente e non è stata modificata. È la migliore da rimpiazzare.
- **(1, 0)**: La pagina è stata utilizzata di recente, ma non è stata modificata. È una buona candidata, ma meno preferibile rispetto a quella (0, 0).
- **(0, 1)**: La pagina non è stata utilizzata di recente, ma è stata modificata. È meno buona da sostituire, in quanto dovrà essere salvata in memoria secondaria per non perdere le modifiche.
- **(1, 1)**: La pagina è stata utilizzata di recente e modificata. È la peggiore da sostituire, poiché richiede di essere salvata in memoria secondaria prima di essere rimossa.

Il principio di sostituzione applicato in questo algoritmo è simile a quello della "Seconda Chance". In questo caso, si seleziona come vittima la prima pagina che appartiene alla classe migliore non vuota (quella con il valore **(0, 0)** se disponibile, poi **(1, 0)**, e così via).

Definizione 10.4.7: Algoritmo di Sostituzione con Reference e Dirty Bit

L'algoritmo, che si basa su questa classificazione delle pagine, è utilizzato in molti sistemi operativi Unix e in macOS. La scelta della pagina da rimpiazzare dipende dallo stato combinato dei bit di riferimento e di modifica, con la preferenza per le pagine meno utilizzate e non modificate.

10.5 Allocazione dei Frame

In un sistema multiprogrammato, la distribuzione dei frame disponibili tra i processi può essere gestita in vari modi:

Definizione 10.5.1: Strategie di Distribuzione dei Frame

- **Allocazione uniforme**: Ogni processo riceve lo stesso numero di frame. Ad esempio, se ci sono n frame e p processi, ogni processo ottiene n/p frame.
- **Allocazione proporzionale**: I frame vengono distribuiti in base alle dimensioni di ogni processo. Ad esempio, se i processi sono di dimensioni diverse, i processi più grandi riceveranno più frame.
- **Allocazione proporzionale in base alla priorità**: La distribuzione dei frame tiene conto della priorità dei processi. I processi con priorità più alta ricevono più frame.
- **Allocazione con riserva di frame**: Alcuni frame devono essere tenuti liberi per consentire l'ingresso di nuovi processi nel sistema.

Definizione 10.5.2: Esempio di Allocazione Proporzionale

Se abbiamo 11 frame disponibili e i processi P1, P2 e P3 hanno dimensioni rispettive di 4, 6 e 12 pagine, la distribuzione dei frame sarà:

- P1: 2 frame,
- P2: 3 frame,
- P3: 6 frame.

Nel caso dell'allocazione proporzionale in base alla priorità, il numero di frame assegnati a ciascun processo dipenderà dalla priorità relativa dei processi (e in alcuni casi dalle loro dimensioni).

Definizione 10.5.3: Allocazione dei Frame e Grado di Multiprogrammazione

Note:-

Qualunque schema di allocazione venga scelto, il numero di frame assegnato a ciascun processo cambierà in funzione del grado di multiprogrammazione.

Definizione 10.5.4: Strategie di Selezione della Vittima per la Rimozione

In caso di page fault, bisogna scegliere quale pagina rimuovere dalla memoria principale:

- **Allocazione globale**: La vittima è scelta fra tutte le pagine in memoria principale, esclusi i frame utilizzati dal sistema operativo. Questo approccio potrebbe rimuovere una pagina di un altro processo rispetto a quello che ha causato il page fault.
- **Allocazione locale**: La vittima è scelta fra le pagine del processo che ha causato il page fault, mantenendo costante il numero di frame allocato a ciascun processo.

Concetto sbagliato 10.1: Problemi

Problemi dell'Allocazione Globale: La strategia globale rende il turnaround di un processo fortemente dipendente dal comportamento degli altri processi, con un elevato rischio di variazione nelle prestazioni da un'esecuzione all'altra.
Problemi dell'Allocazione Locale: Se un processo riceve troppi frame, ciò può ridurre il throughput complessivo del sistema, poiché gli altri processi avranno meno frame disponibili e genereranno più page fault.

Note:-

Si è visto sperimentalmente che l'allocazione globale fornisce in genere un throughput maggiore e riesce a gestire la multiprogrammazione in maniera più flessibile.
L'allocazione globale è di solito preferita per sistemi time sharing, in cui molti utenti possono usare contemporaneamente il sistema.
I sistemi Windows usano l'allocazione locale, mentre Linux e Solaris usano l'allocazione globale delle pagine.

10.5.1 Thrashing (attività di paginazione degenere)

Note:-

Consideriamo un sistema in cui, in un dato momento, ogni processo ha a disposizione un numero ridotto di frame, cioè ogni processo ha in memoria principale (MP) un numero di pagine inferiore rispetto al numero totale di pagine di cui è composto. Supponiamo anche di adottare una allocazione globale dei frame. Va sottolineato che il problema del **thrashing** si può verificare anche nel caso di allocazione locale dei frame.

Definizione 10.5.5: Probabilità di Page Fault

Con poche pagine in RAM per processo, la probabilità che ogni processo generi un page fault è alta. In seguito a un page fault, una pagina vittima viene rimossa dalla memoria principale (MP), probabilmente da un altro processo.
Questo altro processo, a sua volta, avrà ancora meno pagine in memoria, aumentando ulteriormente la probabilità che anche esso generi un page fault a breve. Si innesca così un circolo vizioso in cui:

- Ogni processo genera continuamente page fault,
- I frame vengono "rubati" tra i vari processi,
- La probabilità di page fault aumenta continuamente.

Definizione 10.5.6: Thrashing

Questo fenomeno è noto come **thrashing**, che si verifica quando il sistema tenta di aumentare eccessivamente il grado di multiprogrammazione, ossia cercando di eseguire più processi contemporaneamente per sfruttare al massimo la CPU e incrementare il throughput del sistema.
Tuttavia, oltre una certa soglia, i processi passano più tempo a gestire i page fault generati che ad eseguire realmente il loro lavoro. Di conseguenza, la *utilizzazione della CPU* diminuisce drasticamente e il *throughput* del sistema crolla.

Definizione 10.5.7: Relazione tra Grado di Multiprogrammazione e Thrashing

Il grado di multiprogrammazione è direttamente legato al rischio di thrashing. Aumentando il numero di processi che cercano di essere eseguiti contemporaneamente, cresce anche la probabilità che ognuno di essi abbia pochi frame disponibili. Questo può portare a una situazione in cui il sistema è sopraffatto dalla gestione dei page fault, riducendo di fatto l'efficienza complessiva.
La relazione tra grado di multiprogrammazione e throughput diventa inversamente proporzionale oltre una certa soglia, con l'aumento della multiprogrammazione che inizialmente porta a un miglioramento delle prestazioni, ma che successivamente provoca un forte degrado delle stesse a causa del thrashing.

10.5.2 Cause del Thrashing

Se il livello di utilizzo della CPU di un sistema è troppo basso, lo si può alzare aumentando in grado di permettendo a più utenti di connettersi, e/o di lanciare un maggior numero di processi. In questo modo però, i nuovi processi incominciano a sottrarre pagine ai processi già presenti, per "farsi un po' di spazio". Fino ad un certo punto l'aumento di processi è ben tollerato dal sistema, poiché ciascun processo ha comunque una quantità sufficiente di frame a disposizione da poter girare senza generare troppi page fault. Ma se si esagera, ci si può

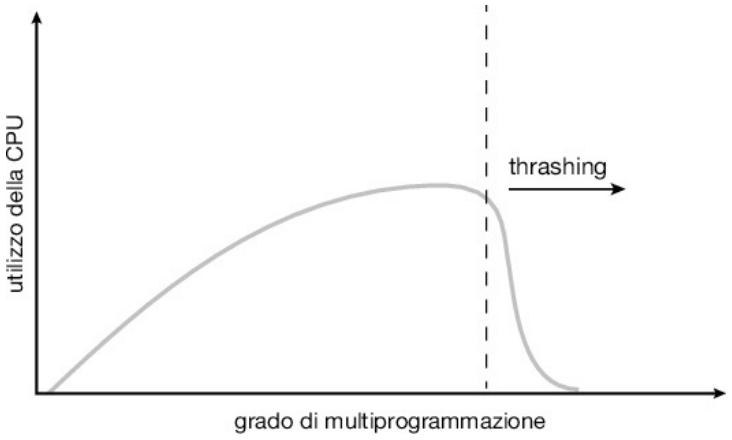


Figure 10.10: 10.20

avvicinare alla soglia del thrashing: molti processi incominciano a generare dei page fault e, come conseguenza, vengono tolti dalla RQ e messi in una coda di wait in attesa della pagina mancante. Risultato? La RQ si svuota, e il livello di utilizzo della CPU scende. Beh, ma se la CPU è sotto-utilizzata, si può lanciare qualche altro processo, o permettere a qualche altro utente di collegarsi... E la situazione non fa che peggiorare!.
Nei moderni sistemi time-sharing, il ciclo perverso è spesso innescato dagli utenti, che lanciano altri programmi senza attendere la fine di quelli già in esecuzione, sperando così di aumentare la percentuale del tempo di CPU globale che riescono ad usare a loro vantaggio.
In definitiva quindi, il thrashing è una sorta di "ingolfamento" del sistema: vogliamo sfruttarlo al meglio "iniettando" più e più processi nel sistema, fino ad arrivare ad un punto in cui i processi si ostacolano a vicenda.
La soluzione giusta sarebbe di diminuire il grado di multiprogrammazione temporaneamente, in modo che i processi non rimossi dalla MP abbiano il tempo di terminare correttamente prima di far (ri)partire gli altri.

10.5.3 Come combattere il Thrashing

In definitiva, quindi, il *thrashing* rappresenta una sorta di "ingolfamento" del sistema: si tenta di sfruttare al massimo le risorse "iniettando" più processi possibili, ma si finisce per raggiungere un punto in cui i processi si ostacolano a vicenda.

Note:-

La soluzione ottimale consiste nel ridurre temporaneamente il grado di multiprogrammazione. Così facendo, i processi ancora presenti in memoria possono completare correttamente la loro esecuzione, prima di permettere l'avvio (o il riavvio) di altri processi.

Osservazioni 10.5.1 Frequenza accettabile dei page fault

È possibile stabilire, ad esempio in base ad osservazioni sperimentali, un livello "accettabile" di frequenza di page fault, per raggiungere le prestazioni desiderate:

- **Se la frequenza osservata è troppo bassa**, si possono rimuovere alcuni frame dai processi e aumentare il grado di multiprogrammazione.
- **Se la frequenza osservata è troppo alta**, occorre diminuire il grado di multiprogrammazione e redistribuire i frame liberati tra i processi ancora attivi.

Note:-

Il *thrashing* può essere prevenuto monitorando attentamente la frequenza dei page fault (come illustrato in figura 10.23).

Adottare una politica di sostituzione locale può contribuire a ridurre il rischio di thrashing, poiché i processi

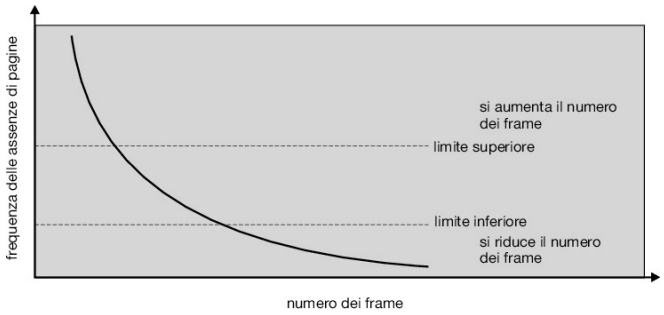


Figure 10.11: frequenctOfPageFault.png

non possono sottrarsi le pagine tra loro. Se un singolo processo entra in thrashing, non danneggia gli altri, sebbene possa utilizzare intensivamente le risorse di I/O su disco.

Tuttavia, se si assegnano troppi pochi frame a ciascun processo per aumentare eccessivamente il grado di multiprogrammazione, esiste il rischio che tutti i processi entrino in thrashing.

Corollario 10.5.1 Prevenzione ottimale del thrashing

La migliore prevenzione del thrashing è dotare il sistema di una quantità sufficiente di memoria principale.

10.6 Dimensioni delle pagine

Sono sempre potenze di 2; ma quanto dovrebbero essere grandi?

- **Pagine piccole** implicano:
  - Tabelle delle pagine (PT) più grandi;
  - meno frammentazione interna;
  - peggiori prestazioni nell'uso dell'HD, poiché il *seek* e la latenza del disco sono costanti (si vedrà meglio nel Capitolo 11);
  - in generale, un maggior numero di page fault (immaginiamo pagine di un solo byte in pure demand paging).
- **Pagine grandi** presentano le considerazioni opposte.

A causa della diminuzione del costo della RAM e dell'aumento degli spazi di indirizzamento fisico disponibili, la tendenza è quella di usare pagine sempre più grandi. Negli anni '80, 4 Kbyte era considerata la dimensione massima accettabile di una pagina, mentre oggi tale valore è normale, e spesso anche superato.

Note:-

L'aumento costante della quantità di RAM nei sistemi moderni ha fatto sì che, nel tempo, il problema del *thrashing* e, più in generale, le problematiche legate alla memoria virtuale abbiano un impatto minore sulle prestazioni complessive dei sistemi.

10.6.1 Struttura dei programmi

Il modo in cui i programmi utilizzano i dati influisce significativamente sul numero di page fault generati.

- Ad esempio, gli **array bidimensionali** sono allocati per riga: se scriviamo codice che accede agli elementi per colonna, aumentiamo il rischio di page fault. Supponiamo che una pagina contenga esattamente una riga dell'array; cosa accade se è allocato un solo frame per contenere una parte dell'array?
- Le **tabelle hash** possono offrire prestazioni scarse con la memoria virtuale, poiché anche dati concettualmente contigui vengono memorizzati in modo sparpagliato.

Consideriamo l'array bidimensionale:

```
char A[1024][1024];
```

Ogni riga è memorizzata in una pagina, e all'array è assegnato un solo frame (di 1024 byte) in RAM. L'array è memorizzato per righe:

```
A[0][0], A[0][1], A[0][2], ..., A[0][1023], A[1][0], ...
```

- **Programma 1**:

```
for (j = 0; j < 1024; j++)
    for (i = 0; i < 1024; i++) A[i][j] = '0';
```

- **Programma 2**:

```
for (i = 0; i < 1024; i++)
    for (j = 0; j < 1024; j++) A[i][j] = '0';
```

Domanda 10.6

Quale programma genera meno page fault? E quanti?
Quale programma genera più page fault? E quanti?

Note:-

Il **Programma 1** genera più fault, perchè accede prima colonna per colonna, $(1024 * 1024)$ page fault.
Il **Programma 2** genera meno fault, perchè accede prima riga per riga, (1024) page fault.

10.7 Gestione nei Sistemi Operativi

10.7.1 Windows

In Windows 10 viene implementata la *demand paging with clustering*: quando una pagina viene caricata in memoria, vengono caricate anche alcune pagine adiacenti, che si presume possano essere usate a breve.

Alla creazione di un processo, vengono assegnati a quest'ultimo due numeri:

- **Insieme di lavoro minimo**: il numero minimo di pagine che il sistema operativo garantisce di allocare in RAM per quel processo (di solito, 50);
- **Insieme di lavoro massimo**: il numero massimo di pagine che il sistema operativo allocherà in RAM per quel processo (di solito, 345).

Il sistema operativo mantiene anche una lista di **frame liberi**, con un numero minimo di frame da mantenere liberi in lista.

- Se un processo P genera un page fault e non ha ancora raggiunto il suo insieme di lavoro massimo, la pagina mancante viene portata in RAM e assegnata a un frame libero.
- Se invece P ha raggiunto il suo insieme di lavoro massimo, viene scelta una **pagina vittima** tra quelle di P, quindi viene applicata una politica di sostituzione locale delle pagine.

Se il numero di frame liberi in RAM scende al di sotto del limite minimo, viene avviata una procedura per liberare spazio:

- Ciascun processo che ha in RAM un numero di pagine superiore al proprio insieme di lavoro minimo vedrà rimosse dalla RAM tutte le pagine in eccesso.
- Nei sistemi con processore Intel, per decidere quali pagine rimuovere viene utilizzato l'**algoritmo della seconda chance**.

10.7.2 Solaris

Note:-

Solaris utilizza una normale *paginazione su richiesta*, assegnando un frame libero in caso di page fault. Un parametro, `lostfree`, associato all'elenco dei frame liberi e pari di solito a 1/64 del numero di frame in cui è suddivisa la RAM, indica il numero minimo di frame liberi desiderati.

Ogni 1/4 di secondo, il sistema operativo controlla se il numero di frame liberi è inferiore a `lostfree`. In tal caso, viene attivato il processo **pageout**, che funziona in due fasi applicando una variante dell'algoritmo della seconda chance:

- **Prima fase**: `pageout` scorre tutte le pagine allocate in RAM azzerando il bit di riferimento di ciascuna.
- **Seconda fase**: scorre di nuovo tutte le pagine e quelle con bit di riferimento ancora a 0 vengono considerate riutilizzabili. Le pagine con *dirty bit* a 1 vengono salvate prima di essere effettivamente riutilizzate.

Se un processo accede a una pagina marcata come "riutilizzabile" e in attesa di essere salvata, la pagina viene semplicemente riassegnata a quel processo.

Il tempo tra le due scansioni effettuate da `pageout` può variare in base ai parametri del sistema operativo, ma è generalmente dell'ordine di alcuni secondi.

- Se `pageout` non riesce a mantenere la quantità di frame liberi a un livello accettabile (stabilito dai parametri di sistema), potrebbe indicare che si sta verificando il fenomeno del **thrashing**.
- In tal caso, il sistema operativo può decidere di rimuovere tutte le pagine di un processo, scegliendo tra quelli che sono rimasti inattivi per il tempo più lungo.

11 Memoria di massa

11.1 Disco Rigido

11.1.1 Struttura

Definizione 11.1.1: Hard Disk (HD)

Un HD è composto da una serie di piatti o "dischi" sovrapposti, con un diametro che varia tra i 4,5 e i 9 cm.

- Ogni piatto è suddiviso in una serie di tracce circolari concentriche.
- Ogni traccia è suddivisa in una serie di settori.
- L'insieme delle tracce posizionate nello stesso punto sui vari piatti prende il nome di *cilindro*.
- Un "braccio del disco" sostiene una testina di lettura/scrittura per ogni piatto: le testine si muovono tutte simultaneamente e si posizionano sui vari settori del piatto corrispondente (simile al braccio di un giradischi).

Note:-

I settori del disco rappresentano l'unità minima di memorizzazione delle informazioni. Storicamente, ogni settore aveva una dimensione standard di 512 byte; tuttavia, dal 2010 molti produttori hanno aumentato la dimensione fino a 4 KB per settore. Ogni settore memorizza un blocco di dati.

- I piatti dell'HD ruotano sincronicamente attorno al loro asse, raggiungendo velocità tra 5400 e 15000 RPM (*rounds per minute*), corrispondenti a circa 250 giri al secondo.
- Ogni piatto ha associata una testina di lettura/scrittura dei settori, che opera a pochi micron dalla superficie del piatto.

Domanda 11.1: Perché il tempo di accesso a un settore varia?

Una testina può leggere o scrivere su un settore solo quando questo si trova esattamente sotto la testina. Pertanto, il tempo di accesso a un settore dipende principalmente da due componenti:

- *Seek time* (tempo di posizionamento): il tempo necessario affinché la testina raggiunga la traccia contenente il settore desiderato.
- *Rotational latency* (latenza rotazionale): il tempo che occorre affinché la rotazione del piatto allinei il settore esatto sotto la testina.

Note:-
A causa della presenza di elementi meccanici, i tempi di accesso sono dell'ordine di alcuni millisecondi.

11.1.2 Mappatura degli indirizzi

Definizione 11.1.2: Modello Logico di HD

Un HD può essere logicamente visto come un array unidimensionale di blocchi logici, ciascuno di 512 (o più recentemente 4096) byte: questa è la più piccola unità di trasferimento dati.

- Ogni settore corrisponde a un blocco logico.
- L'array unidimensionale di blocchi logici viene mappato sequenzialmente nei settori del disco:
  - Il *settore 0* è il primo settore della traccia più esterna del primo piatto (solitamente in posizione superiore o inferiore nella pila dei piatti).
  - Successivamente, i settori vengono numerati consecutivamente lungo la traccia fino a raggiungere i settori delle tracce più interne. La numerazione prosegue in modo analogo nei restanti piatti.

La mappatura tra blocco logico e settore del disco risulta più complessa di quanto sembri, a causa di due fattori principali:

- **Difetti di fabbricazione:** I dischi possono avere settori difettosi. Tali settori vengono nascosti attraverso il meccanismo di mappatura, che associa blocchi logici a settori funzionanti del disco.
- **Differenze di lunghezza delle tracce:** Non tutte le tracce hanno la stessa lunghezza.
  - Le tracce più lontane dal centro del disco sono più lunghe rispetto a quelle interne e possono contenere fino al 40% di settori in più.

11.1.3 Scheduling dei dischi rigidi

Il sistema operativo (SO) riceve frequentemente richieste di accesso al disco da parte dei processi e deve ottimizzare il trasferimento dei dati per migliorare le prestazioni complessive di accesso al disco.

Note:-
Il SO non può influenzare la *latenza rotazionale* del disco, che in media corrisponde a metà del tempo necessario per completare una rotazione. Tuttavia, può ridurre il *seek time medio* complessivo ordinando in maniera strategica le richieste in coda, minimizzando così il movimento delle testine.

Algoritmi di scheduling delle richieste I/O

Esistono diversi algoritmi per gestire lo scheduling delle richieste di I/O del disco. Consideriamo come esempio la seguente sequenza di richieste di accesso, comprese tra la traccia 0 e la traccia 199:

$$\{98, 183, 37, 122, 14, 124, 65, 67\}$$

Note:-
Le tracce potrebbero trovarsi su piatti diversi, dato che tutti i piatti ruotano insieme e le testine si muovono simultaneamente. Tuttavia, per semplicità possiamo supporre l'esistenza di un unico piatto e che la testina sia inizialmente posizionata sulla traccia (o cilindro) numero 53.

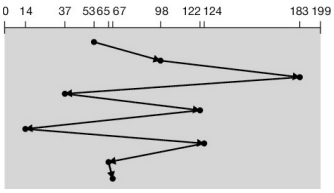


Figure 11.1: Coda delle richieste: 98, 183, 37, 122, 14, 124, 65, 67 In tutto la testina attraversa 640 tracce. Invece che "122 - 14124" era meglio fare "122 - 124 - 14"

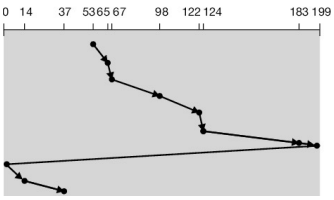


Figure 11.2: Coda delle richieste: 98, 183, 37, 122, 14, 124, 65, 67 La testina attraversa 183 tracce + 200 tracce per tornare indietro, ma questo ritorno richiede poco tempo, perché la testina non deve mai fermarsi e ripartire

C-SCAN (Circular-SCAN)

Fornisce un tempo di attesa per le varie richieste più uniforme di altri algoritmi, anche se non riesce a garantire un tempo medio di attesa minimo.
La testina si muove da un estremo all'altro del piatto, servendo le richieste.
Quando raggiunge l'estremità del piatto, torna immediatamente all'inizio senza servire richieste.
In pratica, **tratta i settori/cilindri come una lista circolare.**

Note:-
Questo è l'algoritmo di scheduling più utilizzato nei sistemi operativi moderni.

11.2 Formattazione del disco

Prima di poter essere utilizzato, un disco rigido deve essere sottoposto a un processo di *formattazione*, che avviene in due fasi principali:

- Questa operazione viene solitamente effettuata dal costruttore dell'HD e ha lo scopo di:
  - Associare un numero univoco a ogni settore.
  - Allocare uno spazio per un codice di correzione degli errori (ECC), utilizzato durante ogni operazione di I/O su quel settore.
- Durante questa fase è possibile definire la dimensione dei blocchi fisici, ad esempio 512 byte o 4096 byte per settore.

Formattazione logica

- Questo processo, gestito dal sistema operativo, è necessario per creare e organizzare il File System.
- Il sistema operativo esegue le seguenti operazioni:
  - Creazione della lista dei blocchi liberi secondo lo schema adottato.
  - Creazione di una directory iniziale, punto di partenza per l'intera struttura del File System.
  - Riservazione di aree specifiche per la gestione diretta da parte del SO:
    - **Boot Block:** il blocco di avviamento.
    - **Aree per gli attributi dei file:** ad esempio, gli *index-node* in Unix o la MFT (*Master File Table*) in Windows.

11.2.1 Il Boot Block

- Contiene il codice necessario per avviare il sistema operativo.
- All'accensione, un piccolo programma residente in ROM istruisce il *disk controller* a trasferire il contenuto del Boot Block nella RAM.
- Una volta trasferito, il controllo passa al codice del Boot Block, che avvia l'intero sistema operativo caricandolo dal disco stesso.

11.3 Gestione dell'area di SWAP

Durante la formattazione logica del disco rigido, il sistema operativo riserva uno spazio per l'*area di Swap*, che funge da memoria virtuale utilizzata per lo scambio di pagine o segmenti tra RAM e memoria secondaria.

11.3.1 Gestione dell'area di Swap

- **Swap come file:**
  - Nel caso più semplice, l'area di Swap può essere un file di grandi dimensioni all'interno del File System.
  - Nei sistemi Windows, lo *swap file* è denominato `pagefile.sys`.
  - Gli utenti possono regolarne la dimensione, ad esempio riducendola in presenza di una grande quantità di RAM, per recuperare spazio sul disco rigido.
- **Swap come partizione dedicata:**
  - Una porzione specifica del disco rigido, chiamata *partizione di Swap*, può essere riservata esclusivamente a questo scopo.
  - Questa partizione è gestita diversamente rispetto a un normale File System, adottando strategie di allocazione ottimizzate per la velocità di accesso.
  - Ad esempio, i blocchi possono essere allocati in modo contiguo per evitare la ricerca di blocchi liberi, riducendo così il tempo necessario per lo scambio.

Dimensionamento dell'area di Swap

Note:-

È fondamentale dimensionare adeguatamente l'area di Swap per garantire che il sistema operativo trovi rapidamente uno spazio libero per lo scambio di pagine e segmenti.

- **Consigli per il dimensionamento:**
  - In Solaris, si raccomanda di dimensionare l'area di Swap in base alla differenza tra lo spazio di indirizzamento logico e quello fisico.
  - In Linux, si suggerisce di utilizzare un'area di Swap pari al doppio della RAM disponibile.
- **Sistemi con più dischi:**
  - In configurazioni multi-disco, è possibile creare un'area di Swap per ciascun disco.
  - Ciò consente di sfruttare le aree di Swap in parallelo, bilanciando il carico di lavoro e migliorando le prestazioni.

11.4 Sistemi RAID

Gli hard disk (HD) e i dischi a stato solido (SSD) sono dispositivi notevolmente più lenti rispetto al processore e alla memoria primaria. Inoltre, il guasto di un disco rigido rappresenta un rischio significativo: in assenza di un back-up, i dati memorizzati possono essere irrimediabilmente persi o, nel migliore dei casi, non disponibili durante i tempi di riparazione.

11.4.1 Introduzione ai sistemi RAID

Definizione 11.4.1: RAID (Redundant Array of Independent Disks)

È un sistema di configurazione della memoria secondaria progettato per migliorare sia le prestazioni sia l'affidabilità degli hard disk.

- RAID si rivela utile in ogni settore, ma è essenziale in contesti critici dove il servizio non può mai interrompersi, come nel settore finanziario e bancario.
- Il concetto di RAID fu introdotto nel 1988 da Patterson, Gibson e Katz, con l'acronimo iniziale *Redundant Array of Inexpensive Disks*, successivamente ridefinito come *Redundant Array of Independent Disks*.
- La controparte dei sistemi RAID è rappresentata da dispositivi SLED (*Single Large Expensive Disk*).

11.4.2 Caratteristiche principali di un sistema RAID

- Un sistema RAID è costituito da un insieme di dischi (detto *disk array*) che viene visto dal sistema operativo come un singolo dispositivo di memorizzazione, più veloce e affidabile di un SLED.
- La gestione dei dischi è demandata al *controller del RAID*, che si occupa di distribuire i dati secondo criteri specifici, senza necessità di modifiche al sistema operativo.
- Questo vantaggio semplifica notevolmente la vita degli amministratori di sistema (*system administrators*).

11.4.3 Idee principali alla base di RAID

Le due idee fondamentali di un sistema RAID sono:

1. **Distribuzione dei dati:** L'informazione è suddivisa su più dischi per parallelizzare parte delle operazioni di accesso e migliorare le prestazioni.
2. **Ridondanza dei dati:** L'informazione è duplicata su più dischi. In caso di guasto di un disco, il sistema può continuare a funzionare, recuperando i dati dal disco di backup.

11.4.4 Livelli di RAID

Differenti schemi di implementazione delle idee sopra descritte hanno portato alla definizione di vari livelli RAID, numerati da 0 a 6.

Note:-

Da qui in poi, ci si discosta parzialmente dalla trattazione del libro di testo.

11.4.5 RAID di Livello 0

Note:-

I sistemi RAID di livello 0, pur essendo inclusi nella famiglia RAID, non offrono ridondanza dei dati, quindi non aumentano l'affidabilità del sistema.

Definizione 11.4.2: RAID Livello 0

Un'architettura RAID in cui il disco virtuale (cioè l'insieme di blocchi logici consecutivi visti dal sistema operativo) viene suddiviso in *strip* (strisce) di k blocchi consecutivi ciascuna. Questa tecnica, chiamata **striping**, distribuisce i dati su più dischi per migliorare le prestazioni.

- Ogni *strip* è identificata da un numero e contiene k blocchi consecutivi.
  - Lo strip 0 contiene i blocchi da 0 a $k-1$.
  - Lo strip 1 contiene i blocchi da k a $2k-1$.

- Gli strip sono distribuiti sui dischi disponibili secondo la formula:

$$\text{numero-strip} \mod \text{dischi-nel-sistema}$$

- Ad esempio, in un sistema con 4 dischi:
  - Il disco 0 conterrà gli strip 0, 4, 8, ...
  - Il disco 1 conterrà gli strip 1, 5, 9, ...
- Se $k = 1$, ogni strip contiene un singolo blocco. In questo caso:
  - Il blocco 0 sarà sul primo settore del primo disco.
  - Il blocco 1 sarà sul primo settore del secondo disco.
  - Il blocco 4 sarà sul secondo settore del primo disco, e così via.

Esempio 11.4.1 (Esempio di richiesta su un RAID Livello 0)
Supponiamo che il sistema operativo richieda la lettura di dati contenuti negli strip 4, 5, 6 e 7.

- Il controller RAID suddividerà la richiesta in quattro letture parallele, una per ciascun disco.
- Su un singolo disco SLED, invece, tutti i settori dei quattro strip dovrebbero essere letti in sequenza.

Di conseguenza, l'operazione sarà completata più rapidamente con il RAID 0.

Osservazioni 11.4.1 Prestazioni e limiti del RAID Livello 0

- **Miglioramenti delle prestazioni:** Un RAID di livello 0 è particolarmente efficiente quando le richieste coinvolgono molti strip consecutivi, e il sistema è composto da un elevato numero di dischi.
- **Limitazioni:**
  - Richieste che riguardano un singolo strip non ottengono alcun miglioramento rispetto a un disco SLED.
  - L'affidabilità del sistema è inferiore a quella di un singolo disco SLED, poiché il *Mean Time To Failure* (MTTF) complessivo diminuisce con l'aumento del numero di dischi.
- **Applicazioni tipiche:** Il RAID 0 è utilizzato in applicazioni che richiedono alte prestazioni ma non necessitano di particolare affidabilità, come lo streaming audio e video.

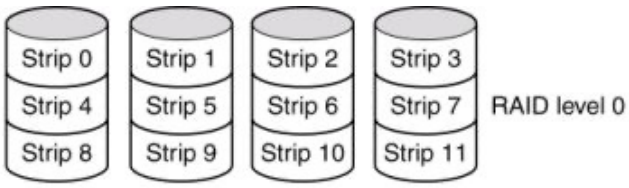


Figure 11.3: Raid Livello Zero

11.4.6 RAID di Livello 1: Mirroring

Definizione 11.4.3: RAID Livello 1

Un'architettura RAID in cui ogni disco di dati D è affiancato da un disco di mirroring che contiene una copia esatta dei dati memorizzati in D. La configurazione più semplice prevede l'utilizzo di due dischi: uno contenente i dati e l'altro la copia esatta.

- La duplicazione dei dati garantisce una maggiore affidabilità:
  - Se un disco si rompe, il sistema utilizza il disco di mirroring senza perdita di dati o interruzione del servizio.
- **Prestazioni:**
  - La scrittura dei dati è leggermente rallentata, poiché deve avvenire contemporaneamente su due dischi.
  - La lettura è più veloce, dato che i dati possono essere letti da entrambi i dischi in parallelo.
- **Limiti:** Il costo di implementazione è elevato, poiché richiede il raddoppio del numero di dischi.

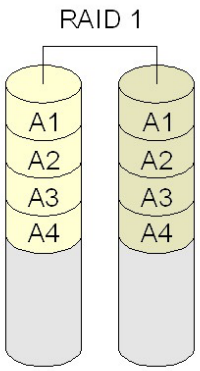


Figure 11.4: Raid Mirror

Esempio 11.4.2 (Combinazione RAID Livello 0+1)
Unendo le caratteristiche dello striping (RAID 0) e del mirroring (RAID 1), si ottengono configurazioni RAID ibride che massimizzano sia le prestazioni che l'affidabilità.

11.4.7 RAID di Livello 01: Striping + Mirroring

Definizione 11.4.4: RAID Livello 01

Un'architettura RAID che combina lo striping (livello 0) e il mirroring (livello 1).

- I dati sono suddivisi in *strip* distribuiti su più dischi (come nel RAID 0).
- Ogni disco è duplicato da un disco di mirroring (come nel RAID 1).

- **Affidabilità:**
  - Quando un disco si rompe, il sistema RAID utilizza il disco di mirroring per accedere ai dati, mantenendo il sistema operativo attivo.
  - Un disco rotto può essere sostituito, e i dati saranno copiati automaticamente dal "gemello" (se è disponibile uno *spare disk*).
- **Prestazioni:**
  - La lettura di un blocco di dati che coinvolge n strip può essere eseguita con n letture in parallelo, sfruttando sia i dischi primari sia quelli di mirroring.
- **Costo:** È una delle soluzioni RAID più costose, poiché richiede la duplicazione di tutti i dischi.

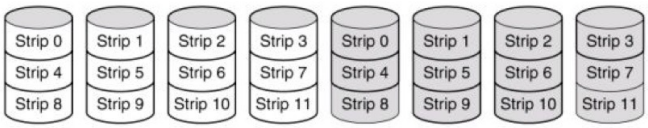


Figure 11.5: Raid Mirro 0+1

Osservazioni 11.4.2 Applicazioni del RAID 01

Il RAID di livello 01 è ideale per contesti in cui l'affidabilità è fondamentale, come la gestione di dati finanziari o bancari.

- Aumentando il numero di dischi coinvolti, migliorano le prestazioni complessive del sistema.

11.4.8 RAID Livello 10 (Mirroring + Striping)

Definizione 11.4.5: RAID Livello 10

Una configurazione RAID che combina il mirroring e lo striping in modo inverso rispetto al livello 01:

- Il mirroring viene eseguito per primo, creando coppie di dischi speculari.
- Successivamente, viene applicato lo striping alle coppie di dischi.

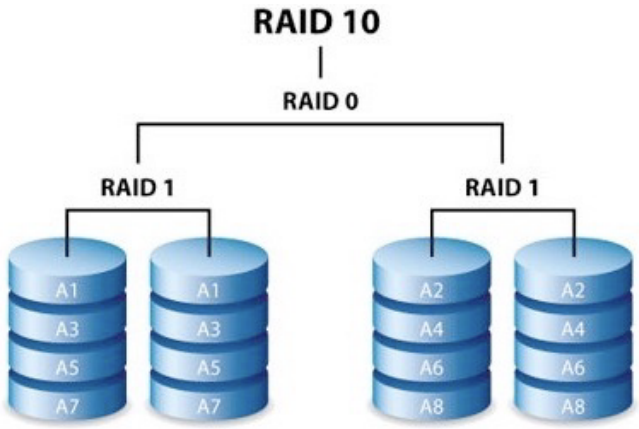


Figure 11.6: Raid 10

- **Prestazioni:** Simili a quelle del RAID livello 01.
- **Affidabilità:**
  - Vantaggi nel recupero dei dati in caso di guasti, poiché ogni coppia di dischi può essere gestita indipendentemente.

11.4.9 RAID Livello 4 (Striping con Parità)

Definizione 11.4.6: RAID Livello 4

Una configurazione RAID che utilizza lo striping a livello di blocchi e calcola uno strip di parità per consentire il recupero dei dati in caso di guasto.

- Gli strip di parità sono memorizzati in un disco dedicato.
- Ogni strip di parità è calcolato usando i corrispondenti strip degli altri dischi.

- **Vantaggi:**
  - Risparmio di dischi rispetto al RAID 01.
- **Svantaggi:**
  - Scritture più lente, poiché ogni modifica richiede l'aggiornamento dello strip di parità.
  - Il disco di parità può diventare un collo di bottiglia, aumentando la probabilità di guasti.

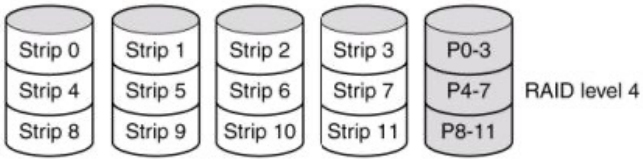


Figure 11.7: Raid livello 4

11.4.10 RAID Livello 5 (Striping con Parità Distribuita)

Definizione 11.4.7: RAID Livello 5

Una configurazione RAID simile al livello 4, ma con gli strip di parità distribuiti tra tutti i dischi, riducendo il carico sul disco di parità.

- Ogni disco contiene sia strip di dati sia strip di parità.

- **Vantaggi:**
  - Migliore distribuzione del carico di scrittura.
  - Ottima combinazione di prestazioni, affidabilità e capacità di memorizzazione.
- **Limiti:**
  - Ricostruzione complessa in caso di guasto, data la distribuzione degli strip di parità.
- **Utilizzo:** Livello RAID più usato per applicazioni generiche.

11.4.11 RAID Livello 6 (Striping con Doppia Parità Distribuita)

Definizione 11.4.8: RAID Livello 6

Una configurazione RAID simile al livello 5, ma con due livelli di parità distribuiti tra i dischi, permettendo di resistere al guasto simultaneo di due dischi.

- **Vantaggi:**

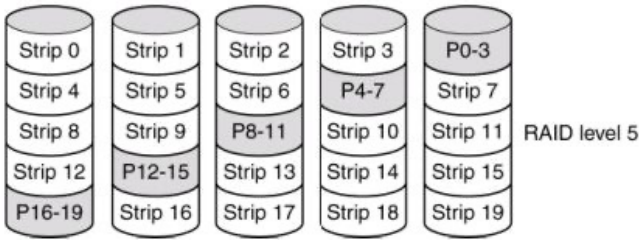


Figure 11.8: Raid Livello 5

  - Maggiore affidabilità rispetto al RAID 5.

- **Svantaggi:**
  - Richiede un disco in più rispetto al RAID 5 per memorizzare la stessa quantità di dati.
  - Maggiore overhead computazionale.

- **Utilizzo:** Poco usato, poiché la rottura contemporanea di due dischi è un evento raro.

11.4.12 Ricostruzione di Dati Persi con la Parità

Definizione 11.4.9: Parità

Una tecnica che consente di recuperare i dati di un disco guasto usando la parità bit a bit calcolata con un'operazione EX-OR tra le stringhe binarie memorizzate nei vari dischi.

- La parità è definita come:

$$\text{parità} = \text{stringa}_0 \oplus \text{stringa}_1 \oplus \cdots \oplus \text{stringa}_m$$

- Se un disco si rompe, il suo contenuto può essere ricostruito:

$$\text{stringa}_j = \text{parità} \oplus \text{stringa}_0 \oplus \text{stringa}_1 \oplus \cdots \oplus \text{stringa}_{j-1} \oplus \text{stringa}_{j+1} \oplus \cdots \oplus \text{stringa}_m$$

Esempio 11.4.3 (Esempio di Recupero con Parità)

Consideriamo tre stringhe binarie di 8 bit:

$$a = 01100011$$
$$b = 10101010$$
$$c = 11001010$$

La parità calcolata è:

$$p = a \oplus b \oplus c = 00000011$$

Se si perde la stringa a, possiamo ricostruirla:

$$a = p \oplus b \oplus c = 01100011$$

11.5 Memorie a Stato Solido (SSD)

Definizione 11.5.1: Memorie a Stato Solido

Memorie permanenti e riscrivibili basate su tecnologia flash, utilizzate come supporto di memoria secondaria. Caratteristiche principali:

- Organizzate in pagine da 2 a 16 Kbyte.
- Operazioni di lettura/scrittura coinvolgono l'intera pagina.
- Limite di circa 100.000 riscritture per pagina.
- Prima della riscrittura, le pagine devono essere cancellate tramite un processo detto *flashing*.

- **Utilizzo:**
  - Dispositivi mobili: smartphone, tablet.
  - Computer, spesso nei portatili di fascia alta:
    - Come sostituto del disco rigido per aumentare velocità e leggerezza.
    - Come complemento al disco rigido.
- **Tipi di supporto:**
  - **SSD (Solid State Disk):** Montati su schede o contenitori installabili nel computer.
  - **Pen Drive:** Incapsulati in contenitori rimuovibili con connessione USB.

11.5.1 Confronto tra Tipi di Memorie

Caratteristica	Hard Disk	Flash Memory	RAM
Costo per GB	~ 0,015$	~ 0,05-0,1$	~ 2-4$
Velocità in Lettura (T = RAM)	~ 1000 × T	~ 4 × T	T
Velocità in Scrittura (T = RAM)	~ 1000 × T	10-100 × T	T

Table 11.1: Confronto tra Hard Disk, Memoria Flash e RAM.

11.5.2 Utilizzo degli SSD

- **Sostituzione dell'hard disk:**
  - Maggiore velocità di accesso.
  - Minore peso per dispositivi portatili.
- **Combinazione con hard disk:**
  - *Cache permanente:* Collocata tra l'hard disk e la RAM.
  - *Secondo hard disk:* Utilizzato per:
    - Sistema operativo.
    - Applicativi.
    - File più utilizzati e area di swap.

    Gli altri file rimangono memorizzati sull'hard disk.
- **Politica di Accesso:**
  - Accesso completamente diretto.
  - Di solito gestito con politica *First Come, First Served (FCFS)*.

99 Esercizi

99.1 Capitolo 5

99.1.1 1

Processo	Durata	priorità
P_1	10	3
P_2	1	1
P_3	2	3
P_4	1	4
P_5	5	2

Table 99.1: Processi con durata e priorità

FCFS:

$\mathbf{P_1}$	$\mathbf{P_2}$	$\mathbf{P_3}$	$\mathbf{P_4}$	$\mathbf{P_5}$

SJF:

$\mathbf{P_2}$	$\mathbf{P_4}$	$\mathbf{P_3}$	$\mathbf{P_5}$	$\mathbf{P_1}$

99.2 Esercizi pre-esame

ESERCIZIO 2 (5 punti) In un SO la tabella delle pagine può contenere al massimo 256 (decimale) entry, e l'offset massimo all'interno di una pagina è FFF (esadecimale). a) Il SO potrebbe dover adottare un sistema di paginazione a due livelli (motivate la vostra risposta)?

c)