

Appunti Sistemi Operativi

Luca Seggiani
2025

1 Lezione del 23-09-25

1.1 Introduzione

Il corso di sistemi operativi riguarda l'ultima parte dello studio delle "architetture", che è partita con l'implementazione hardware in reti logiche, è continuata con lo studio del kernel in calcolatori elettronici, e termina appunto con lo studio dei sistemi operativi. Nello specifico, si considereranno sistemi operativi derivanti dalla famiglia **UNIX**.

Argomento del corso è la conoscenza di tecniche di programmazione usate nello sviluppo del sistema operativo **multiprogrammato** (più *processi* o più *thread*), con riferimento particolare alla programmazione **concorrente**, lo **scheduling** e la **memoria virtuale**.

Il corso mira a dare informazioni generiche utili allo studio di qualsiasi sistema operativo (anche non direttamente derivante da UNIX), in primis rivolte alla compresione di *come mai* una certa soluzione ad un problema è migliore di altre, e quali sono le tecniche che ci permettono di sviluppare soluzioni migliori.

1.1.1 Sistemi embedded e in tempo reale

Ci interesseremo anche ai sistemi **embedded** e soprattutto sistemi in **tempo reale**. Questi rappresentano sistemi *special-purpose* (per distinguere dai sistemi a scopo generale, *general-purpose*), dove dobbiamo rispettare coi nostri algoritmi di scheduling date scadenze temporali.

1.1.2 Programmazione concorrente

Con programmazione concorrente ci riferiamo alle tecniche che ci permettono di gestire più processi che si contendono le solite risorse, adottando politiche più o meno *equi* per i processi, o magari privilegiandole alcune. Obiettivo fondamentale sarà comunque quello di evitare *stalli* o **deadlock** dati da risorse occupate.

1.1.3 Programma del corso

Il corso è strutturato negli argomenti:

- **Concetti introduttivi** su sistemi operativi, architetture hardware e relativi cenni storici, in particolare ci interessano dettagli come la gestione della *pila* e le *interruzioni*;
- **Processi** e la loro gestione, inclusi gli algoritmi di *scheduling preemptive* e *non preemptive*, *prioritari* e *non prioritari* (FCFS, SJF, SRTF, RR). Inoltre si tratta la schedulazione nei sistemi **hard real-time** (RM, RDF);
- **Sincronizzazione** dei processi, quindi *programmazione concorrente*, *competizione* su risorse, e *scambio di informazioni fra processi* (IPC);
- Gestione della **memoria**, quindi *memoria virtuale*, *segmentazione* e *paginazione*;
- Gestione dei **dispositivi** di I/O, cioè i *driver*;

- **File system** su disco, cioè i componenti software che permettono la gestione di strutture di *file*, nella loro struttura sia *logica* che *fisica* di implementazione nei driver e nel sistema operativo;
- **Sicurezza**, quindi meccanismi di *protezione* fra processi, controlli sugli *accessi* sia in memoria che al filesystem, con riferimento al modello della *matrice degli accessi*.

1.1.4 Meccanismi e politiche

Una prima distinzione che possiamo fare è quella fra **meccanismo** e **politica**.

- Si dice **meccanismo** ciò che effettivamente implementato, in maniera sufficientemente veloce e compatta, nel kernel, per fornire il cosiddetto *supporto architetturale* a risorse, dispositivi, ecc...
- Nei sistemi operativi ci interessano invece principalmente le **politiche**, cioè decisioni (che vanno poi implementate) su come gestire *a priori* date risorse, dispositivi, ecc...

1.1.5 Sistemi operativi

Un **sistema operativo** è in primo luogo un *programma software* che ha il compito di fare da intermediario fra *l'utente* e *l'hardware* di un calcolatore.

Far fronte ai bisogni dell'utente significa gestire e consentire l'accesso delle risorse ai *processi* di cui l'utente necessita. In questo individuiamo come obiettivi del sistema operativo:

- Eseguire i *programmi utente*;
- Rendere il sistema facile da usare;
- Utilizzare l'hardware in maniera efficiente.

1.1.6 Programmi

I programmi con cui abbiamo a che fare sono per noi *liste di istruzioni* (tralasciando il fatto che queste siano codificate in linguaggio macchina o in un suo linguaggio mnemonic), ordinate ma che possono presentare salti condizionali che cambiano il normale *flusso di esecuzione*.

Il **comportamento** e quindi i **risultati** di un programma dipendono sì dal codice, ma anche dai **dati** in ingresso allo stesso. In questo possiamo dire che il programma non esiste mai da solo ed è solo la parte **statica** di un processo.

1.1.7 Risorse

Iniziamo quindi a vedere quelle che sono le risorse che dobbiamo fornire ai programmi. Il modello che adottiamo è il più diffuso oggigiorno, cioè quello di *Von Neumann*. Questo modello comprende, a grandi linee:

- La **CPU** o *processore*, che ha il solo scopo di prelevare ed eseguire le istruzioni in maniera sequenziale, alterando il suo flusso come già detto solo nel caso di istruzioni condizionali, o come vedremo nel caso di interruzioni o altre situazioni simili;

- La **memoria principale**, che nell'architettura di Von Neumann contiene sia i dati che il programma in esecuzione, e che deve essere capace di fornire su richiesta alla CPU.

Ricordiamo che questa è spesso *volatile*, cioè i suoi contenuti vengono sostanzialmente invalidati allo spegnimento della macchina. Potremmo interrogarci sul motivo di tale decisione: principalmente diciamo che le ragioni sono economiche, in quanto mantenere l'informazione per lunghi periodi di tempi è solitamente più costoso e delegato a dispositivi (come i dischi) che offrono risparmi in cambio di grandi tempo di accesso (inusuali sulla memoria principale);

- Altre **risorse** che si aggiungono a CPU e memoria, comunque indispensabili per eseguire qualsiasi istruzione. Queste sono:
 - I **dispositivi**, che comprendono ad esempio la memoria di *archiviazione* (il **disco**) e le *periferiche* di interfaccia con l'utente;
 - Le **risorse logiche**, cioè determinate strutture dati in memoria che devono essere fornite in maniera più o meno esclusiva ai processi. Anche gli stessi *file* del *file system* sono risorse logiche.

Risulta chiaro come la gestione delle risorse hardware e logiche è fondamentale anche alla **portabilità** dei programmi, che magari vogliono avere accesso a funzionalità simili su più sistemi operativi (accesso alla tastiera, ai file, ecc...), senza dover necessariamente conoscere l'implementazione interna di tali sistemi operativi.

Abbiamo quindi l'obiettivo di implementare tutte quelle **interfacce** di cui il programma bisogna per presentare all'utente le sue funzioni. Questo include le interfacce grafiche, audio, ecc... per la realizzazione di ambienti visuali e interattivi nei sistemi moderni.

Dal nostro lato, quello del *sistema*, vorremo che le soluzioni tecniche che adottiamo non impattino in maniera negativa le prestazioni o comunque il funzionamento dei programmi che mandiamo in esecuzione.

1.1.8 Struttura stratificata del S/O

La struttura di un sistema operativo può dividersi in più livelli, fra cui:

- Il livello **hardware**, fornito come già detto da risorse come:
 - La **CPU**;
 - La **memoria principale**;
 - Le **periferiche**, fra cui *video*, *disco*, *interfacce di rete*, ecc...

Il livello hardware offre la cosiddetta *intefaccia hardware*, data dalle specifiche secondo cui interagiamo con i dispositivi hardware stessi;

- Il livello **sistema operativo** (o *S/O*), che implementa la gestione delle risorse che studieremo nel corso, e offre a sua volta altre risorse logiche. In particolare notiamo:
 - Gestione della **CPU**;
 - Gestione della **memoria**;
 - Gestione del **file system** e quindi dei *file*;

- Gestione dei **dispositivi** attraverso i *driver*.

Questo livello offre la sua interfaccia attraverso le **chiamate di sistema** o *primitive*, che implementano una certa **API** (*Application Programming Interface*) secondo le quali i programmi utente delegano all'S/O operazioni che non potrebbero normalmente portare avanti da soli (accesso a risorse, schedulazioni temporali, ecc...);

- Il livello delle **applicazioni**, che comprende i programmi utente.

Questa gerarchia implica chiaramente che ogni livello non conosce nulla riguardo al livello successivo, ma si preoccupa solo di fornire un'*interfaccia* conforme alle specifiche. A questo punto è compito del livello successivo stesso rispettare l'interfaccia e farne uso per i suoi scopi.

Il programmatore di **sistema** interagisce con i livelli *hardware* e *S/O*, mentre il programmatore di **applicazioni** interagisce con i livelli *S/O* e *applicazioni*.

Compito dell'*API* è quello di generare per i programmati di applicazioni una macchina *astratta* più semplice da usare, più efficiente e più sicura (cioè realizzare gli obiettivi che ci eravamo posti in 1.1.5). Ricordiamo che per noi sicurezza significa *modelli* che controllano l'accesso da parte dei processi (altresì **soggetti**) alla memoria, e più in generale a tutte le risorse sistema (altresì **oggetti** dei programmi).

1.1.9 Definizione di S/O

Iniziamo a definire più nei dettagli cos'è un S/O.

- Un S/O è un **allocatore di risorse**, cioè gestisce *tutte* le risorse, e decide tra richieste conflittuali di accesso a tali risorse (inviate dai vari processi) al fine di garantirne un uso equo ed efficiente.
- Un S/O è però anche un **programma di controllo**, che controlla l'esecuzione dei programmi e lo stato delle risorse per prevenire usi impropri e stati inconsistenti.

Ricordiamo che in ogni caso l'unico programma effettivamente in esecuzione in ogni momento sulla macchina reale è il **kernel**, cioè nucleo, mentre il controllo viene temporaneamente passato fra programmi utente.

2 Lezione del 24-09-25

2.1 Cenni storici

Le prime macchine calcolatrici "moderne" nascono durante la seconda guerra mondiale, principalmente per scopi crittografici.

Fu nel periodo del secondo dopoguerra che diverse industrie, principalmente dal settore delle macchine da scrivere e di apparecchiature simili, decisamente di sviluppare queste tecnologie per scopi di ricerca e commerciali.

Di pari passo diverse università iniziarono a loro volta a sviluppare architetture e macchine calcolatrici, in questo caso a puro scopo di ricerca. Un esempio locale è quello della **CEP** (*Calcolatrice Elettronica Pisana*), sviluppata dai dipartimenti di matematica e fisica di Pisa (sotto indicazione di Enrico Fermi) per aiutare i ricercatori nei loro calcoli.

Sempre a Pisa fu l'ingegnere Mario Tchou a lanciare, in collaborazione con Olivetti, il progetto che diventò nel 1959 l'**Elea 9003**, fra i primi calcolatori a transistor commerciali (di contro la CEP funzionava a valvole termoioniche).

2.1.1 Sistemi Batch

In queste prime macchine, anche se la possibilità della multiprogrammazione era disponibile, raramente si parlava di "sistemi operativi" veri e propri. I primi sistemi operativi nascono quindi per i mainframe degli anni '60, fra cui notiamo gli **IBM Sistema 360** (e i successivi Sistema 370).

Inizialmente, queste macchine venivano usate in modalità **batch** (più programmi di più utenti eseguiti in sequenza): i primi S/O nascono appunto per permettere l'uso simultaneo (*time-sharing*) della macchina da parte di più utenti.

In ogni caso, già nei primi sistemi batch monoprogrammati si necessitava di diversi componenti effettivamente assimilabili ad un rudimentale sistema operativo:

- Un sistema di programmazione in memoria di massa (all'epoca nastri magnetici);
- Una *Job Control Language* (**JCL**), che esprimeva direttive interpretate da un *Monitor* (antenato delle moderne *shell*);
- Un **BIOS** (*Basic Input Output System*), cioè un insieme di routine per l'interazione con le periferiche.

L'S/O era quindi composto da Monitor + BIOS, che poteva essere configurato per caricare programmi e mandarli in esecuzione. In ogni momento in memoria si trovavano comunque il S/O e al più un programma utente.

2.1.2 Sistemi di spooling

Il prossimo passo è quello dei sistemi di **spooling** (*Simultaneous Peripheral Operation On-Line*). Questi nascono per permettere al programma utente di restare in esecuzione mentre le periferiche (all'epoca molto lente) completano le loro operazioni, bufferrizzando quindi le operazioni di ingresso/uscita.

I sistemi operativi che implementavano lo spooling dovevano quindi arricchirsi per permettere questo tipo di funzionalità.

2.1.3 Sistemi multiprogrammati

Arriviamo quindi ai sistemi **multiprogrammati**, cioè che permettono la gestione contemporanea di più programmi nella memoria principale: per la prima volta oltre al sistema operativo possiamo caricare in memoria più di un singolo programma utente.

I sistemi operativi di questo tipo si dovranno quindi dotare di diverse funzionalità, fra cui *scheduling* dei processi, possibilità di fare **DMA** (*Direct Memory Access*) sulle periferiche, *preemption* dei programmi in esecuzione, *memoria virtuale* per permettere mappature in memoria localmente costanti per ogni programma, ecc...

2.1.4 Sistemi time-sharing

Lo sviluppo di sistemi di tipo multiprogrammato è stato favorito dal fatto che i programmi utente che venivano sviluppati erano sempre più *interattivi*, quindi caratterizzati da fasi temporali distinte:

- **CPU-Burst**, dove il processore lavorava effettivamente sui dati;
- **I/O-Burst**, dove il processore attendeva operazioni I/O dalle periferiche, magari fornendosi del DMA.

Ci spostiamo quindi da un paradigma di esecuzione *sequenziale* ad un paradigma *multi-tasking*, dove il sistema operativo assegna ciclicamente istanti temporali (*quantum*) ai processi in esecuzione.

Il vantaggio dell'esecuzione multitasking è di poter avvicinare fra di loro i CPU-Burst, spostando il controllo della CPU da un processo all'altro quando si incorre in un I/O-Burst.

Per quanto ci riguarda, quindi, la tecnica del **time-sharing** non è che un modo per implementare il *multi-tasking*, cioè un caso particolare della *multiprogrammazione*, caratterizzato da processi in memoria che vengono eseguiti (o almeno hanno l'illusione di essere eseguiti) contemporaneamente. Ricordiamo che l'esistenza di più processi in memoria era di per sé caratteristica del sistema multiprogrammato.

L'idea di sviluppare diversi e sofisticati algoritmi di *scheduling* viene proprio dalla necessità di dover mantenere la CPU in piena attività, cioè eseguire più CPU-Burst possibile, scegliendo in maniera intelligente quali processi mandare in esecuzione (equivalememente, a quali processi assegnare i quantum temporali).

Notiamo che il tempo che la CPU passa a realizzare lo scheduling e i cambi di contesto rappresenta effettivamente **overhead** per il sistema, cioè tempo non passato ad eseguire programmi utente, ma in qualche modo "sprecato" in altri modi. Questo overhead è giustificato solo nel caso in cui le virtualizzazioni che consente permettono una velocizzazione considerevole della macchina.

2.1.5 Sistemi in tempo reale

La storia dei sistemi operativi ha un'interessante tangente nei cosiddetti sistemi **real-time** (*in tempo reale*). Questi sono sistemi dove lo scheduling è *deterministico* e il tempo impiegato ad eseguire un dato processo può quindi essere stabilito prima che questo venga lanciato.

Sistemi di questo tipo sono utili nel caso di calcolatori che interagiscono con *ambienti operativi* reali attraverso **sensori** ed **attuatori**, dove la precisione temporale con cui vengono eseguite certe operazioni è effettivamente importante alla funzione della macchina.

In particolare notiamo due paradigmi possibili per i sistemi real-time:

- **Soft** real-time, che non assicurano ma si impegnano a mantenere le specifiche sopra descritte;
- **Hard** real-time, il cui funzionamento ha come priorità imprescindibile le specifiche sopra descritte.

3 Lezione del 25-09-25

3.1 Richiami architetturali

Riprendiamo alcuni aspetti architetturali di un sistema di elaborazione. L'architettura che consideriamo è quella di *Von Neumann*, modello ancora oggi in uso e composto da:

- La **CPU** (*Central Processing Unit*) o come abbiamo già detto *processore*. Rappresenta un circuito piuttosto complesso che ha però l'unica funzione di *esecutore di istruzioni*.

Le istruzioni che questa esegue possono essere di tipo **CISC** (*Complex Instruction Set*), come ad esempio nell'architettura x86, o di tipo **RISC** (*Reduced Instruction Set*),

come ad esempio nell'architettura ARM. Ricordiamo comunque che nelle moderne implementazioni dell'x86 si traduce comunque in un instruction set RISC a livello architetturale per questioni di ottimizzazione.

Si può infatti dire che è inutile avere molte e complesse istruzioni (CISC) che richiedono molti cicli di clock, quando si possono avere poche e semplici istruzioni (RISC) che ne richiedono pochi: eventuali istruzioni più complesse potranno essere implementate come *subroutine* che usano più istruzioni semplici.

Ricordiamo quindi che la CPU si limita ad eseguire istruzioni, e non conosce (non memorizza) il programma. La poca memoria che ha a disposizione (sotto forma di *registri*) viene usata per mantenere i dati che sta elaborando;

- La **RAM** o *memoria centrale*, o ancora come abbiamo visto *memoria principale*. Questa ha il compito di memorizzare *dati* e *programma* (questo il fulcro dell'architettura di Von Neumann) e di renderli disponibili alla CPU e, come vedremo, anche ad altri dispositivi.

Abbiamo visto che è una memoria *volatile*, quindi che si mantiene solo finché il calcolatore è acceso, e che è una memoria ad *accesso diretto*, cioè si può accedere a qualsiasi locazione in tempo costante (a differenza di memorie di tipo *sequenziale*, ecc...).

Le operazioni che possiamo svolgere sulla memoria sono *lettura* e *scrittura* su locazioni di memoria. Nelle memorie moderne le letture sono *non distruttive*, mentre le scritture (chiaramente) lo sono.

- Qualche tipo di complesso di **I/O**. Questo comprende periferiche come *tastiera*, *porte seriali/parallele*, *interfacce di rete*, ecc...

Un dispositivo particolare che si trova nello spazio di I/O è il **disco** o *memoria secondaria*, a differenza della principale *persistente*, e usata per l'archiviazione di dati a lungo termine. Chiaramente, il tradeoff in questo caso è in termini di tempo (i dischi, anche allo stato solido, sono molto più lenti in tempo di accesso della RAM).

- Un **bus**, o *rete di interconnessione*, che permette a questi componenti di comunicare fra di loro.

Questa comunicazione dovrà essere **bidirezionale**, in quanto ad esempio la CPU deve sia leggere che scrivere dalla RAM: abbiamo visto come bus di questo tipo possono essere implementati sfruttando la logica a 3 stati.

Sperabilmente un bus dovrà contenere un numero consistente di linee. Torniamo all'esempio della CPU che legge in memoria: avremo bisogno di specificare l'*indirizzo* della locazione che vogliamo leggere, e vorremo vederci tornare una o più *parole* (cioè i dati che ci interessano) dalla memoria. Il modo più veloce per effettuare questa operazione è fornirsi di abbastanza linee per specificare sia gli indirizzi che i dati in **parallelo**: un bus *seriale* si dimostrerebbe molto più lento.

A livello logico dobbiamo dire anche che c'è bisogno di un **protocollo**, o comunque una qualche *politica* di gestione del bus.

- Ad esempio, la politica più semplice è quella dove la CPU è l'unica che può iniziare una transazione sul bus: questa è la classica configurazione *master-slave* dove la CPU rappresenta il *master* e memoria e I/O rappresentano gli *slave*;

- Esistono però situazioni dove potremmo volere che i dispositivi (ad esempio il disco) scrivano in memoria, o viceversa sia la memoria a scrivere sui dispositivi. Questo è effettivamente il caso del *DMA*. Avere un bus che supporta più iniziatori di transazioni richiede necessariamente un protocollo che stabilisca chiaramente chi può iniziare in quale momento una data transazione.

Le transazioni avvengono chiaramente in fasi, di cui ne individuiamo almeno 3 nel caso più semplice (singolo master, più slave):

1. Una prima fase di richiesta della transazione da parte dell'*iniziatore*;
2. Una fase di attesa da parte dell'iniziatore del responso dell'*obiettivo*;
3. Una fase dove l'operazione viene effettivamente eseguita, in un determinato lasso di tempo.

Nel caso di più master, abbiamo bisogno di meccanismi più sofisticati che implementino **mutua esclusione** e **sincronizzazione** delle risorse a cui i più iniziatori potrebbero voler accedere. Questo è vero sia a livello *logico* (su risorse logiche o comunque gestite dal S/O) che *elettrico* (2 o più componenti non pilotino mai le stesse linee contemporaneamente, pena fili bruciati).

Facciamo quindi una considerazione su come organizzare lo spazio di memoria e lo spazio dedicato ai registri delle periferiche. Esistono due configurazioni principali:

- **Memory-mapped I/O**: disponiamo i registri di I/O direttamente nello spazio di memoria, usando gli stessi indirizzi per indirizzare sia la memoria che i dispositivi;
- **Port-mapped I/O**: sfruttiamo due spazi, lo *spazio di memoria* e lo *spazio di I/O*, che mantengono separati i due tipi di informazione. Questo può essere fatto agilmente includendo un bit di selezione di spazio nel bus, ed è la soluzione adottata dall'architettura x86.

3.2 CPU

Vediamo nel dettaglio il primo componente, cioè la CPU.

3.2.1 Cicli CPU

Il funzionamento della CPU avviene in maniera **ciclica**: cogliamo più fasi che si ripetono nel tempo da quando questa viene accesa (reset) fino a quando viene spenta.

1. **Prelievo** o *fetch*: si legge la prossima istruzione in memoria, puntata dall'**IP** o **PC** (*Instruction Pointer* o *Program Counter*), e la si porta in un qualche registro interno al processore, pronta ad essere eseguita;
2. **Decodifica** o *decode*: si interpreta il significato dell'istruzione, cioè si individua qual'è effettivamente l'istruzione che dobbiamo eseguire, e si portano all'interno di registri gli eventuali *operandi sorgente* o gli indirizzi degli *operandi destinazione*;
3. **Esecuzione** o *execute*: si esegue effettivamente l'istruzione, direttamente attraverso la rete di controllo della CPU o sfruttando una o più **ALU** (*Arithmetic and Logic Unit*).

Successivamente, il risultato viene (se necessario) riscritto in memoria attraverso un'operazione di *write-back*. Questa fase viene a volte considerata come a sé stante (ad esempio nelle pipeline delle architetture RISC).

3.2.2 Registri CPU

La CPU è dotata di una sua memoria interna formata da locazioni di memoria dette **registri**. Questi si dividono in registri **general**i, riservati alle elaborazioni, e **di stato**, riservati a compiti speciali.

Registri generali

Consideriamo un set estremamente generico di registri:

- **AX, BX, CX e DX** sono i classici registri programmatore a uso generale;
- **ESP** è utilizzato per indirizzare la **pila o stack**, ovvero una parte di memoria con disciplina LIFO che serve a gestire sottoprogrammi.

Registri di stato

Ricordiamo due registri di stato:

- L'**IP** o **PC** (*instruction pointer* o *program counter*). Viene usato per contenere l'indirizzo della locazione dalla quale sarà prelevata la prossima istruzione da eseguire. Il contenuto dell'EIP è fissato al reset iniziale, e impostato sulla prima istruzione da eseguire.
- L'**F** (registro dei *flag*). Consiste di una serie di elementi binari detti **flag**, fra cui ricordiamo:
 - **OF**: flag di overflow (traboccameto) delle operazioni aritmetiche, si imposta se l'ultima operazione, presi gli operandi come interi, ha prodotto un risultato non rappresentabile su n bit;
 - **SF**: flag di segno, impostato quando l'ultima operazione restituisce un complemento a 2 con MSB = 1 (ergo negativo);
 - **ZF**: flag zero, che viene impostato quando l'ultima operazione restituisce qualcosa di nullo;
 - **CF**: flag di carry (riporto), che viene impostato quando l'ultima operazione richiede un riporto o un prestito, ergo presi gli operandi come naturali il risultato non è rappresentabile su n bit;
 - **IF**: flag di interruzioni attivate, quando è attivo il processore risponde alle interruzioni (che approfondiremo in seguito).

Al reset i flag visti finora sono impostati a 0.

3.2.3 Instruction set

Consideriamo un set di istruzioni estremamente basilare. Innanzitutto, possiamo dividere le istruzioni in **operative** e **di controllo**. Possiamo quindi fare ulteriori suddivisioni all'interno di queste categorie:

- **Operative:**

- Di trasferimento;
- Aritmetiche;
- Di traslazione/rotazione;

- Logiche.
- **Di controllo:**
 - Di salto;
 - Di gestione di sottoprogrammi.

Queste categorie andranno quindi a definirsi nelle varie istruzioni **MOV**, **ADD**, ecc... a cui siamo abituati.

Una nota va fatta adesso sulla scomodità data dall'utilizzo di istruzioni CISC: queste si sono sviluppate storicamente secondo il pensiero che era meglio dare più strumenti possibili al programmatore, ma oggi che il codice macchina è quasi esclusivamente compilato la dimensione variabile degli opcode rende difficile tecniche di ottimizzazione come il pipelining.

In ogni caso, per informazioni più approfondite sulla struttura generale del processore considerato si rimanda ai testi specializzati o agli appunti in <https://raw.githubusercontent.com/seggiani-luca/appunti-rl/34228f66db395637bd1824d04f3130b977cc0ce4/master/master.pdf>.

3.3 RAM

Approfondiamo quindi il discorso della RAM. Nel sistema considerato inseriremo un elemento di **cache**, nello specifico fra la CPU e il bus (da cui si accede alla RAM). Il funzionamento della cache è dettagliato in <https://raw.githubusercontent.com/seggiani-luca/appunti-ce/638d3abf2e1d473632b575401582203c3b113c82/master/master.pdf>, e per quanto ci riguarda possiamo dire che funge da unità di "*memoizzazione*" dei dati (quando vengono richiesti), più veloce della RAM.

4 Lezione del 30-09-25

Continuiamo la discussione della memoria RAM.

Avevamo introdotto il meccanismo della cache come una sorta di "*memoizzazione*" dei dati in occasione del primo accesso. In verità, nei moderni processori (dal Pentium in poi) abbiamo due cache separate:

- La **I-cache**, cioè *cache istruzioni*;
- La **D-cache**, cioè *cache dati*;

Il vantaggio di distinguere fra cache istruzioni e cache dati è che la I-cache non ha bisogno di essere ricoppiata in memoria alla fine dell'utilizzo, e probabilmente deve mantenere zone di memoria molto specifiche, per cui ha senso non rallentarne l'operazione chiedendole di mantenere anche informazioni sui dati.

4.0.1 Gerarchie di memoria

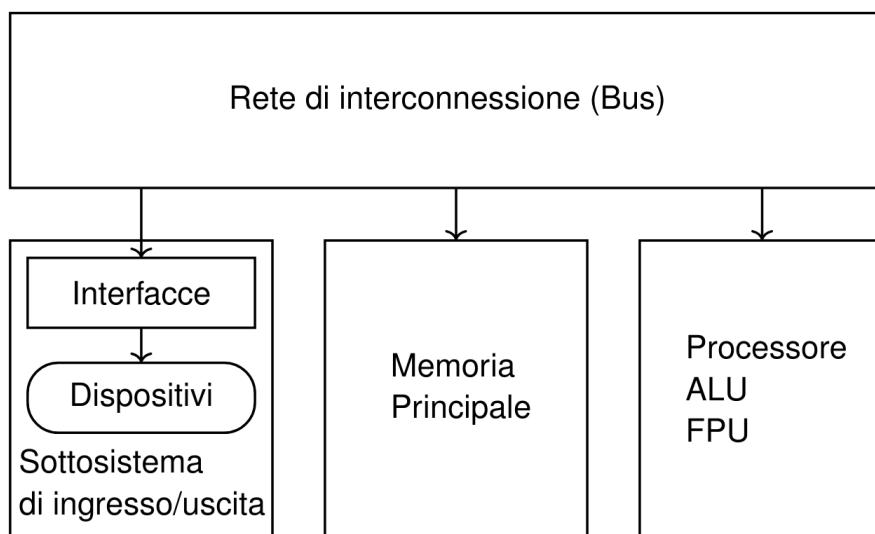
Fra registri, RAM, dispositivi a blocchi, ecc... abbiamo visto diverse fonti di *memoria* che un calcolatore può utilizzare. Potrebbe avere senso organizzare queste memorie in una struttura gerarchica, magari per *dimensione crescente* in avanti (e di conseguenza per [velocità] all'indietro):

1. Registri interni;

2. Cache;
3. RAM;
4. Dischi;
5. Nastri.

4.1 Schema a blocchi di un semplice calcolatore

Possiamo quindi, dopo aver visto tutte le componenti che lo compongono, vedere lo schema a blocchi di un semplice calcolatore:



Vediamo quindi come questi componenti comunicano fra di loro:

- La rete di interconnessione (bus) serve tutti (a scapito della direzione delle frecce, può supportare la comunicazione *da e a* componenti);
- Il processore e la RAM si trovano sul bus;
- I dispositivi, cioè i trasduttori col mondo esterno, comunicano con il sistema attraverso le loro *interfacce*, che obbedisce da un lato alle regole del bus e dall'altro alle specifiche del dispositivo stesso per permettere la comunicazione.

4.2 Interfacce

Abbiamo visto come fra il calcolatore ed ogni dispositivo si trovi un'apposita *interfaccia*.

Di base, ogni interfaccia è caratterizzata da più registri (accessibili nello spazio di I/O), che possono essere scritti o letti dal calcolatore per dare o ottenere informazioni dal dispositivo. Notiamo che letture e scritture sui registri delle interfacce possono essere distruttive: spesso il dispositivo implementa particolari funzioni che vengono lanciate da operazioni di questo tipo (un registro che si azzera dopo esser letto, ecc...).

Nel caso più semplice, in ogni caso, un'interfaccia dispone di almeno 3 registri:

- Registro di **stato**, che segnala lo stato corrente dell'interfaccia (se è di uscita può segnalare che è pronta a ricevere dati, se di entrata che ci sono dati pronti, ecc...);

- Registro di **controllo**, che permette al calcolatore di comandarne l'operazione (se di entrata può impedire che nuovi dati arrivino in ingresso, ecc...);
- Uno o più **buffer dati**, resi accessibili attraverso un registro di lettura. Solitamente si dice **TBR** (*Transfer Buffer Register*) il registro che accede al buffer di uscita e **RBR** (*Receive Buffer Register*) il registro che accede al buffer di entrata. Nel caso di interfacce di ingresso/uscita TBR e RBR stanno alla stessa porta dello spazio di I/O, e quale viene reso disponibile al processore varia in base al tipo di operazione che esso richiede (TBR per uscita, RBR per ingresso).

4.3 Interruzioni

Veniamo quindi al meccanismo dell'**interruzione**. Nella formulazione originale di Dijkstra queste servivano a risparmiare al processore l'attesa "attiva" (*busy wait*) dei bit di stato delle interfacce, delegando questo invece ad una segnalazione esplicita da parte dell'interfaccia, che viene *gestita* dal processore mettendo in esecuzione un determinato *handler* di interruzione.

Per gestire correttamente le interruzioni abbiamo bisogno di un po' di infrastruttura in più:

- Una nuova fase processore, successiva all'esecuzione, che si occupa di controllare le richieste di interruzioni in arrivo (nei sistemi x86 la richiesta, che è stata inoltrata da un sottosistema detto APIC);
- Una zona di memoria dove viene allocata la **IVT** (*Interrupt Vector Table*), che associa ad un indice (cioè il tipo di interruzione) l'inizio dell'handler relativo a tale tipo;
- Una nuova istruzione, **IRET**, che si occupa di ritornare da un gestore di interruzione.

Non potremmo usare la semplice RET in quanto ogni interruzione salva dello stato aggiuntivo oltre al semplice IP sulla pila: di base, salveremo anche il registro dei FLAG.

4.3.1 Tipi di interruzione

Abbiamo visto nel corso di calcolatori elettronici come il meccanismo delle interruzioni può essere sfruttato per implementare molta più funzionalità di quelle relative alla gestione dei dispositivi. In particolare, i calcolatori moderni dispongono di più tipi di interruzioni:

- Interruzioni **esterne**, del tipo che abbiamo appena visto, che si distinguono ulteriormente in:
 - Interruzioni esterne **mascherabili**, cioè che possono essere ignorate variando il flag IF;
 - Interruzioni esterne **non mascherabili**, cioè che vengono sempre gestite;
- Interruzioni **interne**, cioè lanciate da situazioni interne al processore (eccezioni);
- Interruzioni **software**, che possono essere lanciate dal programmatore attraverso l'apposita istruzione **INT**.

4.4 Meccanismi di protezione

Veniamo quindi ai meccanismi tipici del S/O in sé per sé. Se vogliamo la separazione fra processi e S/O che gestisce quei processi (e quindi ha l'accesso prioritario alle risorse di sistema), dobbiamo separare l'operazione del calcolatore in due modalità principali:

- Modalità **utente**: usata per la normale esecuzione dei programmi, non è possibile accedere a tutte le risorse di sistema;
- Modalità **supervisor**: usata per lo svolgimento delle chiamate sistema (primitive), tutte le risorse di sistema sono disponibili.

Importante è che il passaggio da modo utente a modo supervisor richieda al programma in esecuzione in modo utente di "abbandonare" il controllo, cedendolo ad una primitiva sistema. Vedremo come questo si può implementare agilmente sfruttando il meccanismo di interruzione.

4.5 Componenti del S/O

Vediamo quindi, come abbiamo fatto per l'hardware, quelli che sono i **componenti** del S/O e come questi sono organizzati.

Prendiamo come riferimento un sistema *Unix*, in quanto più semplice ed elegante per i nostri scopi.

Adottando un approccio *top-down*, dove per *top* intendiamo lo spazio dell'utente, vediamo le seguenti componenti:

- L'**userspace**, cioè gli applicativi utente veri e propri;
- Gli strumenti che il S/O fornisce all'utente per la gestione del sistema, cioè:
 - La **shell**;
 - **Compilatore e linker**;
 - Le **librerie** sistema (che si rivolgono ad API, ecc...).
- Il **kernel**, cioè la parte del S/O che effettivamente gestisce il sistema. Qui troviamo:
 - Il sottosistema **file**, che gestisce il filesystem su uno o più dispositivi a blocchi;
 - A sua volta il sottosistema file interagisce con i **driver** dispositivo (in particolare coi driver dei dispositivi a blocchi), che hanno il compito di gestire a livello hardware il comportamento dei dispositivi;
 - Inoltre, troviamo il sottosistema **controllo processi**, composto da:
 - * Funzionalità **IPC** (*Inter Process Communication*) per la comunicazione fra processi;
 - * Lo **scheduler**, che decide quali processi mandare in esecuzione;
 - * Il sottosistema di **gestione memoria**, che gestisce lo spazio in memoria principale allocato per ogni processo, interagendo col meccanismo della *memoria virtuale*.
- Infine, il *kernel* si appoggia all'**hardware** della macchina.

4.5.1 Modello gerarchico

Un modello più complesso per S/O potrebbe elaborare su questa struttura, prevedendo più livelli intermedi di kernel che implementano *macchine virtuali* via via più vicine all'hardware. Ognuna di queste fornirà al livello superiore funzioni (effettivamente chiamate sistema) sempre più astratte, che implemetteranno nel complesso le chiamate sistema rese disponibili ai processi utente.

4.5.2 Modello client-server

Un'altro modello possibile per sistemi distribuiti in rete è quello di avere più *nodi* collegati alla stessa rete. Ogni nodo disporrà del suo kernel, e in esecuzione su quel kernel avrà uno specifico processo (utente o sistema).

La funzionalità del S/O sarà quindi implementata interrogando la rete per il servizio richiesto: sarà quindi compito della macchina su quella rete che effettivamente implementa tale servizio rispondere e fornire, appunto, il servizio.

In ogni caso non analizzeremo sistemi di questo tipo in questo corso, relativi più che altro a sistemi su *cloud*, e quindi all'ambito delle reti informatiche.

4.6 Gestione processi

Informalmente, il termine processo viene usato per indicare un programma in esecuzione sulla macchina.

- Rappresenta la *sequenza di eventi* generati dall'elaboratore durante l'esecuzione;
- Identifica la più piccola *unità di esecuzione* dentro un S/O multiprogrammato: questo consentirà l'esecuzione di *più* processi concorrenti;

Un processo va necessariamente *descritto*, cioè bisogna definire un **descrittore** che lo rappresenta. Del processo ci interessa:

- Il **codice** del programma che esegue;
- I **dati**;
- Il valore dell'**IP**;
- Lo stato dei **registri**;
- Lo **stack**.

Inoltre, ad un certo processo potranno essere associate delle risorse:

- **Memoria** utilizzata;
- **File** aperti;
- **Dispositivi** di I/O a cui ha accesso.

4.6.1 Processi in memoria

Il processo in memoria ha a disposizione il suo *spazio di indirizzamento virtuale*. Viene detto *virtuale* perché verrà allocato in una memoria centrale fisica, le cui locazioni potrebbero non corrispondere esattamente con la memoria offerta al processo (attraverso il meccanismo della *memoria virtuale*).

Partendo dal basso, le regioni di memoria fornite al processo nel suo spazio di indirizzamento saranno:

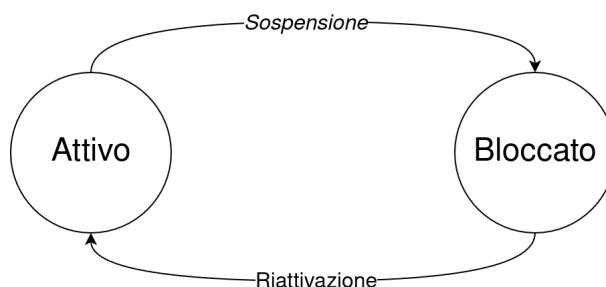
1. `text`: contiene il codice del processo;
2. `data`: contiene i dati statici del programma (sezione `data` e `bss`, che contiene lo spazio riservato a variabili statiche non allocate);
3. `heap`: l'heap del processo, dove vengono allocati oggetti in memoria dinamica;
4. `stack`: si sviluppa verso il basso, rappresenta la pila del processo in esecuzione.

5 Lezione del 02-10-25

Continuiamo la discussione dei processi:

5.0.1 Stato dei processi

In un **S/O monoprogrammato** il processo può trovarsi in uno di due stati:



- **Attivo**: il processo stato creato, è in esecuzione ed ha ancora istruzioni da eseguire.

Creare un processo significa in primo luogo allocare le strutture dati che lo descrivono. In seguito, si deve allocare un po' di memoria al processo stesso per contenere i suoi dati (istruzioni, pila, ecc...) come visto in 4.6. Abbiamo però che allocare particolari descrittori al processo quando questo è l'unico in esecuzione sarebbe inutile: effettivamente un sistema monoprogrammato può essere tale solo se il processo in esecuzione è in qualche modo il S/O.

A questo punto il processo è l'unico in esecuzione sulla CPU, e resterà tale fino alla fine del suo *lifetime*. Dovrà però *bloccarsi* se vuole accedere a risorse sistema: farà ciò usando una *chiamata a sistema*;

- **Bloccato**: il processo o qualche altro attore ha causato un qualche evento che ha determinato il passaggio di controllo al S/O (richiesta risorse di I/O, di risorse logiche, interruzioni esterne, ecc...).

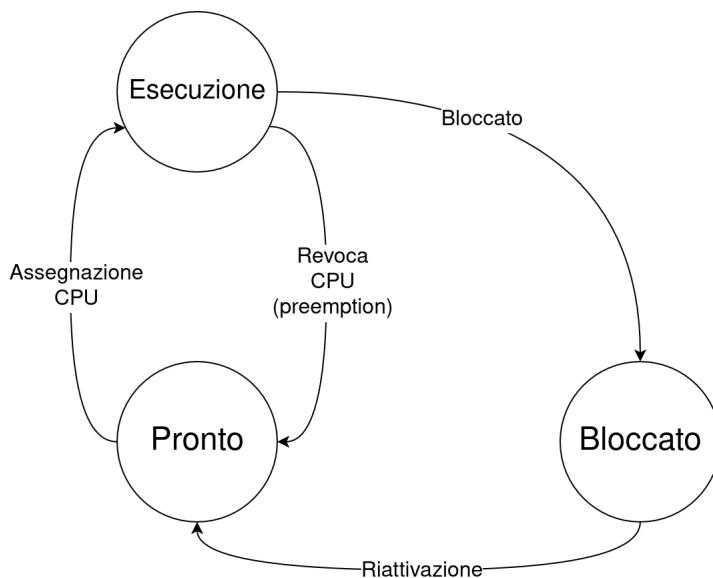
A questo punto sarà nuovamente un evento esterno (interruzione esterna, azione dell'utente) a riportare in esecuzione un processo, creandone uno nuovo se lo scorso aveva terminato, o rimettendo il corrente in esecuzione se si era bloccato su una operazione di I/O o simile.

Notiamo che la memoria del processo potrebbe cambiare anche questo è bloccato: ad esempio se si hanno dispositivi che operano in DMA.

La transizione fra attivo e bloccato è detta *sospensione*, mentre fra bloccato e attivo è detta *riattivazione*.

Possiamo dire che gli stati *attivo* e *bloccato* sono in qualche modo in corrispondenza con le fasi descritte in 2.1.4 di **CPU-Burst** e **I/O-Burst**.

Se il numero di CPU è minore del numero dei processi, cioè in un sistema **monoprocesso**, ci dotiamo di più stati:



- **Pronto**: in questo caso il processo è in attesa del tempo del processore. Abbiamo che nella maggior parte delle implementazioni questo stato è rappresentato da una *coda* di descrittori di processo, la cosiddetta **coda pronti** (utile per implementare politiche prioritarie ↔ code prioritarie);
- **Esecuzione**: il processo è in esecuzione sulla CPU;
- **Bloccato**: il processo è in attesa, come nell'esempio precedente.

In questo caso, come nel precedente, la memoria processo potrebbe essere modificata da dispositivi in DMA. Potrebbero poi essere modificati, in particolari situazioni, i descrittori di processo: magari a causa di operazioni di altri processi col S/O (pipe fra processi, ecc...).

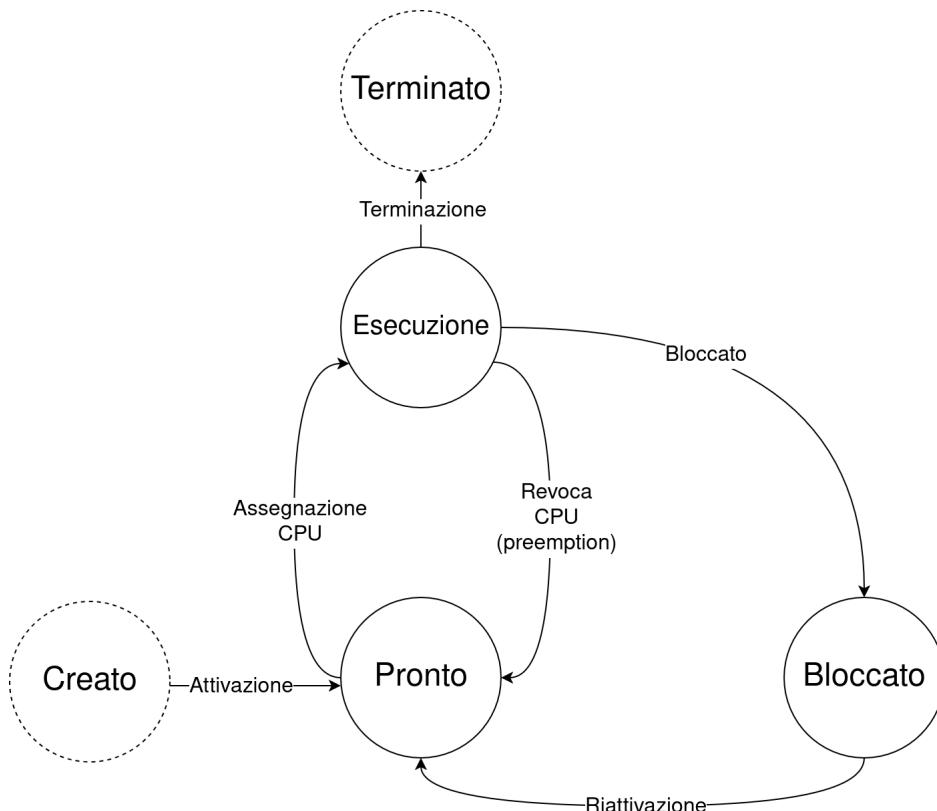
La transizione da pronto e esecuzione è detta *assegnazione CPU*, mentre la contraria è detta *revoca CPU*, o in inglese **preemption**. Notiamo che quest'ultima transizione non è prevista da tutti i sistemi.

L'operazione di assegnazione CPU è eseguita da un componente del S/O detto **scheduler**. Lo scheduler viene eseguito quando il processo in esecuzione cambia stato, quindi

in caso di *revoca CPU* o *sospensione*: sostanzialmente ogni volta che si richiede un nuovo processo da mettere in esecuzione. Questo comprende anche la situazione in cui il processo corrente *termina*.

Abbiamo poi che la transizioni allo stato bloccato sono le stesse dell'esempio precedente, con l'unica differenza che la *riattivazione* del processo non lo mette in esecuzione ma nello stato *pronto*.

C'è inoltre un'altro stato, oltre a quello di *terminazione* prima accennato, di cui dobbiamo tener conto: quello in cui il processo viene *creato*, cioè la transizione di un nuovo processo allo stato di *pronto* (detta *attivazione*):



In questo caso possiamo anticipare che in S/O che adottano politiche *prioritarie* allo scheduling potrebbe essere necessario eseguire lo scheduler (per mettere in esecuzione un nuovo processo di priorità più alta).

5.0.2 Descrittori di processo

Abbiamo introdotto come la gestione dei processi bisogna di apposite strutture dette *descrittori di processo* (in inglese **PCB**, *Process Control Block*).

Questa dovrà associare ad ogni processo:

- Nome del processo: questo è il solito **PID** (*Process IDentifier*), ed identifica univocamente ogni processo. Richiedere che i PID siano univoci è una *politica* che li trasforma in una *risorsa*: l'S/O dovrà impegnarsi a gestire i PID in modo che non avvengano collisioni;
- Stato del processo: codifica uno degli stati definiti prima;

- Modalità di servizio: questo riguarda la priorità (se implementata) o il tipo di scheduling che si usa per gestire il processo;
- Informazioni sulla gestione della memoria: conterrà puntatori alla memoria dedicata al processo;
- Contesto del processo: l'immagine dei registri all'ultima sospensione del processo;
- Utilizzo delle risorse: conterrà puntatori alle risorse logiche e fisiche a cui ha accesso il processo;
- Identificazione del processo successivo: questo serve semplicemente ad implementare, come abbiamo accennato, le code prioritarie di processi (come ad esempio la *coda pronti*).

Una volta definito il descrittore di processo, potremo volerci fornire di:

- La coda dei processi pronti;
- Una o più code per i processi bloccati (solitamente una coda è associata a una risorsa su cui i processi si bloccano);
- Un registro di qualche tipo che contiene il processo attualmente in esecuzione.

5.0.3 Cambio di contesto

Il **cambio di contesto** è l'operazione attraverso cui l'uso del processore viene commutato da un processo all'altro. Questo consiste in:

1. Salvataggio del contesto del processo in esecuzione nel suo descrittore (cioè salvataggio di *stato*);
2. Inserimento del descrittore di processo corrente in coda *bloccati* o *pronti*.
3. Selezione di un nuovo processo da mettere in esecuzione e caricamento nel registro del descrittore del processo corrente (*short term scheduling*);
4. Caricamento del contesto del nuovo processo e cessione a questi del controllo.

Notiamo che per realizzare il cambio di contesto al S/O (che si occupa poi di portare avanti queste operazioni) abbiamo bisogno di funzionalità implementate in hardware: il "processo" S/O è semplicemente il processo in esecuzione al tempo del cambio, che è forzato a passare al contesto sistema nell'istante in cui si mette ad eseguire codice sistema. Questo si riassume nella lista di operazioni ai passi (1) e (4).

Questo significa che per realizzare un S/O si necessita di un processore che implementi il cambio di contesto (x86 dal 286, ecc...).

Possiamo iniziare ad approfondire il modo in cui indicizziamo il processo. Dal punto di vista concettuale, vorremo usare una tripla:

$$S = \langle \text{PID}, \text{UID}, \text{GID} \rangle$$

composta da **PID** (*Process IDentifier*, già visto), e **UID** e **GID** (*User IDentifier* e *Group IDentifier*), che rappresentano il proprietario del processo.

In caso di cambi di contesto al contesto sistema, quello che faremo è semplicemente cambiare UID e GID in modo da eseguire lo stesso processo, ma con privilegi diversi.

5.0.4 Creazione e terminazione di processi

Vediamo le operazioni che si possono svolgere sui processi.

Avremo che vorremo supportare *gerarchie* di processi, dove un processo (padre) può richiedere la creazione di un nuovo processo (figlio). Questo significherà che ogni processo sarà figlio di un altro processo, e potrà a sua volta essere padre di altri processi. Chiaramente, le informazioni relative alle relazioni parentali saranno mantenute da S/O nei descrittori di processi.

Se un processo termina, di base:

- Il padre può rilevare il suo stato di terminazione;
- Tutti i suoi figli terminano.

5.0.5 Processi concorrenti

Specifichiamo come più processi vengono eseguiti su una stessa macchina. Abbiamo che questi possono alternarsi (*interleaving*), tipico delle macchine a singolo processore, o eseguire effettivamente l'uno contemporaneamente all'altro. Questo è tipico delle macchine a più processori.

Nel caso di più processi in esecuzione contemporaneamente (diciamo processi *concorrenti*) vengono a verificarsi alcune problematiche:

- **Processi indipendenti:** se due processi P_1 e P_2 sono indipendenti, cioè non influenzano l'uno l'esecuzione dell'altra, dobbiamo assicurarci che questo resti vero, cioè dobbiamo risolvere il *problema della riproducibilità*. Visto che le risorse sistema sono condivise, dobbiamo infatti assicurarci che non ci siano effetti collaterali sensibili da un processo o dall'altro.
- **Processi interagenti:** se due processi P_1 e P_2 devono interagire fra di loro, possono farlo in maniera *esplicita* (per **cooperazione**), scambiandosi messaggi, o *implicita* (per **competizione**), magari competendo per la stessa risorsa in mutua esclusione.

6 Lezione del 07-10-25

6.1 Nucleo

Il **nucleo** o *kernel* è il cuore di un S/O, il componente che ha il compito di realizzare l'astrazione della *CPU virtuale*. Nel caso di sistemi monoprocesso, vogliamo dividere il tempo fra i processi per dargli l'illusione di essere gli unici in esecuzione sulla macchina.

6.1.1 Scheduling

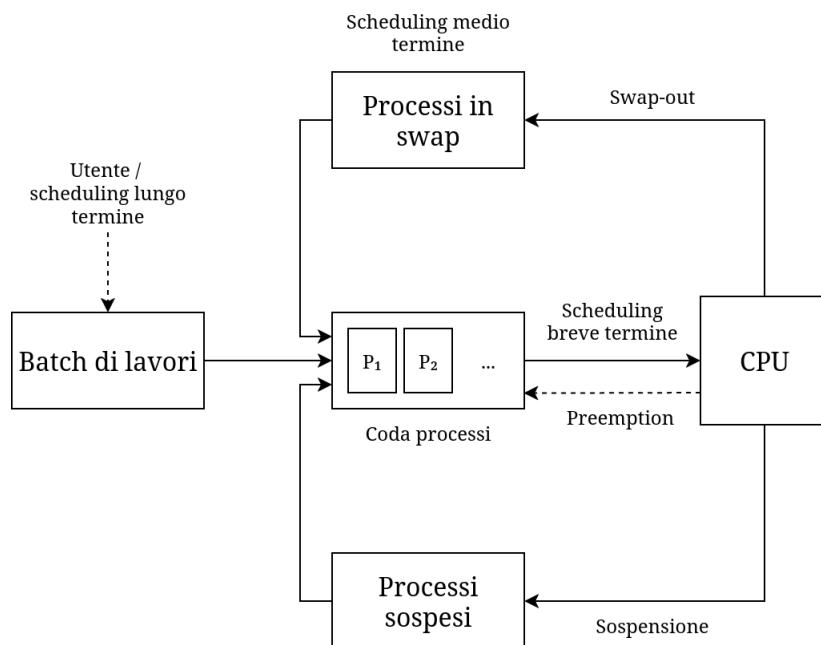
Lo scheduling è l'attività secondo la quale il sistema operativo effettua delle scelte fra quali processi caricare in memoria centrale e a quali assegnare la CPU.

Ci sono 3 diverse attività di scheduling:

- **Breve** termine: lo scheduling propriamente detto, il processo attraverso cui si assegna la CPU. Può essere *preemptive* e *non preemptive* (con o senza diritto di revoca). Solitamente viene invocato molto frequentemente (millisecondi);

- **Medio** termine (*swapping*): il trasferimento temporaneo in memoria secondaria dei processi. Si usa quando la memoria centrale dispone di memoria minore della somma di quella richiesta dai vari processi. Viene invocato più di rado (secondi, minuti);
- **Lungo** termine: la scelta di quali processi caricare dalla memoria secondaria in memoria centrale. Rappresenta un componente importante dei sistemi *batch* multiprogrammati, oggi come oggi quindi sui *server* e meno sulle macchine personali;

Vediamo quindi una schematica che mostra dove queste attività di scheduling avvengono nell'architettura vista:



I processi possono in genere classificarsi in:

- Processi vincolati da **I/O**: passano più tempo a fare I/O burst piuttosto che CPU burst (che sono tanti e piccoli);
- Processi vincolati da **CPU**: passano più tempo a fare calcoli, hanno pochi e lunghi CPU burst.

6.2 Algoritmi di scheduling

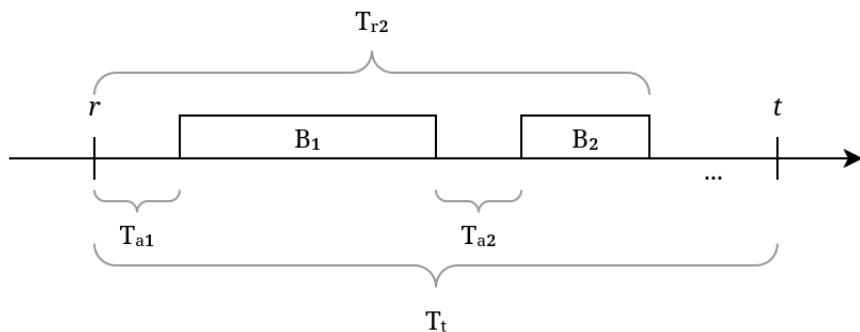
Gli algoritmi di scheduling che vedremo saranno:

- **FCFS (First Come First Served)**: è *non prioritario* e *non preemptive*, consiste nel assegnare la CPU sempre al primo processo arrivato;
- **SJF (Shortest Job First)**: è *prioritario* e *non preemptive*, consiste nell'assegnare la CPU al processo più breve;
- **SRTF (Shortest Remaining Time First)**: è *prioritario* e *preemptive*, rappresenta sostanzialmente la versione con revoca del precedente;
- **RR (Round Robin)**: è *non prioritario* e *preemptive*: si basa sull'assegnare quanti temporali ugualmente ad ogni processo;

- Schedulazione **su base prioritaria**: introdurremo qui l'idea di *priorità* per ogni processo;
- Schedulazione **a code multiple**: prevediamo più code, che possiamo distinguere usando gli algoritmi sopra descritti, o come vedremo sarà conveniente, assegnando una *priorità* ad ogni cosa;
- Schedulazione di sistemi **in tempo reale**. Questi sono algoritmi che devono assicurare la terminazione deterministica dei processi. In questo vedremo gli algoritmi:
 - **RM (Rate Monotonic)**;
 - **EDF (Earliest Deadline First)**.

6.2.1 Valutazione degli algoritmi di scheduling

Iniziamo a vedere alcune metriche per la valutazione degli algoritmi di scheduling:



- **Utilizzazione** della CPU o *efficienza*, cioè definito ΔB_i il tempo di burst e T il quanto di tempo totale, vogliamo un efficienza E :

$$E = \frac{\sum \Delta B_i}{T} < 1$$

il più possibile vicina a 1;

- **Tempo medio di completamento** (o tempo di *turnaround*), cioè il tempo che passa prima che il processo possa completare la sua operazione (terminando). Prendiamo l'istante in cui il processo entra in coda pronti come *r* (da *richiesto*) e quello in cui termina come *t* (da *termina*). Il tempo di completamento *T_t* sarà ovviamente:

$$T_t = t - r$$

- **Produttività** (o frequenza di *throughput*), definita come il numero medio di processi completati nell'unità di tempo, cioè l'inverso del tempo medio di completamento:

$$P = \frac{1}{T_t}$$

- **Tempo di risposta**, valutato dall'istante in cui un processo entra in coda pronti *r* all'istante in cui risponde, cioè termina un CPU burst (solitamente il primo). Purtroppo, non tutto il tempo di completamento *T_c* è dedicato al processo, ma

questo viene eseguito, come sappiamo, in più burst (diciamo B_1, B_2, \dots). Il tempo di turnaround T_t sarà allora il tempo trascorso fra r e la fine di un burst B_i , cioè:

$$T_{ri} = \text{end}(B_i) - r$$

- Tempo di **attesa**, cioè la somma dei tempi di attesa posti fra i vari burst:

$$T_a = \sum t_{\alpha i}$$

Con riferimento ai sistemi interattivi, spesso ci interessa il tempo di attesa iniziale, cioè quello fra l'inserimento in coda pronti e l'inizio del primo CPU burst, o in relazione al nostro schema:

$$T'_a = t_{\alpha 1} = r - \text{begin}(B_1)$$

Dovrebbe essere chiaro che il tempo di *attesa* si distingue dal tempo di *risposta*, in quanto:

- Il tempo di attesa è quello visto dal *processo* prima che questo possa iniziare la computazione;
- Il tempo di risposta è quello visto dall'*utente* prima di vedersi tornare un primo risultato (si presume che alla fine dei CPU burst inizia un I/O burst che porta avanti qualche operazione di lettura o scrittura da periferiche, tangibile per l'utente).
- Rispetto dei **vincoli temporali**, utile principalmente negli algoritmi di scheduling in *tempo reale*.

Fra queste metriche, tempo di **risposta** e di **attesa** sono relativi principalmente ai *sistemi interattivi*, mentre il rispetto dei **vincoli temporali** è relativo ai sistemi in *tempo reale*.

Chiameremo poi O_v l'**overhead** associato all'esecuzione dello scheduler. Ricordiamo che in ogni caso in questa fase stiamo parlando di scheduling a breve termine.

6.2.2 Algoritmo FCFS

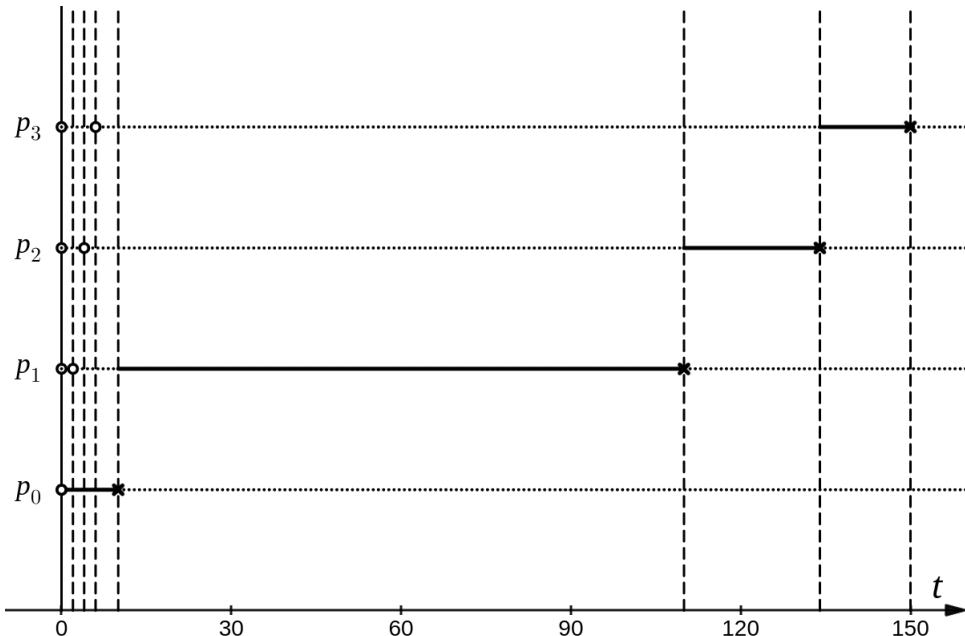
Nell'algoritmo **FCFS** (*First Come First Served*) assegnamo la CPU al primo processo in coda pronti. Sostanzialmente, trattiamo la coda pronti come una coda **FIFO** (*First In, First Out*). Questo lo rende non prioritario e non preemptive.

Quello che otteniamo è un efficienza teorica pari a $E \sim 1$ (c'è un piccolo overhead $O_v \sim 0$ dato dal cambio di contesto), ma generalmente prestazioni piuttosto limitate. Questo è dovuto al fatto che i tempi di attesa (e di conseguenza di completamento) dei processi sono completamente aleatori, e non si fa alcuna scelta informata mirata a minimizzarli.

Vediamo ad esempio il comportamento ottenuto con la sequenza di processi:

Processo	T richiesta	C esecuzione
p_0	0	10
p_1	2	100
p_2	4	24
p_3	6	16

Su un grafico con il tempo t alle ascisse, avremo:



Questo risulta in un tempo medio di attesa di:

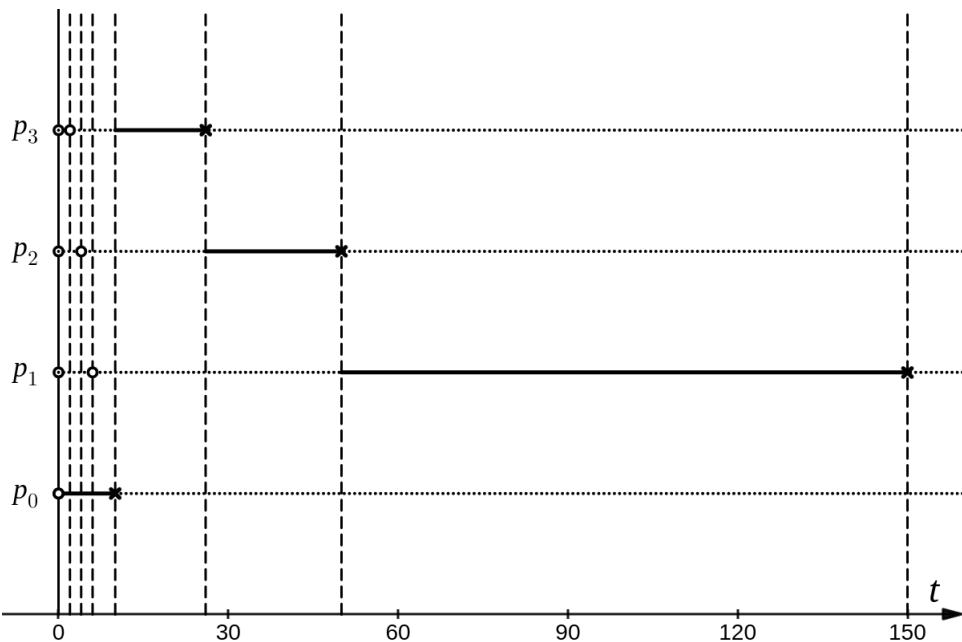
$$\tilde{t}_a = \frac{t_{a0} + t_{a1} + t_{a2} + t_{a3}}{4} = \frac{0 + 8 + 106 + 128}{4} = 60.5$$

Vediamo come questo comportamento può cambiare radicalmente cambiando l'ordine di richiesta dei processi. Poniamo infatti di avere la sequenza:

Processo	T richiesta	C esecuzione
p_0	0	10
p_1	6	100
p_2	4	24
p_3	2	16

dove semplicemente si è invertito l'ordine degli ultimi 3 processi.

Sul grafico avremo:



Questo risulta in un tempo medio di attesa di:

$$\tilde{t}_a = \frac{t_{a0} + t_{a1} + t_{a2} + t_{a3}}{4} = \frac{0 + 44 + 22 + 8}{4} = 18.5$$

Chiaramente molto meglio del caso precedente.

Abbiamo quindi che l'algoritmo è utile per sistemi batch, dove l'unica cosa che ci interessa è uso massimo della CPU (che ci assicura), ma largamente da evitare per sistemi interattivi, e soprattutto per sistemi real-time. Questo è vouto all'aleatorietà legata al momento della richiesta dei processi, che rende impossibile rispondere celermente o fare qualsiasi tipo di promessa sul tempo di turnaround.

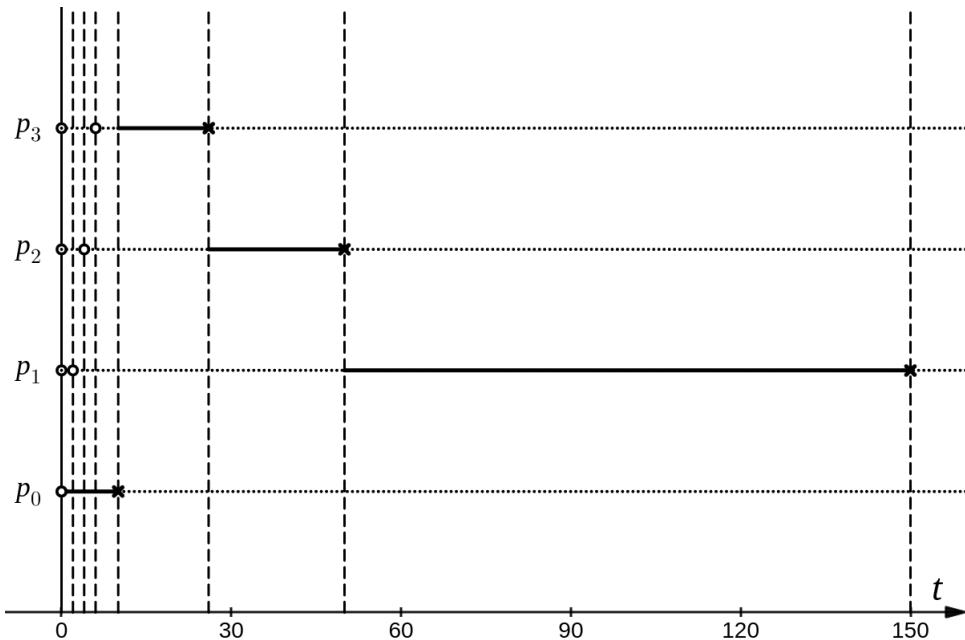
6.2.3 Algoritmo SJF

L'algoritmo **SJF** (*Shortest Job First*) implementa una **priorità statica**: si fa l'ipotesi di conoscere il **tempo di CPU** utilizzato da ogni processo, e assegnare priorità maggiori a processi con tempo di esecuzione minore. Di contro, non è preemptive, cioè una volta assegnata la CPU non può revocarla. Su come il S/O conosce il tempo di esecuzione non facciamo per adesso ipotesi.

Simuliamo anche questo algoritmo, usando la stessa tabella del primo caso in 6.2.2:

Processo	T richiesta	C esecuzione
p_0	0	10
p_1	2	100
p_2	4	24
p_3	6	16

Osserviamo che applicando l'algoritmo si ottiene il flusso di esecuzione:



Cioè esattamente quello che avevamo visto come caso ottimo del FCFS (che ricordiamo aveva tempo medio $\bar{T}_a = 18.5$), senza che i processi siano arrivati necessariamente nell'ordine ottimo.

Facciamo una nota sulla priorità statica: ad ogni chiamata dello scheduler questo può sapere solo i tempi di esecuzione dei processi attualmente in esecuzione, cioè si potrebbe mandare in esecuzione un processo con tempo di esecuzione maggiore quando ne entra in coda pronto ne entra uno con tempo minore. In questo caso, per la natura *non preemptive* dell'algoritmo, bisogna lasciare che questo esegua prima di mettere il nuovo arrivato in esecuzione.

Adoperando questo algoritmo si minimizza (nel senso matematicamente ottimo) il tempo di attesa medio dei processi, in quanto si cerca di arrivare il prima possibile al processo successivo (svolgendo adesso il più veloce). Uno svantaggio sarà chiaramente che i processi che dimostrano tempi di esecuzione lunghi verranno eseguiti sempre per ultimi.

6.2.4 Algoritmo SRTF

Abbiamo introdotto l'algoritmo **SRTF** (*Shortest Remaining Time First*) come una versione *preemptive* del SJF (in questo rimane comunque prioritario).

Visto che è preemptive, viene eseguito *ogni volta* che cambiano le condizioni di scelta (non soltanto quando la CPU è libera, come nel caso dei non preemptive, ma ogni che un nuovo processo entra in esecuzione). Può per questo motivo eseguire senza innescare cambi di contesto.

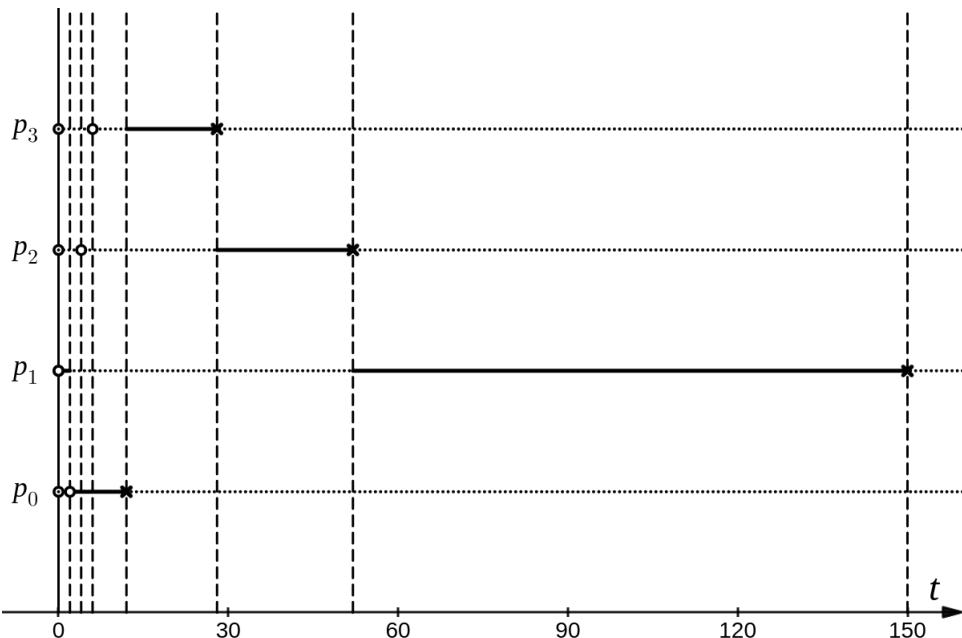
Dovremmo adattare la nostra ipotesi di conoscenza del tempo di CPU ad un'ipotesi di conoscenza del **tempo rimanente** per ogni processo: in questo caso se il processo appena entrato ha tempo rimanente minore di quello del processo attualmente in esecuzione, conviene sfruttare la preemption. Di nuovo, per adesso non facciamo assunzioni su come ricaviamo tale euristica.

L'unica considerazione che ci conviene fare è che aggiornare le previsioni temporali ad ogni evento che cambia le condizioni di scelta chiede allo scheduler di fare più conti, e quindi aumenta leggermente l'overhead O_v . Ampia letteratura dimostra che l'approccio è comunque conveniente.

Simuliamo questo algoritmo con la tabella, modificata dalle precedenti mandando per primo in esecuzione il processo p_1 , con tempo di esecuzione maggiore:

Processo	T richiesta	C esecuzione
p_0	2	10
p_1	0	100
p_2	4	24
p_3	6	16

Osserviamo che applicando l'algoritmo si ottiene il flusso di esecuzione:



Questo risulta in un tempo medio di attesa di:

$$\tilde{t}_a = \frac{t_{a0} + t_{a1} + t_{a2} + t_{a3}}{4} = \frac{2 + 0 + 28 + 12}{4} = 10.5$$

prima del primo burst.

Ciò che è importante in questo caso è che il processo p_1 viene sospeso con preemption per permettere l'esecuzione dei CPU burst di p_0 , p_2 e p_3 , molto più veloci. In questo modo il sistema risulta nel complesso molto più responsivo.

L'SRTF migliora la risposta in tempo reale del SJF, permettendo una riduzione sia dei tempi di turnaround che dei tempi di attesa medi in caso di richieste di esecuzione di processi non ottimali.

Un problema del SRTF, come avevamo visto nel SJF, è la **process starvation**: in genere, negli algoritmi in base prioritaria, si rischia che i processi a priorità minore (in questo caso quelli con tempo rimanente maggiore) non vengano mai serviti e rimangano a lungo in coda pronti, rendendo il sistema meno responsivo.

Anche qui la letteratura ci rende noto che la priorità dei processi in SRTF è **monotona crescente**: man mano che i processi eseguono, il loro tempo rimanente diminuisce e quindi la priorità aumenta. Questo effetto aiuta a ridurre la process starvation.

6.2.5 Stima dei tempi in SJF e SRTF

Chiariamo la questione di come si possono fare previsioni informate sul **tempo di esecuzione** (in SJF) e il **tempo rimanente** (in SRTF).

Un'approccio, preso ad esempio il caso del **tempo di esecuzione**, è quello di usare la tecnica della **esponenziale**. Si fa una stima iniziale s_i del tempo di burst t_i esimo. Preso un parametro a con $0 < a < 1$, si aggiorna ad ogni terminazione del processo (quindi facendo delle *osservazioni* per ogni esecuzione del processo) la stima come:

$$s_{n+1} = at_n + (1 - a)s_n$$

Presi $a \sim 0$ si ha che le stime deviano malvolentieri da quella iniziale, mentre con $a \sim 1$ si ha che le stime sono molto volubili rispetto alle osservazioni fatte.

Modelli statistici più complessi possono dare previsioni più accurate, sempre tenendo conto del fatto che lo scheduler deve eseguire con overhead $O_v \sim 0$, o almeno il più piccolo possibile.

Una volta noto il *tempo di esecuzione*, il **tempo rimanente** si può stimare considerando il tempo che il processo ha impiegato finora e sottraendolo dal tempo di esecuzione totale (se non rinunciando al tener conto se tale tempo è stato usato in CPU o I/O burst, purtroppo un'euristica è un'euristica).

7 Lezione del 08-10-25

Continuiamo la discussione degli algoritmi di scheduling.

7.0.1 Algoritmo RR

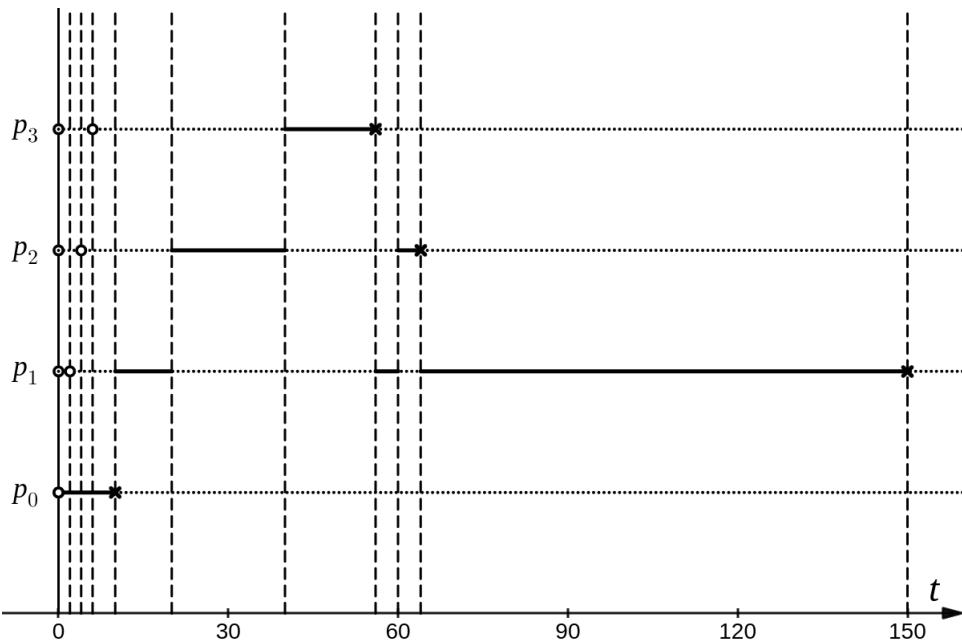
L'algoritmo **RR** (*Round Robin*) è preemptive e non prioritario, e si basa su un meccanismo molto semplice: ad ogni processo viene dato un quanto temporale prefissato, e via preemption si passa al processo successivo quando tale quanto viene esaurito.

Questo lo rende molto efficiente: l'overhead O_v è minimo, quasi al pari di FCFS (leggermente più alto in quanto i cambi di contesto sono più frequenti, uno ogni quanto temporale).

Simuliamo questo algoritmo usando la sequenza di processi richiesti già vista in 6.2.2, e imponendo un quanto temporale di $\Delta T = 20$. La tabella dei tempi di richiesta e esecuzione dei processi sarà:

Processo	T richiesta	C esecuzione
p_0	0	10
p_1	2	100
p_2	4	24
p_3	6	16

In questo caso la *timeline* avrà un aspetto del genere:



Vediamo che il tempo medio, calcolato come:

$$\tilde{t}_a = \frac{t_{a0} + t_{a1} + t_{a2} + t_{a3}}{4} = \frac{0 + 10 + 20 + 40}{4} = 17.5$$

sarà più o meno nella norma rispetto a SJR o STRF.

La cosa importante è, come lo era stato per STRF, la *responsività* che l'algoritmo riesce a realizzare: un processo non resta mai in coda pronti per un tempo maggiore al quanto temporale moltiplicato per i processi in esecuzione meno 1.

Dal punto di vista implementativo, organizzeremo i quanti temporali sfruttando una componente hardware detta **timer**: questo è considerato a tutti gli effetti una periferica, ed invia (dopo un'opportuna configurazione) interruzioni esterne periodiche. Ogni volta che il processore riceve tale interruzione, mette in esecuzione lo scheduler, che provvede a cambiare il contesto al prossimo processo. Chiaramente, lo scheduler può comunque essere messo in esecuzione da eventi comuni come la terminazione di processi.

Per realizzare l'esecuzione ciclica si usa una semplice coda pronti dove lo scheduler estreia sempre dalla testa e inserisce sempre in fondo alla coda.

Come abbiamo visto dall'esempio, l'RR rende molto semplice stimare il tempo di attesa: se ci sono N processi in esecuzione, $N - 1$ saranno in coda pronti in qualsiasi momento, quindi un dato processo aspetterà:

$$T_a = (N - 1)q$$

dove q è il quanto di tempo.

Chiaramente il T_a diventa troppo grande se ci sono troppi processi.

7.1 Code multilivello

Abbiamo quindi discusso i 4 algoritmi di scheduling fondamentali che avevamo introdotto in 6.2. Ognuno di questi ha pro e contro distinti, ed è più adeguato in una o un'altra situazione.

Per questo motivo nei sistemi operativi moderni si preferisce implementare lo scheduling attraverso **più algoritmi** di scheduling, che gestiscono ognuno la situazione che più conviene.

Un modo elegante di realizzare ciò è mantenere **più code**, una per ogni algoritmo di scheduling. Sistemi di questo tipo vengono detti **multilevel queue** o *code multilivello*.

7.1.1 Code di feedback

Una variante interessante delle code multilivello è dato dalle *code di feedback*. Queste nascono per gestire più gerarchie di processi dai requisiti di esecuzione diversi (I/O bound, più interattivi, e CPU bound, più lenti).

In questo caso si prevede una struttura di code, ad esempio come la seguente:

1. Coda RR *veloce*, con quanto $\Delta T = 10$;
2. Coda RR *più lenta*, con quanto $\Delta T = 20$;
3. Coda FCFS, la *più lenta*.

Le code vengono ordinate per priorità decrescente.

Un processo richiesto viene messo nella coda RR più veloce: se al momento della prima revoca CPU non è riuscito a terminare il suo primo CPU burst, viene spostato nella coda RR più lenta. La procedura si ripete finché il processo non è giudicato come *non interattivo* e spostato nella coda FCFS.

In questo modo si riescono a sviluppare sistemi che si *"adattano"* in qualche modo a diversi tipi di processi, scegliendo per ognuno la coda (e quindi la politica di schedulazione) più adatta.

Facciamo qualche altra considerazione: poniamo che ci sia un processo molto interattivo (magari il processo che si occupa di disegnare l'ambiente grafico) nella prima coda, e un processo CPU bound molto lento (magari un software di calcolo scientifico) in coda FCFS.

Avremo che il processo interattivo farà molti I/O burst (scrittura a video, lettura dati da mouse e tastiera, ecc...) e probabilmente i suoi CPU burst finiranno prima del quanto temporale offerto dalla coda RR. Sarà in questi istanti che il processo lento potrà entrare in esecuzione nella coda FCFS.

L'approccio non è ideale in quanto presenta alcuni difetti:

- Potrebbero verificarsi casi dove il processo in coda FCFS non riesce ad eseguire di fronte a una grande massa di processi veloci che arrivano in coda RR (*starvation*): per questo motivo gli S/O moderni implementano altri meccanismi, come *l'aging*;
- Un'altra problematica è data dal fatto che l'ultima coda è non preemptive: quando un processo dal lungo CPU Burst entra in coda FCFS vi resta finché non ha finito di eseguire, bloccando il resto del sistema.

Possiamo risolvere questo problema rendendo l'FCFS vagamente *preemptive*: visto che ci sono altre due code prima di essa, possiamo cogliere l'occasione del lancio di un nuovo processo per rimettere in esecuzione lo scheduler. Questo meccanismo non è propriamente necessaria per le prime 2 code in quanto il quanto di tempo è limitato (e quindi prima o poi il nuovo processo viene servito), mentre è fondamentale per la coda FCFS che potrebbe bloccare anche per diverso tempo.

Un ultima considerazione che vogliamo fare è se *conviene* riportare i processi dalle code di livello inferiore (più lente) nelle code di livello superiore (più veloci) quando queste si sgombrano: in generale, abbiamo che la letteratura non lo trova particolarmente vantaggioso. Questo perché un processo che è finito in una coda meno interattiva è probabilmente *meno interattivo*, per cui può godere di CPU burst più lunghi e deve restare nella coda dove si trova.

In caso di *aging*, di contro, questo processo è necessario e più che naturale: spostiamo i processi che sono da molto tempo in code inferiori verso le code superiori per forzarne l'esecuzione.

8 Lezione del 14-10-25

8.1 Schedulazione real-time

Veniamo quindi a come implementare la schedulazione nei sistemi in **tempo reale**. Avevamo detto che questi erano sistemi principalmente di tipo *embedded*, cioè incorporati, non general-purpose ma *special-purpose* atti a governare sistemi esterni (sistemi di controllo per veicoli, macchinari industriali, ecc...).

8.1.1 Esecuzione ciclica

Abbiamo che la caratteristica principale di sistemi di questo tipo è il tipo di periferiche con cui interagiscono: invece di periferiche multiple e variabili (come nei sistemi general-purpose), avremo un insieme fisso di dispositivi di ingresso (detti *sensori*) e di uscita (detti *attuatori*).

Questo porta ad un paradigma di esecuzione fortemente periodico: si campiona il sistema esterno attraverso i sensori, compie una qualche elaborazione, e aggiornano gli attuatori per rispondere a quanto rilevato.

Ciò significa che i processi messi in esecuzione devono rispettare il periodo dell'esecuzione ciclica, e produrre il loro risultato entro date *scadenze* date dal periodo corrente e il numero di altri processi in esecuzione. Occorre allora avere un controllo preciso e granulare sul tempo che impiegano a terminare.

8.1.2 Deadline

In questo caso prevederemo, dopo l'istante di richiesta r di un processo, una certa deadline d , calcolata come:

$$d = r + \Delta d$$

dove Δd è il tempo massimo di esecuzione del processo.

- In un sistema *soft real-time* si cerca di fare il possibile per assicurare che il processo termini prima di d ;
- In un sistema *hard real-time* la terminazione del processo prima di d è prerogativa dell'integrità dell'intero sistema.

In particolare, riguardo al paradigma di esecuzione ciclica accennato nello scorso paragrafo, avremo che per un processo che deve eseguire ciclicamente con periodo Δt , per ogni istante di richiesta r_i l'istante di richiesta successivo sarà calcolato come:

$$r_{i+1} = r_i + \Delta t$$

In questo caso sarà fondamentale rispettare la diseguaglianza:

$$d_i < r_{i+1} \Leftrightarrow \Delta d_i < \Delta t$$

data d_i come deadline dopo la richiesta r_i .

Estendo il concetto a sistemi multiprogrammati, avremo che dati n processi il periodo T di aggiornamento simultaneo di ogni processo in esecuzione sarà:

$$T = \text{MCM}(\Delta t_i)$$

con t_i i periodi di ogni processo: semplicemente si prende il minimo comune multiplo.

Ritornando all'idea dei CPU burst, se il processo si svolge in più CPU burst C_1, C_2, \dots dovrà quindi essere che:

$$T_e = \sum C_i < \Delta d_i$$

cioè che almeno il tempo di esecuzione del processo sia minore del tempo massimo di esecuzione per rispettare la deadline corrente.

8.1.3 Algoritmo RM

Iniziamo quindi a vedere alcuni algoritmi di scheduling per sistemi real-time. Il primo che vediamo è l'**RM** (*Rate Monotonic*). Questo consiste semplicemente ad assegnare una priorità statica *monotonica crescente* ai processi in base al *rate*, cioè la frequenza, del loro ciclo di esecuzione. Questo equivale ad assegnare una proprietà inversamente proporzionale al periodo t del processo:

$$p \propto \frac{1}{t} = f$$

Visto che la proprietà è statica, chiaramente l'algoritmo è non preemptive.

Facciamo quindi l'esempio dell'esecuzione dell'algoritmo, ipotizzando due processi p_a e p_b :

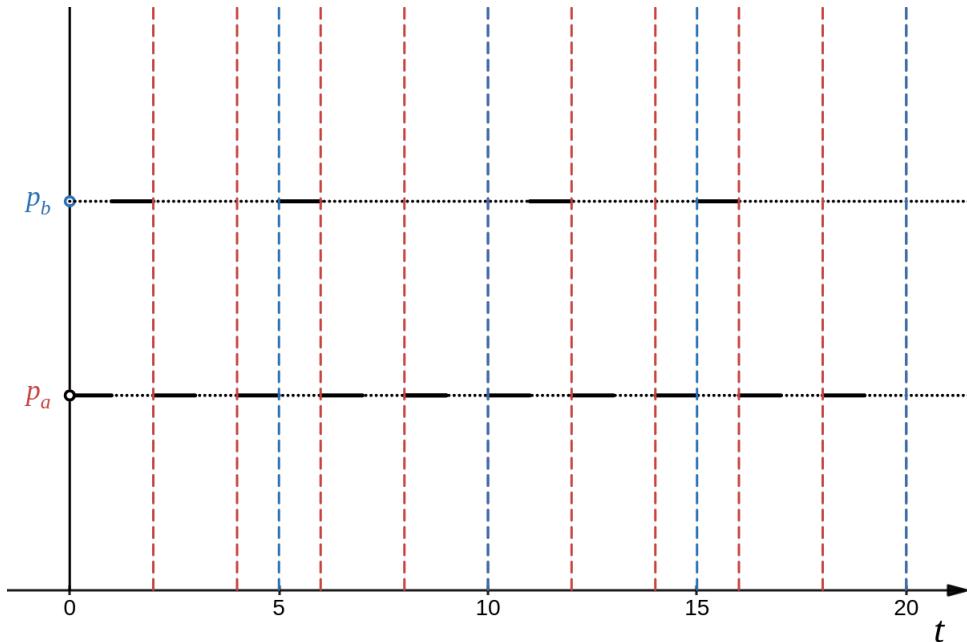
Processo	Δt periodo	C esecuzione
p_a	2	1
p_b	5	1

Da questo è fra l'altro immediato che, con $\Delta t_a = 2$ e $\Delta t_b = 5$, il periodo complessivo di sistema T è:

$$T = \text{MCM}(\Delta t_a, \Delta t_b) = \text{MCM}(2, 5) = 10$$

Vedremo come questo periodo determina anche il periodo dell'attività dello scheduler.

Simulando l'esecuzione si ha, colorando in rosso le deadline di p_a e in blu quelle di p_b :



Vediamo quindi come riusciamo a rispettare tutte le deadline. Un problema è che, ad esempio all'istante 10, si sono fatti 3 cicli da un unità temporale a vuoto, cioè l'efficienza E è:

$$E = \frac{10 - 3}{10} = 70\%$$

Questo non è immediatamente sbagliato: significa solo che il sistema ha abbastanza risorse da soddisfare ampiamente le richieste in arrivo. Potrebbe diventare un problema quando vogliamo *"stringere"* le temporizzazioni in modo da far fronte ad un maggior numero di processi, o processi con CPU burst più consistenti.

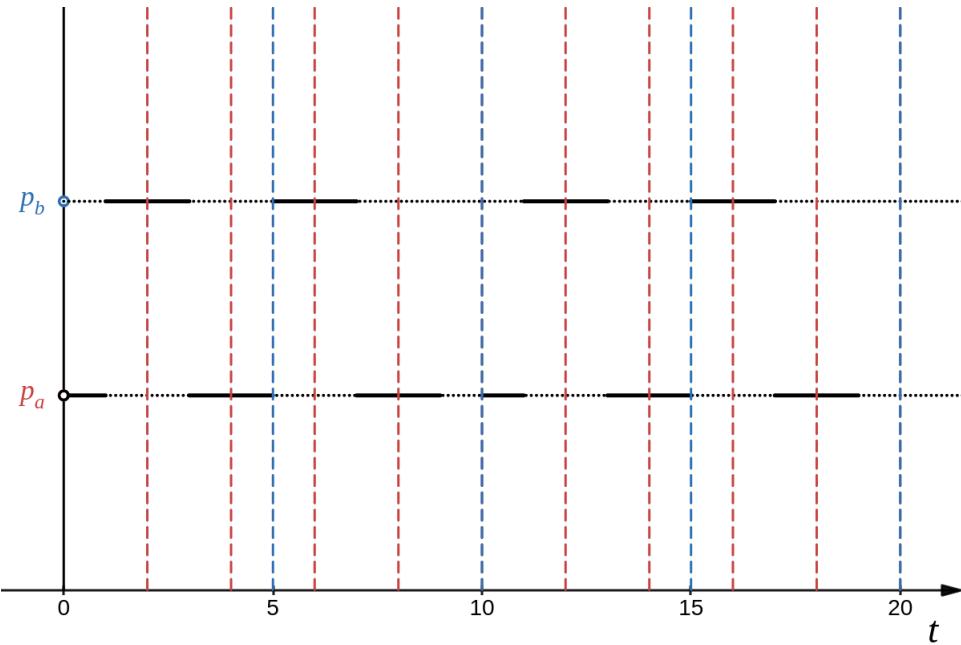
9 Lezione del 15-10-25

Continuiamo la discussione dell'algoritmo **RM** (*Rate Monotonic*).

Volevamo vedere gli effetti che si ottenevano quando si aumentava la pressione sulla CPU sfruttando questo algoritmo. Prendiamo allora gli stessi processi p_a e p_b della scorsa lezione, ma raddoppiamo il tempo di esecuzione del processo p_b :

Processo	Δt periodo	C esecuzione
p_a	2	1
p_b	5	2

Simulando l'esecuzione si ha, colorando le linee di periodo come nello scorso esempio:

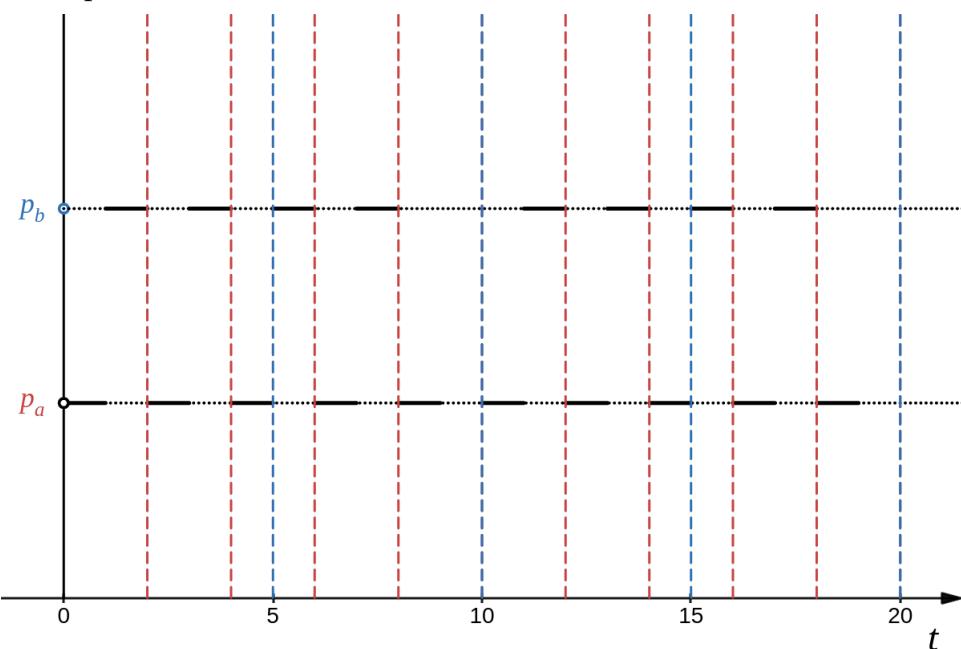


Vediamo quindi come siamo arrivati al 90% di utilizzo della CPU, e come il fatto che l'algoritmo è non preemptive significa che quando p_b accede alla CPU, la tiene anche oltre la linea di periodo di p_a (che ha comunque tempo di eseguire prima della linea successiva).

9.0.1 Algoritmo RM "preemptive"

Potremmo introdurre la preemption nell'algoritmo RM. In questo caso, ad ogni periodo riportiamo in esecuzione il processo con priorità più alta.

Vediamo quindi la timeline che otteniamo applicando questa versione con preemption all'esempio della scorsa sezione:



Notiamo che questo algoritmo di scheduling introduce un overhead maggiore della versione non preemptive, dato dai maggiori cambi di contesto. Inoltre, almeno in questo caso, non varia particolarmente in utilizzo CPU o in efficacia generale.

Comunque, possiamo osservare che mantiene i processi ancora più lontani dalla deadline, che è generalmente un comportamento desiderabile.

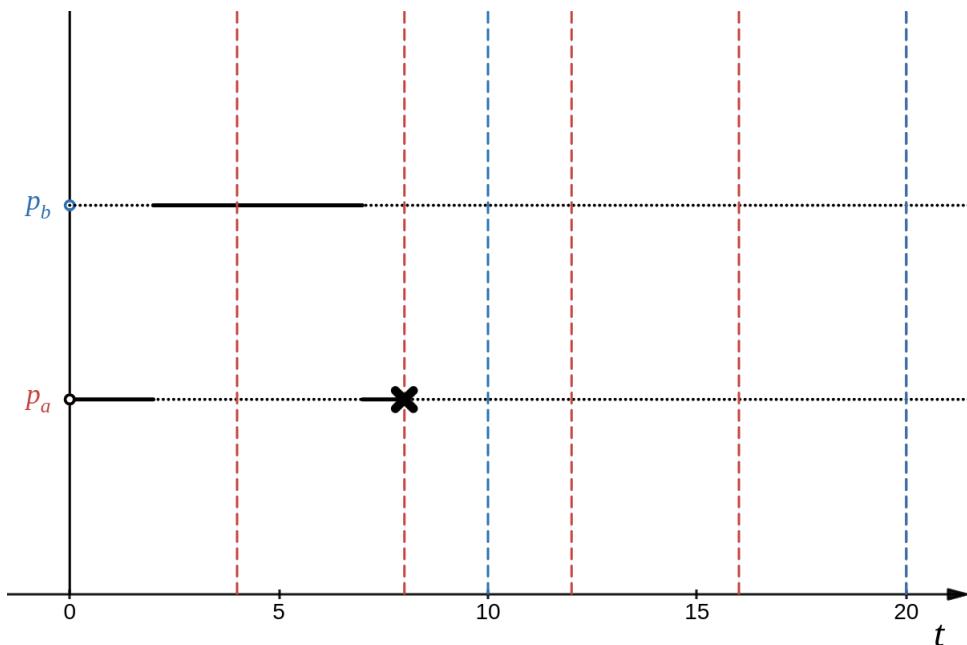
Abbiamo quindi che per gli algoritmi a priorità *statica*, il RM è ottimo: se un insieme di processi è schedulabile a priorità statica in real-time, allora lo è con RM. Di contro, se non è schedulabile con RM, non esiste nessun algoritmo a priorità statica che può schedularlo.

9.1 Processi non schedulabili staticamente

Approfondiamo cosa significa, per un insieme di processi, essere *schedulabili a priorità statica*. Prendiamo la tabella di processi:

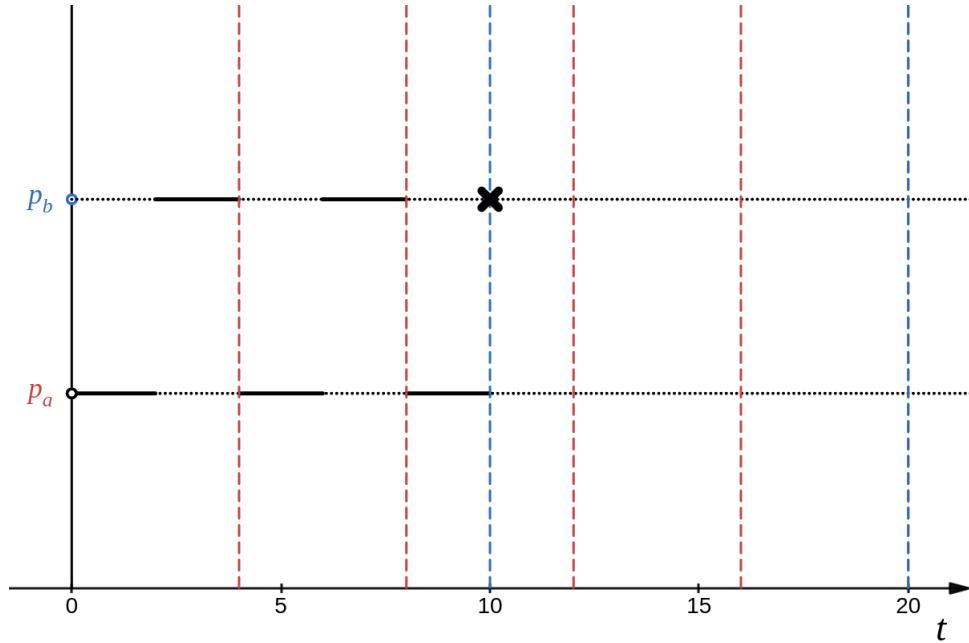
Processo	Δt periodo	C esecuzione
p_a	4	2
p_b	10	5

Vediamo come ogni processo chiede di eseguire per metà del suo periodo. Se usiamo la schedulazione RM in questo caso, otteniamo:



Dove all'istante $t = 8$ non siamo riusciti a completare il CPU burst di p_a , e si è quindi giunti ad un *overflow*: un fallimento della schedulazione che non ha rispettato la deadline.

Pensiamo allora di utilizzare l'algoritmo RM "preemptive" visto nella scorsa szione. In questo caso, si avrà:



A questo punto è p_b ad andare in overflow! All'istante $t = 10$ infatti non siamo riusciti a completare le sue 5 unità temporali per completare l'esecuzione.

Abbiamo chiaramente incontrato un insieme di processi non schedulabili con priorità statica, e nemmeno introducendo la preemption abbiamo risolto il problema: dovremo trovare una qualche altra soluzione.

9.1.1 Trattazione matematica

Abbiamo che, nel caso dei due processi p_a e p_b , il minimo che dobbiamo rispettare per poter in primo luogo eseguire i processi nel periodo di sistema è:

$$n_a C_a + n_b C_b \leq T$$

dove ricordiamo T è il m.c.m. fra i periodi t_a e t_b , e n_a e n_b sono rispettivamente il numero di volte in cui i processi p_a e p_b entrano in esecuzione per periodo di sistema. In particolare, questi valori si possono calcolare dai periodi dei processi Δt_a , Δt_b , come:

$$n_a = \frac{T}{\Delta t_a}, \quad n_b = \frac{T}{\Delta t_b}$$

Sostituendo, si ha quindi:

$$\frac{T}{\Delta t_a} C_a + \frac{T}{\Delta t_b} C_b \leq T \implies \frac{C_a}{\Delta t_a} + \frac{C_b}{\Delta t_b} \leq 1$$

Possiamo quindi generalizzare quanto trovato alla (ovvia) legge:

$$U = \sum_{i=0}^{n-1} \frac{C_i}{T_i} \leq 1$$

per n processi arbitrari, dove U viene detto **fattore di utilizzazione**.

Nell'esempio considerato finora, questo valore è:

$$U = \frac{2}{4} + \frac{5}{10} = 1$$

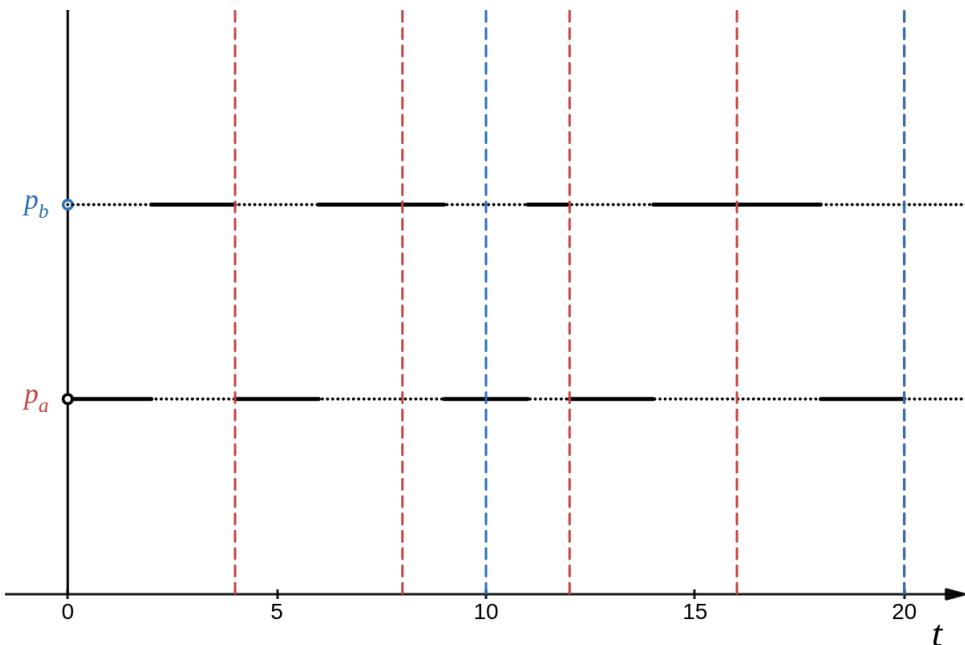
per cui i processi sono schedulabili. Ci manca da trovare un algoritmo che li sappia schedulare.

9.2 Algoritmo EDF

L'algoritmo **EDF** (*Earliest Deadline First*) è un algoritmo di schedulazione real-time, preemptive e a priorità dinamica.

Consiste nel mandare in esecuzione il processo che è più vicino alla sua deadline. Come in tutti gli algoritmi a priorità dinamica, lo scheduler viene messo in esecuzione al cambio dei criteri di scelta (quindi quando un processo entra in coda pronti). In ogni caso, se due processi so trovano ugualmente vicini alla deadline al momento dell'esecuzione dello scheduler, si opta per ridurre i cambi di contesto al minimo e mantenere in esecuzione quello che sta già eseguendo.

Vediamo come questo algoritmo si applica all'esempio schedulabile visto finora:



Vediamo come otteniamo il 100% dell'utilizzazione CPU, e riusciamo a schedulare i processi senza overflow.

All'istante $t = 8$, infatti, il processo p_b è più vicino di p_a alla deadline, e quindi viene mantenuto in esecuzione (fino a $t = 9$ dove termina). In questo modo si riesce ad evitare che il suo CPU burst venga "tagliato" prima che esso possa rispettare la sua deadline.

Abbiamo quindi trovato un'algoritmo che risolve i problemi che avevamo incontrato con RM: possiamo anticipare che questo algoritmo è ottimo fra gli algoritmi di schedulazione in real-time a priorità dinamica.

Una considerazione può essere fatta sull'overhead che introduciamo, almeno per l'esempio sopra. Abbiamo detto che sì, si cerca di mantenere al minimo i cambi di contesto, ma c'è comunque un certo overhead dato dall'esecuzione dello scheduler ad ogni creazione di processo.

In questo, potremmo aggiornare il modello introdotto in 9.1.1 come segue:

$$U = \dots \leq 1 - O_v$$

dove O_v è un fattore temporale che tiene conto dell'overhead.

9.3 Thread

Introduciamo semplicemente il concetto di **thread** (o *processo leggero*). Avevamo detto che un processo è al contempo:

- Un elemento che possiede delle *risorse*;
- Un elemento a cui viene *assegnata* la CPU (si conserva lo *stato* e si usa uno *scheduler* per decidere quando caricarlo).

Possiamo separare questi due aspetti:

- Definiamo **processo leggero**, o *thread*, l'elemento a cui viene assegnata la CPU;
- Di contro, definiamo **processo pesante**, o *task*, l'elemento che possiede le risorse.

Un processo pesante può essere composto da più thread, ognuno dei quali rappresenta effettivamente un flusso di esecuzione a sé stante. Tutti i thread possono però accedere alle risorse del loro processo pesante (incluso *spazio di indirizzamento*, file aperti, ecc...).

10 Lezione del 21-10-25

10.1 Tassonomia di Flynn

Prima di venire alla sincronizzazione dei processi, vediamo brevemente la **classificazione delle architetture** attraverso la *tassonomia di Flynn*.

Questa è una classificazione che vede un sistema di elaborazione da 2 punti di vista ortogonali:

- La capacità di avere più flussi di **esecuzione**: si possono distinguere **SI** (*Single Instruction Stream*) e **MI** (*Multiple Instruction Stream*). Questo concetto è vicino a quello di *thread* visto nella scorsa sezione;
- La capacità di avere più flussi di **dati**: si possono distinguere **SD** (*Single Data Stream*) e **MD** (*Multiple Data Stream*);

Abbiamo quindi la prima distinzione:

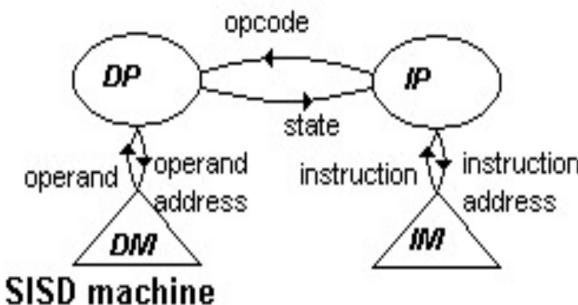
	SI (Single Instruction Stream)	MI (Multiple Instruction Stream)
SD (Single Data Stream)	Macchine SISD	Macchine MISD
MD (Multiple Data Stream)	Macchine SIMD	Macchine MIMD

Iniziamo a vedere dove si collocano le macchine che conosciamo.

10.1.1 Macchine SISD

Le macchine SISD rappresentano le tradizionali macchine *sequenziali* e *monoprocessoressi* definite dall'architettura di Von Neumann. In questo caso si ha un solo flusso di istruzioni, ciascuna agente su al più un flusso dati, e ad ogni istante temporale si esegue una singola istruzione.

Vediamo una schematizzazione di questa architettura coerente con Flynn:



Da questa schematizzazione notiamo:

- Un'unità di elaborazione **dati**, detta **DP** (*Data Processor*), che interagisce ottenendo operandi e fornendo indirizzi di operandi (e sperabilmente dati) con uno stream **dati**, detto **DM** (*Data Memory*);
- Un'unità di elaborazione **istruzioni**, detta **IP** (*Instruction Processor*), che interagisce ottenendo istruzioni e fornendo indirizzi di operazioni con uno stream **istruzioni**, detto **IM** (*Instruction Memory*).

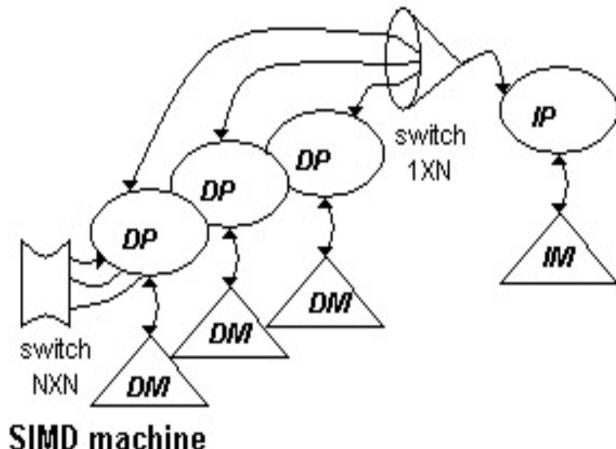
IP e DP interagiscono scambiandosi codifiche di istruzioni ($IP \rightarrow DP$) e variazioni di stato (della memoria dati, $DP \rightarrow IP$).

Possiamo notare che la cosiddetta *architettura Harvard* è un architettura che prevede, come dalla schematizzazione sopra, una forte separazione fra stream istruzioni e dati (di contro alla Von Neumann, che prevede un'unica fonte di memoria per dati e istruzioni). Entrambe le architetture sono classificate come SISD, la Harvard è più usata in sistemi real-time mentre la Von Neumann è ancora oggi più usata nei sistemi general purpose.

10.1.2 Macchine SIMD

Le macchine SIMD sono formate da unità di elaborazione multiple, che eseguono le stesse istruzioni *contemporaneamente*, ma su flussi di dati differenti.

Una schematizzazione simile a quella sopra riportata è la seguente:



Abbiamo che questa moltiplicazione dei flussi dati su cui si elabora si ha moltiplicando le unità di elaborazione dati (i **DP**), facendole obbedire ad una singola unità di elaborazione istruzioni (l'**IP**). Vediamo innanzitutto come si realizza la sincronizzazione fra questi DP: si prevede uno *switch* $1 \times N$ che porta le codifiche di istruzioni dall'**IP** a tutti i **DP**.

Per l'interazione fra le **DP** prevediamo poi uno *switch* $N \times N$ che le collega. Chiaramente questo *switch* sarà inefficiente, e vorremo usarlo il meno possibile.

Architetture di questo tipo possono essere *regolari* o create *ad hoc* sulla base della struttura del problema: nel caso di architetture regolari (cioè che rispettano la topologia fisica) non si hanno conflitti, e questo le rende efficienti e poco costose.

Le applicazioni di un'architettura di questo tipo sono nel caso di operazioni fortemente vettorizzate, come ad esempio nelle applicazioni grafiche e multimediali. Inoltre, questo è il tipo di architettura che incontriamo spesso per macchine che devono portare avanti moli massiccie di computazione come i *supercomputer*.

Riassumendo, possiamo dire che l'architettura SIMD prevede 2 tipi di parallelismo:

- **Parallelismo temporale:** c'è un meccanismo di *pipeline*, cioè fasi diverse di un'unica istruzione sono eseguite in parallelo in differenti moduli connessi in cascata.
- **Parallelismo spaziale:** i medesimi passi sono eseguiti contemporaneamente su un array di processori perfettamente uguali, sincronizzati da un solo controllore.

Lato programmatore possiamo prevedere due paradigmi per la compilazione di programmi pensati per l'esecuzione su macchine SIMD:

- Il primo modo è non riscrivere il codice, sapendo che un programma pensato come scalare su macchina sequenziale (SISD), in esecuzione su macchina SIMD sarà vettoriale.

Avremo quindi che, ad esempio:

```

1 // somma scalari
2 c = a + b

```

su macchina SIMD diventerà:

```

1 // soma vettori (!)
2 C = A + B

```

- Nel caso si voglia essere più esplicativi nel tipo di operazioni che facciamo, possiamo implementare un **compilatore vettoriale**: l'idea è che questo riconosca automaticamente quando le istruzioni SIMD potrebbero tornare utili per parallelizzare delle operazioni vettoriale, e inserisca quindi le operazioni necessarie.

Ad esempio, potremmo volere:

```

1 for(int i = 0; i < 100; i++) {
2     c[i] = a[i] + b[i];
3 }
4 // qui riconosciamo che il ciclo e' vettorizzabile, e quindi lo
   vettorizziamo

```

10.1.3 Macchine MISD

In una macchina MISD vogliamo avere più flussi di istruzioni che lavorano contemporaneamente su un unico flusso dati.

Abbiamo che questa categoria è sostanzialmente vuota: il parallelismo fra più istruzioni in esecuzione sullo stesso flusso dati si ha effettivamente nei processori moderni solo attraverso il meccanismo della **pipeline**.

Se prevediamo che il processore debba svolgere per ogni istruzione più fasi, fra cui ad esempio:

1. Prelievo istruzione;
2. Decodifica;
3. Prelievo operandi;
4. Esecuzione;
5. Scrittura.

Avremo che questo potrà, disponendo di più unità di elaborazione atte a completare ognuna di queste fasi, parallelizzare come segue:

	Prelievo istruzione	Decodifica	Prelievo operandi	Esecuzione	Scrittura
t_0	i				
t_1	$i + 1$	i			
t_2	$i + 2$	$i + 1$	i		
t_3	$i + 3$	$i + 2$	$i + 1$	i	
t_4	$i + 4$	$i + 3$	$i + 2$	$i + 1$	i

In questo, a regime (nella tabella tempo t_4) si avranno più di un'unità di elaborazione a lavoro contemporaneamente, e potremmo dire di aver realizzato in qualche modo il paradigma MISD.

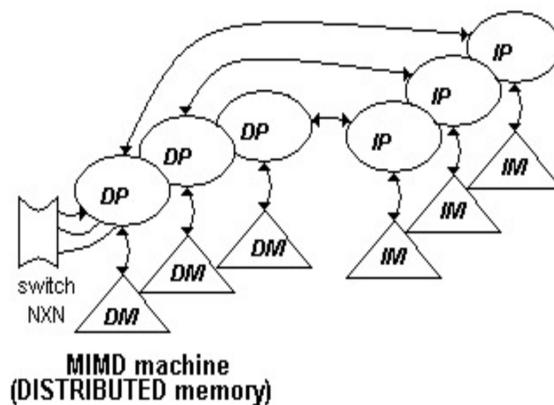
Facciamo tra l'altro una nota sull'efficienza: se l'esecuzione di un'istruzione senza pipeline richiedeva un tempo Δt , e il throughput era $\frac{1}{T}$, prevedendo una pipeline ad n stadi si riesce ad arrivare a $\frac{1}{T} \times n$ (assunto che ogni stadio richieda lo stesso tempo e si riesca a ridurre l'overhead dato da n troppo grandi).

10.1.4 Macchine MIMD

Le macchine MIMD rappresentano per noi la categoria più interessante più flussi di istruzioni sono in esecuzione contemporaneamente su più processori, elaborando insieme di dati distinti, privati o condivisi.

Ne prevediamo due tipologie principali:

- **DM-MIMD** (*Distributed Memory MIMD*), cioè a *memoria distribuita*. Queste si schematizzano come segue:



In questo caso abbiamo più coppie IP-DP (con relative memorie IM e DM), che rappresentano sostanzialmente più macchine SISD. Uno switch $N \times N$ permette quindi la comunicazione fra le unità.

Abbiamo quindi che tra i nodi non esiste memoria condivisa e ogni nodo esegue indipendentemente un flusso di istruzioni su un differente insieme di dati, memorizzati su spazi differenti. La comunicazione è realizzata mediante una sottorete dedicata (appunto, lo switch).

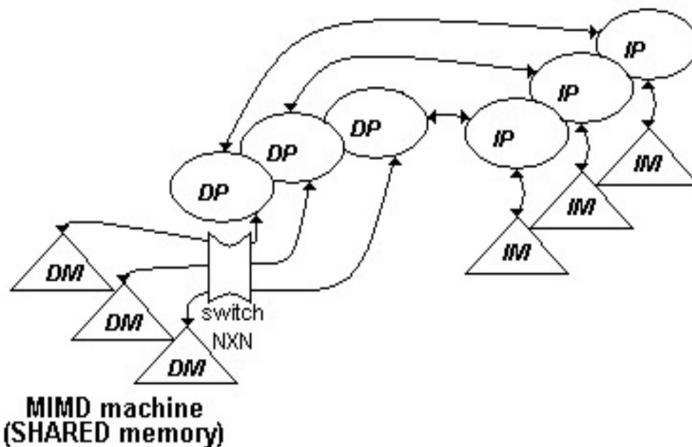
Applicazioni di questa architettura sono ad esempio una qualsiasi *rete di calcolatori* (come sia sviluppato lo switch non è specificato dal modello, e per noi potrebbe essere anche un router internet). Questo è il caso dei *cluster* di workstation, realizzati solitamente attraverso *Ethernet*. In questo richiediamo dalle macchine che compongono la rete 2 caratteristiche principali:

1. **High-availability**: in caso di guasti, la computazione può migrare da un nodo all'altro;
2. **Load-balancing**: i task da eseguire sono allocati nei nodi che hanno il minor carico.

Architetture MIMD più stabili sono poi le reti di interconnessione regolari e dirette (iper cubi, mesh, torus), attraverso cui i nodi si scambiano informazioni secondo il paradigma del *message passing* ("scambio di messaggi"). Queste macchine sono molto scalabili e si prestano bene ad algoritmi ad elevata località: sono stati costruiti cluster composti anche da milioni di unità sequenziali.

Una variante del DM-MIMD è il **DM-MIMD MPP** (*Massively Parallel Processing*). Questo è un paradigma utile in applicazioni scientifiche e particolari contesti di calcolo commerciale-finanziario. In un sistema MPP si ha:

- Migliaia di nodi (CPU standard, ognuna con la propria memoria e la propria copia del SO)
- Una rete di interconnessione custom molto potente (larga banda e bassa latenza). Affinché l'elaborazione MPP dia effettivi vantaggi occorre disporre di software capace di partizionare il lavoro e i dati su cui opera tra i vari processori.
- **SM-MIMD** (*Single Memory MIMD*), cioè a *memoria condivisa*. Queste si schematizzano come segue:



Sono quindi macchine sempre *multiprocessore*, ma dove le varie unità di elaborazione dati comunicano attraverso uno switch $N \times N$ con un *pool* unico di memoria (formato anche da più flussi di memoria, ma visti allo stesso modo da ogni unità di elaborazione dati).

Questo approccio è meno scalabile, realizza la comunicazione fra processori condividendo aree di memoria, e richiede una rete di interconnessione (lo switch $N \times N$ estremamente efficiente). Vogliamo che il numero N di processori sia piccolo ($N < 100$), affiché si possa avere stretto accompiamento fra i nodi. Si incorre chiaramente in problemi di competizione fra unità di elaborazione (mutua esclusione e sincronizzazione) che potrebbero impattare le prestazioni.

10.1.5 Confronto fra SIMD e MIMD

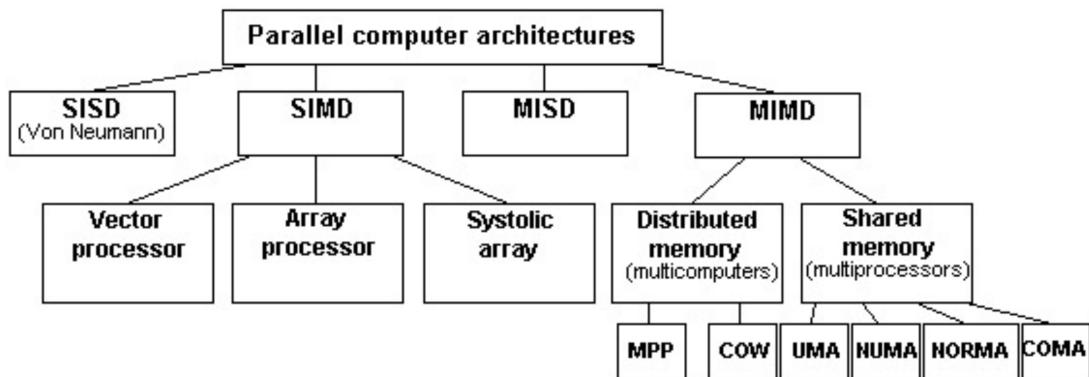
Possiamo fare quindi un confronto fra le architetture apperentemente simili, SIMD e MIMD:

- Le SIMD richiedono meno hardware delle MIMD (c'è un'unica unità di controllo, o *elaborazione istruzioni*);
- Le MIMD usano spesso processori general-purpose, quindi costano meno delle SIMD (che richiedono processori o ISA particolari, quindi meno hardware ma più specializzato);
- Le SIMD usano meno memoria delle MIMD (una sola copia del programma in memoria);

- Le MIMD godono di una grande flessibilità in termini di modelli computazionali supportati (si pensi *client-server*, *P2P*, ecc...). Di contro, è piuttosto semplice modificare un programma sequenziale perché esegua su architettura SIMD in maniera vettorizzata.

10.1.6 Sintesi della tassonomia di Flynn

Possiamo quindi vedere un grafico riassuntivo delle tassonomie viste:



Dove si notano le 4 categorie principali (SISD, SIMD, MISD e MIMD) ed alcune sottocategorie (non abbiamo parlato di tutte le sottocategorie, per una trattazione più completa si rimanda alla letteratura).

10.2 Tipologie di interconnessione

Dopo aver discusso le architetture descritte dalla tassonomia di Flynn, vediamo i tipi principali di **interconnessione** che si possono avere fra unità di elaborazione (o in generale nodi).

10.2.1 Bus



Il **bus** è la più semplice rete di interconnessione che abbiamo visto. Rappresenta una configurazione semplice ed affidabile (a meno che non si rompa il bus tutto funziona), dove ogni nodo a *grado 1* (tutti i nodi sono direttamente connessi al bus e a nient'altro), e il *diametro* è 1 (la distanza massima è data dal solo bus). Il numero totale di *link* di cui abbiamo bisogno è sempre 1: il bus stesso è l'unico link di cui necessitiamo (i nodi devono solo collegarsi a tale link).

Il problema è chiaramente la *competizione* sull'accesso al mezzo, che è massima: si hanno spesso problemi di mutua esclusione sulle stesse risorse, ad esempio quando più nodi vogliono accedere alla stessa risorsa contemporaneamente e devono farlo attraverso un unico bus.

10.2.2 Array lineare

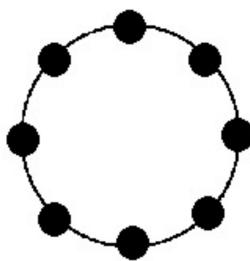


L'**array lineare** espande in qualche modo l'idea del bus: in un certo senso vogliamo partizionare il bus in tanti link nodo a nodo. In questo caso il grado del primo e dell'ultimo nodo sarà 1, mentre quello dei restanti nodi sarà 2. Il diametro sarà $n - 1$ per n nodi, e il numero totale di link $n - 1$ (quelli necessari a legare ogni nodo con i nodi adiacenti).

Il vantaggio di questo approccio è la competizione, che viene ridotta al minimo (ogni coppia adiacente di nodi può comunicare indipendentemente dagli altri). Più nello specifico, nel caso ideale possiamo avere fino a $\frac{N}{2}$ competizioni contemporanee e parallele (appunto, una per ogni coppia).

I nodi dovranno quindi fornire servizi di *routing*, cioè permettere a nodi da un capo dell'array di comunicare con nodi dall'altro capo, prendendosi a carico in qualche modo il messaggio da comunicare. Questo è chiaramente a scapito della robustezza: se un nodo si rompe due parti dell'array lineare rimangono separate.

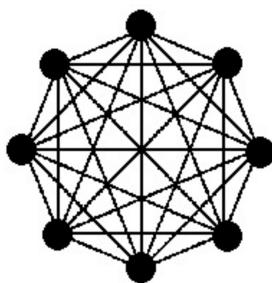
10.2.3 Ring



Il **ring** è un array lineare chiuso su stesso (dove si sono collegati i nodi estremi, cioè quelli con grado 1). Il grado diventa quindi 2 per tutti i nodi. Il diametro subisce la prima caratteristica fondamentale del ring: per raggiungere un dato nodo si hanno a disposizione due direzioni anziché una, per cui il diametro complessivo è $\frac{N}{2}$.

Anche la tolleranza ai guasti migliora, in quanto un nodo guasto non pregiudica necessariamente l'integrità del sistema (ne servono 2 per isolare una parte della rete).

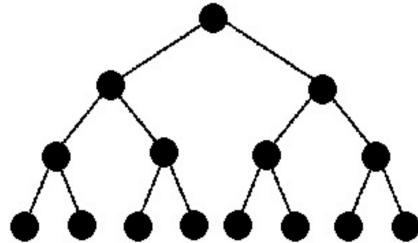
10.2.4 Connessione completa



La **connessione completa** (o *tutti-a-tutti*) è la rete di interconnessione più, appunto, *connessa*, che vediamo. In questo caso ogni nodo comunica direttamente con ogni altro nodo: il grado è per tutti i nodi $N - 1$ e il diametro è 1 (si arriva direttamente al nodo desiderato). Svantaggioso è chiaramente il numero totale di link, che cresce come $N^{\frac{N-1}{2}}$: l'approccio chiaramente non è scalabile!

Dobbiamo quindi trovare modelli per reti di interconnessione che presentino parte dei vantaggi della connessione completa (alta tolleranza ai guasti, bassissimo diametro), riducendo però il numero di link e quindi aumentando la scalabilità.

10.2.5 Albero binario



Si può pensare di ordinare i nodi secondo un **albero binario**. In questo caso vorremo definire l'*altezza* dell'albero come $h = \log_2(N)$ per N nodi e i gradi come:

- 2 per la radice;
- 1 per le foglie;
- 3 per tutti gli altri nodi.

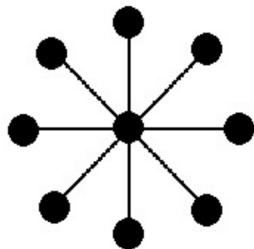
Il diametro sarà (senza dimostrazione) $2 \times (h - 1)$ e il numero totale di link $N - 1$ ($O(N)$ anziché $O(N^2)$, già migliore).

Questo tipo di rete permette una facile comunicazione fra nodi sugli stessi sottoalberi, mentre per comunicazioni fra sottoalberi distinti porta alla *congestione dei rami alti*: questo la rende poco scalabile. In particolare, più ci avviciniamo alla radice minori saranno i link (e quindi maggiore il carico sul singolo link). Inoltre, i nodi dovranno fare da router, e quindi più ci avviciniamo alla radice più i nodi hanno responsabilità di router sempre maggiori. Questo culmina sulla radice stessa, che chiaramente è sottoposta ad un carico non indifferente e rappresenta il punto più debole dell'architettura ad albero binario.

Per quanto riguarda la tolleranza ai guasti, vale lo stesso discorso: più in alto (verso la radice) avviene il guasto, maggiori sono le conseguenze per il sistema. Come caso limite, se si guasta la radice si incorre in un partizionamento in due dell'intero sistema.

Soluzioni alternative si possono avere sfruttando alberi, anziche binari, *n-ari*, cioè con n figli per nodo.

10.2.6 Stella



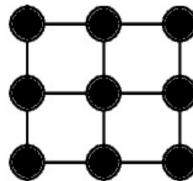
Le reti di interconnessione a **stella** prevedono un singolo nodo centrale che connette $n - 1$ nodi periferici. Il grado di tale nodo centrale sarà quindi $n - 1$, mentre i nodi periferici

avranno 1. Il diametro sarà 2 (dobbiamo passare sempre dal nodo centrale, a meno che non si voglia parlare col nodo centrale stesso). Il numero di link è ridotto ($N - 1$), e quindi da questo punto di vista il sistema è vantaggioso.

Il difetto più grande è chiaramente la presenza di un singolo nodo centralizzato soggetto a guasti o sovraccarichi. Questo è il classico problema del *single point of failure* delle architetture client-server: possiamo infatti intendere il nodo centrale come un *server* e i nodi periferici come *client* di tale server.

Abbiamo quindi che per quanto si possa rendere potente il nodo server, questo dovrà portare tutto il carico della rete (bassa scalabilità), e un guasto del server rappresenterà una mancanza di servizio per tutti i nodi client (bassa robustezza).

10.2.7 Mesh bidimensionale



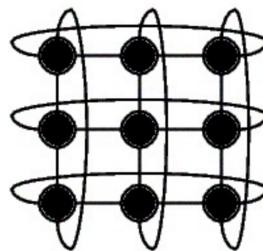
La struttura a **mesh bidimensionale** prevede di disporre i nodi su una griglia bidimensionale. In questo caso il lato della griglia sarà $r = \sqrt{N}$ e il grado dei nodi sarà:

- 2 per i nodi ai vertici,
- 3 per i nodi "centrali" ai lati (diciamo nodi agli *spigoli*);
- 4 per tutti gli altri nodi.

Il diametro sarà $2 \times (r - 1)$, e il numero totale di link $(2 \times (n - 2) \times r)$.

La resistenza ai guasti di queste configurazioni è buona, ma può essere migliorata. Vediamo come.

10.2.8 Toro bidimensionale



Collegando i nodi ai lati opposti di una struttura a mesh bidimensionale si ottiene una struttura a **toro bidimensionale**. In questo caso il grado è 4 per tutti i nodi, ma il diametro migliora: va sostanzialmente come r (dove r è il lato calcolato come $r = \sqrt{N}$, uguale alla rete bidimensionale). Il numero totale di link è invece stabile a $2N$.

Questa topologia risulta ben scalabile e notevolmente resistente ai guasti: sostanzialmente rappresenta per la mesh bidimensionale quello che il ring rappresenta per l'array lineare.

Architetture come le ultime viste (mesh, tori, ecc...) possono essere estese a più dimensioni, portando a *iper cubi*, *iper tori*, ecc...

10.3 Metriche di prestazione

Finiamo questa sezione del programma parlando di alcune **metriche** per le **prestazioni** di architetture descritte secondo la tassonomia di Flynn.

- Chiamiamo **speed-up** (S) il rapporto fra il tempo di esecuzione sequenziale sul tempo di esecuzione parallelo (SIMD o MIMD, cioè):

$$S = \frac{T_1}{T_N}$$

Questa metrica rappresenta quindi il *guadagno* in velocità che si ha passando ad un sistema multiprocessore. Idealmente vorremmo $S = N$, cioè *lineare* col numero di unità di elaborazione, ma come vedremo a causa dell'overhead introdotto da più processori dobbiamo accontentarci di $S < N$, con $S \approx N$.

Il valore dello speed-up dipende dalle applicazioni, ma anche dall'architettura: nelle SIMD spesso $S \approx N$, mentre nelle MIMD è difficile far crescere S (non è facile far lavorare pienamente tutte le CPU, o trasferire efficientemente fra di esse).

- L'**efficienza** (E) è invece definita come lo speed-up sul numero di unità di elaborazione, cioè:

$$E = \frac{S}{N}$$

Come prima, vorremmo $E = 1$, ma dobbiamo accontentarci di $E < 1$, con $E \approx 1$.

Per giustificare come mai non si può mai avere $S = N$ (o equivalentemente, $E = 1$), ci viene incontro la *legge di Amdhal*. Questa dice che un parallelismo "perfetto" (nelle varie attività compiute da un calcolatore) non è mai raggiungibile in quanto saranno *sempre* presenti sequenze di programmi *intrinsecamente seriali*. Semplificando, si può dire che ci saranno sempre degli intervalli di tempo sequenziale (T_{seq}), impiegati ad eseguire istruzioni non parallelizzabili come operazioni di I/O, costrutti condizionali, algoritmi intrinsecamente sequenziali, ecc....

Più nello specifico, la legge di Amdahl ridefinisce lo speed-up come:

$$S = \frac{T_1}{T_{\text{seq}} + \frac{T_1 - T_{\text{seq}}}{N}}$$

cioè il guadagno di speed-up è dato solo dalle parti parallelizzabili del programma ($T_1 - T_{\text{seq}}$), mentre le parti seriali (T_{seq}) non hanno grandi guadagni.

Il problema è che, prendendo il limite per $N \rightarrow \infty$, si ha:

$$\lim_{N \rightarrow \infty} S = \lim_{N \rightarrow \infty} \frac{T_1}{T_{\text{seq}} + \frac{T_1 - T_{\text{seq}}}{N}} = \frac{T_1}{T_{\text{seq}}}$$

cioè a dominare lo speed-up sono le parti sequenziali (le parti non sequenziali vengono abbattute velocemente, quelle sequenziali mai e rimangono come overhead complessivo dell'intero sistema).

10.3.1 Multitasking

In conclusione, vogliamo dire che il *multitasking* è quasi sempre vantaggioso (anche intuitivamente), ed è di notevole importanza anche nelle macchine per mantenere alto lo sfruttamento delle CPU (ciò che avevamo chiamato *efficienza CPU*).

Il vincolo che vogliamo rispettare sarà, quindi, di base:

$$P >> N$$

cioè avere più processi che unità di elaborazione.

10.4 Sincronizzazione processi

Iniziamo quindi a vedere la **sincronizzazione** fra processi, in particolare con riferimento ai *tipi* di interazione che ci possono essere fra processi, e i problemi di **mutua esclusione** e **sincronizzazione**.

10.4.1 Tipi di interazione

Ricordiamo che i processi possono interagire fra di loro secondo 2 modalità:

- **Cooperazione:** quindi per sincronizzazione diretta o esplicita, cioè definita dai programmi;
- **Competizione:** quindi per sincronizzazione indiretta o implicita, non definita dal codice dei programmi ma causata da tentativi di accesso *simultaneo* a risorse limitate.

Nominiamo poi l'**interferenza**, rappresentata da errori dipendenti dal tempo.

Per l'interazione fra processi faremo riferimento a 2 modelli principali:

- A **memoria comune:** in questo caso prevediamo n processi con risorse private (cioè spazi di indirizzamento privati per ognuno), ma che possono accedere a risorse *condivise* in un terzo spazio di indirizzamento comune per tutti.

Se vogliamo ricondurci alla tassonomia di Flynn, questo è un esempio di macchina multiprocessore (o monoprocessoresso in *timesharing*...) di tipo MIMD (se vogliamo SM-MIMD): più unità di elaborazione sugli stessi dati (tralasciando il fatto che ogni unità ha poi i suoi dati privati);

- A **scambio di messaggi:** in questo caso prevediamo n processi in esecuzione su unità di elaborazione distinte (ancora, multiprocessore o monoprocessoresso in *time-sharing*) con le loro risorse (memorie) locali, che possono comunicare fra di loro attraverso un meccanismo di *scambio di messaggi*.

Notiamo di nuovo che non è necessario che le unità siano necessariamente distinte, e ricordiamo che non sono i processi a comunicare in sé per sé, ma le unità di elaborazione a supportare un meccanismo che permette tale comunicazione.

Questo è sempre un esempio di architettura MIMD (se vogliamo DM-MIMD), simile ad esempio a quella che ci permettono i sistemi multicalcolatore connessi in rete (più macchine distinte, con le loro risorse, che comunicano fra di loro).

11 Lezione del 22-10-25

11.1 Mutua esclusione

Analizziamo il problema della **mutua esclusione** studiando il seguente pseudocodice C, implementante una semplice struttura *stack*:

```

1 T stack[n];
2 int top = -1;
3
4 // inserisci in cima
5 void insert(T y) {
6     top++;
7     stack[top] = y;
8 }
9
10 // estrai dalla cima
11 T extract() {
12     T temp = stack[top];
13     top--;
14     return temp;
15 }
```

Se le variabili *stack* e *top* si trovano in memoria condivisa, o in altre la struttura *stack* creata si trova in memoria condivisa, potremmo incorrere in situazioni dove più operazioni sullo *stack* vengono iniziate contemporaneamente (le funzioni che operano sullo *stack* vengono chiamate contemporaneamente), e lo scheduler interlaccia le operazioni in un modo che rende lo *stack* inconsistente.

Ad esempio, con 2 processi *p*₁ e *p*₂, il primo chiamante *insert()* e il secondo chiamante *extract()* potremmo avere la timeline di esecuzione:

```

1 t0: top++;           // p1
2 t1: temp = stack[top]; // p2
3 t2: top--;           // p2
4 t3: stack[top] = y;   // p1
```

con conseguenze chiaramente disastrose! Si estre da una zona della pila non allocata e si inserisce sovrascrivendo la cima.

Chiaramente, sappiamo che nella maggior parte dei sistemi reali le istruzioni *top++*, ecc... non saranno eseguite atomicamente, ma lo saranno le istruzioni assembler che implementano tali istruzioni di alto livello. In ogni caso, se il sistema può fallire visto dal livello alto, possiamo essere sicuri che fallirà anche più probabilmente al livello più basso.

Siamo quindi di fronte al classico problema della mutua esclusione, dove vogliamo limitare l'accesso ad una determinata risorsa ad un solo processo per volta, o equivalentemente vogliamo rendere *atomiche* le operazioni su tale risorsa.

11.1.1 Soluzione (scorretta) software

Una prima idea potrebbe essere, *in software*, di introdurre un prologo alle funzioni che dovranno essere atomiche. Tale prologo avrà il compito di controllare un determinato flag di "prenotazione" sulla risorsa, ad esempio come:

```

1 prologo:
2     while(occupato == 1);
3     occupato = 1;
4     // sezione critica
```

```

5 epilogo:
6     occupato = 0;

```

Dobbiamo però renderci conto che a questo punto i due processi si trovano a condividere una nuova variabile in memoria condivisa, cioè il flag `occupato` stesso. Questo significa che possiamo incorrere nuovamente in situazioni dove lo stato diventa inconsistente, ad esempio, assumendo `occupato` inizialmente uguale a 0:

```

1 t0: while(occupato == 1); // p1, passa
2 t1: while(occupato == 1); // p2, passa
3 t2: occupato = 1        // p2
4 t3: occupato = 1        // p1
5 // p1 e p2 sono entrambi in sezione critica!

```

Notiamo inoltre che quello che abbiamo implementato è effettivamente una *busy wait* (attesa attiva) sulla variabile `occupato`, che nella programmazione di sistemi operativi è inaccettabile (rappresenta overhead inutile che potrebbe essere dedicato ad altri processi).

11.1.2 Soluzione hardware

Prevediamo allora una modifica *hardware* che ci permetta di risolvere il problema: magari una nuova istruzione assembler detta *test-and-set* (con mnemonica `TSL`). La `TSL` accetterà un registro e un indirizzo in memoria, e il suo funzionamento sarà il seguente:

- Carica il valore all'indirizzo nel registro;
- Imposta il valore nell'indirizzo a 1.

Immaginiamo che realizzare questo tipo di istruzione in sistemi multiprocessore richiederà l'aggiunta al bus di una nuova linea, detta *lock*, che permette ad un processore di bloccare il bus finché non ha terminato la sua operazione.

Quello che potremo fare è quindi implementare 2 routine assembler:

```

1 lock(x):
2     TSL registro, x
3     CMP registro, 0
4     JNE lock
5     RET // torna al chiamante, entra in sezione critica
6
7 unlock(x):
8     MOVE x, 0
9     RET

```

Il fatto che la `TSL` imposta subito, e soprattutto atomicamente, il valore all'indirizzo `x` a 1, cioè blocca subito la risorsa, ci permette di stare sicuri che nessun'altro avrà l'opportunità di "rubare" la risorsa occupata. Anche schedulando a livello istruzioni assembler (che è ciò che si fa nella realtà), dopo la `TSL` lo stato del sistema è consistente: il processo che ha diritto alla risorsa vi accederà, gli altri no.

I vantaggi di questo approccio sono che funziona, e funziona anche su sistemi multiprocessori (assunta memoria condivisa e bus dotato di linea *lock*). Gli svantaggi sono che siamo comunque costretti a fare una *busy wait*, anche se questa è più tollerabile della precedente (la combinazione `TSL`, `CMP` e `JNE` si sbrigava in pochi cicli di clock).

In questo caso, prologo ed epilogo visti da un linguaggio di alto livello come il C avranno il seguente aspetto:

```

1 prologo:
2   lock(x)
3   // sezione critica
4 epilogo:
5   unlock(x)

```

11.2 Semafori

I semafori rappresentano strumenti generali per la soluzione di problemi di sincronizzazione.

Un semaforo s è un oggetto alla quale è associato un intero non negativo, $s.value$, con valore iniziale $s_0 \geq 0$. Al semaforo è associata una lista di attesa, $s.queue$, nella quale sono posti i descrittori dei processi che attendono l'autorizzazione a procedere.

Le primitive sul semaforo coinvolgono il contatore $s.value$, e sono 2:

- **wait(s)**: questa si occupa di decrementare, se > 0 , $s.value$. Altrimenti mette il chiamante in attesa. In pseudocodice:

```

1 void wait(s) {
2   if(s.value == 0) {
3     // metti il chiamante in attesa
4     insert(s.queue, /* chiamante */);
5   } else {
6     s.value--;
7   }
8 }

```

- **signal(s)**: questa si occupa di incrementare $s.value$, e se c'erano processi in attesa risveglierne uno. In pseudocodice:

```

1 void signal(s) {
2   if(!isEmpty(s.queue)) {
3     primo = extract(s.queue); // e' una coda fifo
4     // inserisci primo in coda pronti
5   } else {
6     s.value++;
7   }
8 }

```

11.2.1 Semafori di mutua esclusione

I semafori permettono di realizzare la mutua esclusione. Chiamiamo questo tipo di semafori **mutex**.

Per realizzare un mutex basta inizializzare un semaforo col valore $s_0 = 1$. In questo caso basterà inserire prologo ed epilogo nelle funzioni che vogliamo rendere atomiche:

```

1 prologo:
2   wait(mutex);
3   // sezione critica
4 epilogo:
5   signal(mutex);

```

Quello che succede è che alla prima `wait()` il semaforo si svuota e tutti i processi successivi dovranno aspettare nella coda `mutex.queue`. Quando il processo finisce la sua operazione, esegue la `signal()`, liberando la risorsa per il prossimo processo in attesa.

12 Lezione del 23-10-25

12.0.1 Atomicità delle primitive semaforiche

Ora che il problema della mutua esclusione è effettivamente risolto, interroghiamoci su come rendere effettivamente atomiche le `wait()` e `signal()`.

- In ambiente monoprocesso, basterà rendere tali funzioni *primitive* di sistema, e quindi eseguirle con le interruzioni disattivate, per assicurarne l'atomicità.
- In ambienti multiprocessore, si potrebbero invece avere collisioni fra le `wait()` e `signal()` chiamate da processi concorrenti in esecuzione parallela.

Approfondiamo il problema: posto un semaforo di mutex, ad esempio, questo si troverà in memoria condivisa. Più di un processore potrà accedere alla memoria condivisa attraverso il bus. Se ci limitiamo a rendere "atomiche" le primitive `wait()` e `signal()` disattivando le interruzioni su *un* processore, non risolviamo il caso dove più processori vogliono accedere al semaforo contemporaneamente.

Possiamo risolvere il problema usando la tecnica introdotte in 11.1.2, cioè do-tandoci di un meccanismo hardware di *bloccaggio* del bus, fornito dall'istruzione `TSL`. Potremo infatti usare le primitive `lock()` e `unlock()` per bloccarci sulla risorsa semaforo, cioè dicendo:

```

1 prologo:
2 {
3     lock(mutex);
4     wait(mutex);
5     unlock(mutex);
6 }
7 // sezione critica
8 epilogo:
9 {
10    lock(mutex);
11    signal(mutex);
12    unlock(mutex);
13 }
```

Notiamo che questa soluzione non è propriamente corretta: infatti il processo non chiamerà la `unlock()` nel prologo finché non verrà svegliato dalla `wait()`: questo significa che potrebbe tenersi il lock sul semaforo, impedendo ad altri processi di segnalare sul semaforo e liberarlo!

La soluzione corretta sarà quindi quella di ridefinire le primitive semaforiche come segue:

– `wait(s):`

```

1 void wait(s) {
2     lock();
3     if(s.value == 0) {
4         // metti il chiamante in attesa
5         insert(s.queue, /* chiamante */);
6     } else {
7         s.value--;
8     }
9     unlock();
10 }
```

```

- signal(s):
  1 void signal(s) {
  2   lock();
  3   if(!isEmpty(s.queue)) {
  4     primo = extract(s.queue); // e' una coda fifo
  5     // inserisci primo in coda pronti
  6   } else {
  7     s.value++;
  8   }
  9   unlock();
10 }
```

12.1 Produttori e consumatori

Ipotizziamo adesso una situazione dove:

- Un processo, detto **produttore**, deposita un messaggio in un *buffer*;
- Un’altro processo, detto **consumatore**, preleva il messaggio dal *buffer*.

La policy sul buffer sarà la seguente:

- Il produttore non deve inserire un messaggio nel buffer se questo è pieno;
- Il consumatore non deve prelevare un messaggio dal buffer se questo è vuoto.

Possiamo usare 2 semafori per realizzare una prima soluzione:

- `spazio_disponibile`, con $s_0 = 1$, segnalerà quando il buffer è vuoto;
- `messaggio_disponibile`, con $s_0 = 0$, segnalerà quando il buffer è pieno.
- A questo punto il processo produttore dovrà controllare che il buffer sia vuoto (e aspettare che lo sia se non lo è), inserire il messaggio e segnalare che un nuovo messaggio è disponibile. In pseudocodice:

```

1 // produttore
2 do {
3   // produci messaggio
4   wait(spazio_disponibile);
5   buffer.insert(messaggio);
6   signal(messaggio_disponibile);
7 } while(!fine);
```

- Il consumatore dovrà invece controllare che il buffer abbia un nuovo messaggio (e aspettare che lo abbia se non lo ha), prelevare il messaggio e segnalare che il buffer è nuovamente vuoto. In pseudocodice:

```

1 // consumatore
2 do {
3   wait(messaggio_disponibile);
4   messaggio = buffer.extract();
5   signal(spazio_disponibile); // prima segnala e poi consuma!
6   // consuma messaggio
7 } while(!fine)
```

12.1.1 Più produttori e consumatori

Complichiamo la situazione: introduciamo un buffer ad n elementi, e prevediamo la presenza contemporanea di più produttori e consumatori.

In questo caso dovremmo assicurare, oltre che la sincronizzazione coi due semafori appena visti, la mutua esclusione attraverso un mutex. Inoltre, dovremmo prevedere che il semaforo `spazio_disponibile` abbia $s_0 = n$, e non 1. Possiamo fidarci che il meccanismo dei semafori assicura il corretto ordinamento dei processi (per ogni messaggio che inseriamo, si libera uno e un solo consumatore).

- Avremo quindi che lo pseudocodice del produttore sarà:

```

1 // produttore
2 do {
3     // produci messaggio
4     wait(spazio_disponibile);
5     {
6         wait(mutex);
7         buffer.insert(messaggio);
8         signal(mutex);
9     }
10    signal(messaggio_disponibile);
11 } while(!fine);
```

- Mentre lo pseudocodice del consumatore sarà:

```

1 // consumatore
2 do {
3     wait(messaggio_disponibile);
4     {
5         wait(mutex);
6         messaggio = buffer.extract();
7         signal(mutex);
8     }
9     signal(spazio_disponibile); // prima segnala e poi consuma!
10    // consuma messaggio
11 } while(!fine)
```

12.1.2 Semafori distinti

Accorgiamoci che in questo sistema, i produttori si bloccano su `spazio_disponibile`, i consumatori si bloccano su `messaggio_disponibile`, ed entrambi si possono bloccare sul `mutex`. Questo non è particolarmente elegante e può portare a situazioni di rallentamento.

Possiamo risolvere questo problema usando, anziché uno, 2 semafori di mutex.

- In questo caso lo pseudocodice del produttore sarà:

```

1 // produttore
2 do {
3     // produci messaggio
4     wait(spazio_disponibile);
5     {
6         wait(mutex_prodotto);
7         buffer.insert(messaggio);
8         signal(mutex_prodotto);
9     }
10    signal(messaggio_disponibile);
11 } while(!fine);
```

- Mentre lo pseudocodice del consumatore sarà:

```

1 // consumatore
2 do {
3     wait(messaggio_disponibile);
4     {
5         wait(mutex_consumatore);
6         messaggio = buffer.extract();
7         signal(mutex_consumatore);
8     }
9     signal(spazio_disponibile); // prima segnala e poi consuma!
10    // consuma messaggio
11 } while(!fine)

```

Questo chiaramente ci porta a dover fare delle considerazioni sulle modalità in cui si implementa il buffer. In particolare, vorremo che le operazioni `insert()` ed `extract()` siano completamente disaccoppiate e non possano collidere: questo perché la configurazione adottata permette a queste di essere eseguite contemporaneamente (l'una dal produttore e l'altra dal consumatore, che si bloccano su semafori diversi).

Se si adotta la classica implementazione ad array, questo problema non si pone. Se si usa una struttura più sofisticata come una lista, la situazione è più complicata. Vediamo nel dettaglio.

La soluzione che possiamo immaginare è di avere una lista con puntatore alla coda, dove le inserzioni (produttore) si fanno in coda, e le estrazioni (consumatore) si fanno in testa.

In questo caso, per liste con più di un elemento, operazioni di estrazione ed inserzione agiranno su oggetti completamente distinti in memoria, e non avremo problemi. Il problema sarebbe però quando si vuole avere un inserzione ed un'estrazione parallela su una lista con un solo elemento.

12.2 Primitive di comunicazione

Tralasciamo per adesso i sistemi in memoria condivisa, e parliamo dei sistemi distribuiti, composti da nodi con memorie locali. Questo è ad esempio il caso delle reti di calcolatori.

Il problema che ci poniamo è come sfruttare un certo **canale di comunicazione** orientato per realizzare due primitive, la primitiva `send(destinazione, messaggio)`, e la primitiva `receive(origine, messaggio)`. Notiamo che la `send()` è *asincrona* (o ugualmente, non *sincrona* o non *bloccante*): quando si ritorna dalla chiamata, non si può avere la sicurezza che il messaggio sia stato recapitato. Una primitiva *bloccante*, di contro, avrebbe sospeso il chiamante fino all'arrivo del messaggio: questo chiaramente implica l'attesa di un ACK da parte del destinatario.

La `receive()` è invece necessariamente bloccante: il processo chiamante viene messo in attesa finché un messaggio non è stato effettivamente ricevuto e può essere recapitato.

Una soluzione più esplicita per il programmatore a cui forniamo le `send()` e `receive()` potrebbe essere quella di imporre il ricevimento dell'ACK, cioè:

```

1 // invia
2 send(destinazione, messaggio);
3 // aspetta l'ACK, e' bloccante
4 ack = receive(destinazione);

```

Questo sfrutta il fatto che la `receive()` è bloccante e effettivamente risolve il nostro problema, permettendoci di mantenere la `send()` asincrona. Chiaramente, però, richiede al programmatore di scrivere codice più complicato (e corretto!).

12.2.1 Formato del messaggio

Per dotarci di primitive di comunicazione, abbiamo bisogno di stabilire un **formato standard** per i messaggi che andiamo ad inviare. Questo è solitamente diviso in:

- **Intestazione:** contiene informazioni su:
 - **Origine** del messaggio;
 - **Destinazione** del messaggio;
 - **Tipo** del messaggio;
 - **Lunghezza** (in byte) del messaggio;
 - *Informazioni di controllo* varie sul messaggio.
- **Corpo:** contiene il messaggio vero e proprio, o *payload*, che vogliamo trasmettere.

12.2.2 Produttori e consumatori remoti

Vediamo un primo esempio di come le primitive di comunicazione potrebbero essere usate, ad esempio per realizzare un sistema produttore e consumatore.

In particolare, vediamo 2 varianti di comunicazione, **diretta simmetrica** e **diretta asimmetrica**:

- Comunicazione **diretta simmetrica**:

- Lato produttore si avrà:

```

1 pid consumatore = /* ... */;
2 main() {
3     msg mess;
4     do {
5         produci(&mess);
6         send(consumatore, mess);
7     } while(!fine);
8 }
```

- Lato consumatore si avrà:

```

1 pid produttore = /* ... */;
2 main() {
3     msg mess;
4     do {
5         receive(produttore, &mess);
6         consuma(M);
7     } while(!fine);
8 }
```

Vediamo come in questo tipo di comunicazione non è necessario prevedere un *buffer*: si invia un messaggio per volta e si aspetta, lato consumatore, per ogni messaggio.

- Comunicazione **diretta asimmetrica**:

- Lato produttore si avrà:

```

1 pid consumatore = /* ... */;
2 main() {
3     msg mess;
4     do {
5         produci(&mess);
6         send(consumatore, mess);
7     } while(!fine);
8 }

```

- Lato consumatore si avrà:

```

1 main() {
2     msg mess;
3     pid produttore;
4     do {
5         receive(&produttore, &mess);
6         consuma(M);
7     } while(!fine);
8 }

```

In questo caso il produttore non è già noto al consumatore, che invece si mette in ascolto per il primo messaggio disponibile.

12.2.3 Modello client-server

Il discorso fatto finora su produttori e consumatori può essere sviluppato introducendo il paradigma (probabilmente già noto) **client-server**.

In questo caso prevediamo più processi, detti **client** (o *clienti*) che richiedono servizi ad un solo processo, detto **server** (o *servitore*). I client accedono al server tramite una determinata **porta**, che per quanto ci riguarda si occupa anche di *bufferizzazione* delle richieste dei client prima che queste arrivino al processo server vero e proprio.

13 Lezione del 28-10-25

13.1 Lettori e scrittori

Sempre sull'argomento della sincronizzazione fra processi, vediamo l'esempio di più **scrittori** che vogliono scrivere su una risorsa che è letta da più **lettori**.

Iniziamo a vedere quali politiche vogliamo assicurare:

1. Chiaramente, fra gli scrittori c'è una stretta politica di mutua esclusione: non si può scrivere in 2 o più contemporaneamente.
2. Anche fra lettori e scrittori deve esserci mutua esclusione (non possiamo leggere ciò che è inconsistente perché ci si sta scrivendo);
3. In tutti gli altri contesti, vorremmo permettere di avere più lettori contemporanei.

Per risolvere il problema della mutua esclusione fra scrittori (1) prevediamo un semaforo di mutex `sem wrt = 1`, che viene prelevato in fase di scrittura:

```

1 proc writer {
2     wait(wrt); // mutex
3
4     // scrivi
5
6     signal(wrt); // mutex
7 }

```

Il semaforo `wrt`, inizialmente pensato per la mutua esclusione fra scrittori (1), può essere usato anche dai lettori per risolvere il problema (2). Il problema in questo caso sarà che non assicureremo la politica (3) di letture contemporanee: facendo la `wait()` su `wrt` sblocchiamo infatti un lettore per volta.

Dotiamoci quindi di un contatore `int readCount = 0`, che tiene conto dei processi lettori che attualmente stanno leggendo la risorsa. Proteggiamo quindi il contatore con un nuovo semaforo di mutex `sem mutex = 1`.

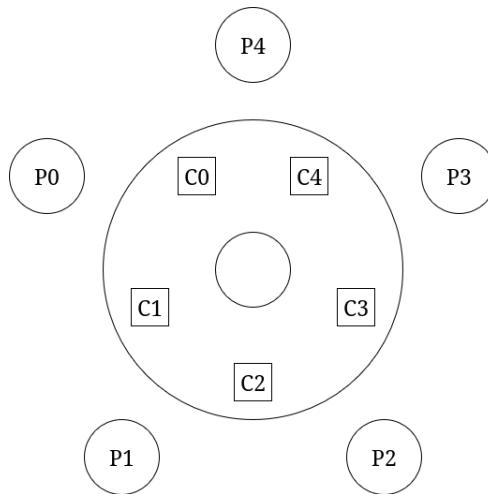
```

1 proc reader {
2     wait(mutex);
3     {
4         readCount++;
5         if(readCount == 1) wait(wrt); // solo il primo aspetta
6     }
7     signal(mutex);
8
9     // lettura
10
11    wait(mutex);
12    {
13        readCount--;
14        if(readCount == 0) signal(wrt); // solo l'ultimo rilascia
15    }
16    signal(mutex);
17 }
```

In questo caso saranno solo rispettivamente il primo e l'ultimo lettore a prendersi la briga di fare la `wait()` e quindi la successiva `signal()` sul semaforo condiviso con gli scrittori.

13.2 Problema dei 5 filosofi

Veniamo quindi ad un esempio celebre di programmazione concorrente.



Ipotizziamo una situazione dove 5 filosofi p_0, p_1, \dots sono seduti ad una tavola circolare, al centro della quale è posta una scodella di riso. Sulla tavola, una alla destra di ogni filosofo, ci sono esattamente 5 bacchette c_0, c_1, \dots . Ogni filosofo per mangiare ha bisogno di due bacchette. Il problema è: come possono i filosofi coordinarsi per mangiare tutti, e quindi ottenere tutti ciclicamente le 2 bacchette?

Contemporaneamente, possono mangiare al massimo 2 filosofi: ad esempio, se sta mangiando p_0 , possono mangiare contemporaneamente solo p_2 o p_3 .

Vediamo il comportamento del singolo filosofo. Questo potrà trovarsi in uno di 3 stati:

```

1 enum State {
2     THINKING, // non ha fame
3     HUNGRY,   // sta cercando di ottenere 2 bacchette
4     EATING    // sta mangiando
5 }
```

1. Un primo approccio può essere quello di dotarsi di un semaforo di mutex per bacchetta, cioè avere `sem chopstick[5] = 1`. In questo caso lo pseudocodice del filosofo sarà:

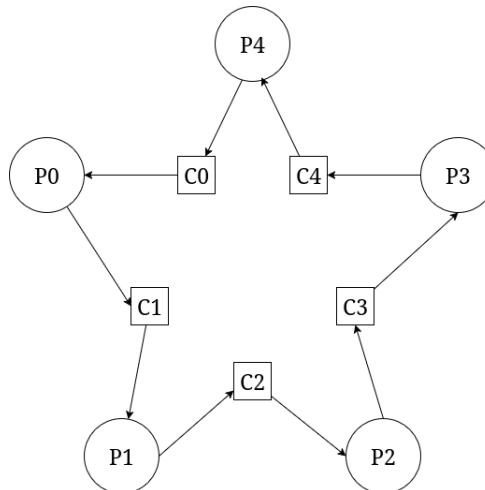
```

1 proc philosopher {
2     // pensa
3
4     // aspetta bacchette
5     wait(chopstick[i]); // sx
6     wait(chopstick[(i + 1) % 5]); // dx
7
8     // mangia
9
10    // rilascia bacchette
11    signal(chopstick[(i + 1) % 5]); // dx
12    signal(chopstick[i]); // sx
13 }
```

Questo approccio può però portare a *deadlock* nel caso in cui tutti i processi riescano ad effettuare la prima `wait()`: in questo caso si troveranno con la bacchetta a sinistra presa da loro, e quella a destra presa dal vicino, col risultato che nessuno può procedere e i filosofi muoiono di fame. Peccato!

Possiamo modellizzare questa situazione con un grafo, dove nodi circolari rappresentano i **filosofi** (processi), e nodi quadrati rappresentano le **bacchette** (*risorse*). Useremo le frecce da processi e risorse per rappresentare l'**attesa** di una risorsa (chiamata di `wait()`), e le frecce da risorse a processi per rappresentare il **possesso** di una risorsa (`wait()` terminata).

Vediamo che la situazione di deadlock appena descritto in questo caso è rappresentata da un **ciclo** nel grafo processi-risorse:



Per risolvere questo problema, vorremo prima approfondire il concetto di deadlock, e quindi implementare appropriate tecniche di deadlock **detection** (*rilevamento* di deadlock) e deadlock **avoidance** (*risoluzione* o *prevenzione* di deadlock).

13.3 Monitor

I **monitor** rappresentano un'astrazione di alto livello che permettono la sincronizzazione di processi. Sostanzialmente, sono strutture dati contenenti funzioni (*operazioni*) che vengono eseguite in mutua esclusione (cioè in maniera *atomica*) all'interno di un certo contesto (dove si condivide codice di *inizializzazione* e *dati*).

All'interno di un monitor prevediamo variabili di **condizione**, su cui sono permesse operazioni di `wait()` e `signal()`. Chiaramente queste variabili di condizioni saranno visibili solo all'*interno* del monitor, cioè del codice delle operazioni definite dal monitor.

- La `wait()` mette in attesa un processo finché un altro non esegue una `signal()`;
- La `signal()` sveglia i processi che erano in `wait` sulla variabile di condizione. La particolarità della `signal()` è che non ha effetti se nessuno è in stato di `wait()`.

13.3.1 Gestione delle variabili di condizione

Assumiamo che un processo *P* invochi `x.signal()` sulla variabile di condizione *x*, mentre un altro processo *Q* si trova nello stato `x.wait()`. Cosa dovrebbe succedere a questo punto?

Ci sono due opzioni:

- **Signal and wait:** *P* aspetta che *Q* esca dal monitor o si metta in attesa di un'altra condizione: è il caso *preemptive*;
- **Signal and continue:** *Q* aspetta che *P* esca dal monitor o si metta in attesa di un'altra condizione.

Noi adotteremo la soluzione *signal and wait*.

13.3.2 Monitor per problema dei 5 filosofi

Vediamo quindi come un monitor può essere usato per risolvere il problema dei 5 filosofi.

```

1 monitor Philosophers {
2     enum State {
3         THINKING, // non ha fame
4         HUNGRY,    // sta cercando di ottenere 2 bacchette
5         EATING     // sta mangiando
6     }
7     State state[5];
8
9     // variabili di condizione
10    condition self[5];
11
12    void pickup(int i) {
13        state[i] = HUNGRY; // hai fame
14        test(i); // puoi mangiare?
15        if(state[i] != EATING) self[i].wait(); // se puoi mangia
16    }
17

```

```

18 void putdown(int i) {
19     state[i] = THINKING; // non stai piu' mangiando
20     test((i - 1) % 5); // sx puo' mangiare?
21     test((i + 1) % 5); // dx puo' mangiare?
22 }
23
24 void test(int i) {
25     if(
26         (state[(i - 1) % 5] != EATING) && // sx non sta mangiando?
27         (state[i] == HUNGRY) &&           // hai fame?
28         (state[(i + 1) % 5] != EATING)       // dx non sta mangiando?
29     ) {
30         state[i] = EATING; // mangia
31         self[i].signal();
32     }
33 }
34
35 initialization_code() {
36     for(int i = 0; i < 5; i++) {
37         state[i] = THINKING;
38     }
39 }
40 }
```

A questo punto il processo filosofo sarà molto semplice:

```

1 process philospher {
2     Philosophers.pickup();
3
4     // mangia
5
6     Philosophers.putdown();
7 }
```

Abbiamo che questo approccio è privo di deadlock. Infatti, i processi filosofi non provano ad ottenere le bacchette finché non hanno la sicurezza di poter prendere *entrambe* le bacchette.

14 Lezione del 29-10-25

14.1 Implementazione di un monitor

Veniamo quindi a come si può effettivamente implementare un monitor come descritto nella scorsa lezione.

Avere più funzioni in mutua esclusione significa effettivamente usare un semaforo di mutex `sem mutex = 1`, e avere il seguente prologo ed epilogo di funzione per ogni funzione interna al monitor:

```

1 func_monitor() {
2     // prologo
3     wait(mutex);
4
5     // corpo func
6
7     // epilogo
8     signal(mutex);
9 }
```

Il problema diventa quindi la gestione delle variabili di condizione:

- Ricordiamo che `x.wait()` vuole che il processo attuale si sospenda;
- `x.signal()` potrebbe invece bloccare il processo e passare ad un altro (*signal and wait*) oppure continuare col processo corrente (*signal and continue*). Noi, come anticipato in 13.3.1, useremo la prima politica.

Avremo quindi un semaforo inizializzato a zero su ogni variabile di condizione (ad esempio `sem x_sem = 0`) per il blocco, e un contatore dei processi bloccati sulla variabile (ad esempio `int x_count = 0`).

- A questo punto la `x.wait()` sarà:

```

1 x.wait() {
2     x_count++;
3
4     signal(mutex); // devo sbloccare il mutex
5     wait(x_sem);
6 }
```

Il problema è che facciamo una `signal(mutex)`, quando in verità vorremmo segnalare di proseguire ai processi già interni al monitor. Modifichiamo allora il monitor, introducendo un semaforo `sem next = 0` e un contatore `int next_count = 0` per i processi "bloccati" al suo interno.

Prologo ed epilogo saranno allora:

```

1 func_monitor() {
2     // prologo
3     wait(mutex);
4
5     // corpo func
6
7     // epilogo
8     if(next_count > 0) {
9         signal(next); // prima fai uscire i processi interni
10    } else {
11        signal(mutex); // poi apri il monitor ad altri
12    }
13 }
```

A questo punto il codice della `x.wait()` potrà essere:

```

1 x.wait() {
2     x_count++;
3
4     if(next_count > 0) { // c'e' qualcuno in attesa
5         signal(next);
6     } else { // non c'e' nessuno, sblocca il monitor
7         signal(mutex);
8     }
9
10    wait(x_sem);
11    x_count--; // uscito da wait()
12 }
```

- Implementiamo quindi la `x.signal()` secondo la politica *signal and wait*:

```

1 x.signal() {
2     if(x_count > 0) {
3         signal(x_sem);
```

```

4
5     next_count++;
6     wait(next); // sono uno dei processi del monitor
7     next_count--;
8 }
9 }
```

Ci dovrebbe quindi essere chiaro il funzionamento del monitor come un ambiente "ristretto" per i processi del sistema dove lo scheduling non è necessariamente FCFS (o qualsiasi fosse l'algoritmo usato dallo scheduler del sistema).

14.1.1 Conditional wait

Un'altra possibile politica che si può adottare all'interno dei monitor è la cosiddetta **conditional wait**, nella forma `x.wait(c)` dove `c` è un *numero di priorità*. I processi con numero di priorità più piccolo (priorità più alta) vengono schedulati per primi.

Un esempio dove potrebbe essere utile usare tale costrutto è il seguente, dove si implementa un monitor con il compito di allocare una certa risorsa:

```

1 monitor ResourceAllocator {
2     boolean busy;
3     condition x;
4     void acquire(int time) {
5         if(busy) x.wait(time);
6         busy = TRUE;
7     }
8
9     void release() {
10        busy = FALSE;
11        x.signal();
12    }
13
14     initialization code() {
15        busy = FALSE;
16    }
17 }
```

In questo caso prendiamo come argomento `time`, cioè il tempo per cui occupiamo la risorsa (meno tempo → più priorità).

14.2 Deadlock

Veniamo quindi alla trattazione vera e propria dei **deadlock**, o *blocchi critici*, che avevamo introdotto in 13.2.

Di base, questi sono situazioni dove ciascun processo, in un insieme di processi, detiene una risorsa e ne desidera una di un altro. Sul grafo di allocazione, equivalentemente, significa che abbiamo un *ciclo*.

Ricordiamo che ci eravamo posti di implementare appropriate tecniche di **deadlock detection** (*rilevamento* di deadlock) e deadlock **avoidance** (*risoluzione* o *prevenzione* di deadlock).

Notiamo adesso che in verità esistono casi dove si possono avere cicli nel grafo di allocazione, ma non avere deadlock: questo è il caso di risorse con **istanze multiple**. In quest'caso, un ciclo in un grafo di allocazione simboleggia la *possibilità* di aver deadlock, ma la situazione potrebbe comunque risolversi (lo si verifica per osservazione del grafo).

Una soluzione banale al problema del deadlock è obbligare il programmatore a dichiarare subito tutte le risorse di cui il programma ha bisogno: a questo punto, lato S/O, si potrà realizzare via mutex su tali risorse un sistema di bloccaggio che eviterà sempre i deadlock. Il problema è chiaramente che tale vincolo è estremamente restrittivo, e renderebbe non solo molto scomodo per il programmatore programmare una data applicazione, ma in generale abbatterebbe l'efficienza dell'intero sistema.

14.2.1 Condizioni di deadlock

Esistono 4 condizioni necessarie affinché si verifichi un deadlock:

1. **Mutua esclusione**: solo un processo per volta può usare una data risorsa;
2. **Hold and wait**: un processo che ha ottenuto almeno una risorsa si mette in attesa di altre risorse ottenute da altri processi;
3. **No preemption**: una risorsa può essere rilasciata solo *volontariamente* dai processi che la ottengono, quando questi completano la loro operazione sulla stessa;
4. **Attesa circolare**: esiste un insieme $\{p_0, p_1, \dots, p_n\}$ di processi in attesa tali che p_0 aspetta una risorsa di p_1 , p_1 aspetta una risorsa di p_2 , ..., p_{n-1} aspetta una risorsa di p_n e p_n aspetta una risorsa di p_1 .

Si ha che difficilmente si può lavorare sulle prime 3 condizioni: la mutua esclusione deriva dal fatto che ci sono fisicamente risorse limitate nel sistema, e non vogliamo imporre al programmatore vincoli su come e quali risorse vogliono ottenere, o constringerli a programmare meccanismi di recupero dal ritiro di risorse (*preemption* su risorse).

15 Lezione del 06-11-25

15.0.1 Cicli su grafi di allocazione

Ricordiamo i fatti di base sul rilevamento di deadlock attraverso l'ispezione dei **grafi di allocazione**.

- Se il grafo non contiene cicli, abbiamo detto non c'è possibilità di deadlock;
- Se il grafo contiene cicli, dobbiamo discriminare sulla presenza di *istanze multiple* di risorsa:
 - Se si ha una sola istanza per tipo di risorsa, allora si ha deadlock;
 - Se si hanno più istanze per tipo di risorsa, allora c'è la possibilità, ma non la sicurezza, di avere deadlock.

15.0.2 Metodi di gestione dei deadlock

Ricordiamo quindi i metodi che avevamo previsto per gestire i deadlock:

- Potremmo assicurarci che il sistema non entri mai in uno stato di deadlock: questo è il modello di **deadlock prevention** (*prevenzione statica*).

Questo è l'approccio usato dalla maggior parte dei sistemi operativi, che preferiscono lasciar eseguire i processi senza attivarsi in maniera dinamica per gestire eventuali deadlock. Di contro, si spera (e si progettano sistemi in maniera tale) che tali situazioni non si verifichino;

- Potremmo permettere al sistema in normale operazione di trovarsi in stato di deadlock, rilevare tale deadlock e iniziare delle procedure di recupero: questo è il modello di **deadlock detection** e **avoidance** (*prevenzione dinamica*).
Questo modello è più usato nei sistemi embedded, dove si possono fare ipotesi più stringenti sui processi in esecuzione.

15.1 Prevenzione dinamica, deadlock avoidance

Vediamo quindi nel dettaglio la prevenzione dinamica di deadlock. In questo caso dobbiamo permettere al sistema di avere a disposizione alcune informazioni a priori:

- Nel caso più semplice vogliamo che ogni processo dichiari il numero massimo di risorse di ogni tipo di cui ha bisogno;
- Un **algoritmo** di prevenzione statica deve essere messo in piedi, e deve eseguire in maniera dinamica per prevenire situazioni di attese circolari. Il più celebre di questi algoritmi è l'*algoritmo del banchiere*;
- Lo **stato** dell'allocazione di risorse è definito dal numero di risorse disponibili ed allocate, e dal numero massimo di risorse che i processi possono richiedere (che abbiamo detto il S/O deve sapere).

15.1.1 Stato sicuro

Definiamo cosa intendiamo per **safe state** (o *stato sicuro*) del sistema.

Ogni volta che un processo richiede una risorsa disponibile, il S/O deve decidere se l'allocazione immediata di tale risorsa lascia il sistema in uno stato sicuro.

Il sistema si dice in stato sicuro se esiste una sequenza $\{p_0, p_1, \dots, p_n\}$ di tutti processi tale che per ogni p_i , le risorse che p_i può richiedere possono essere allocate con le risorse correntemente disponibili più le risorse allocate ai p_j con $j < i$.

Questo significa che:

- Se p_i ha bisogno di risorse e queste sono disponibili, non c'è problema;
- Se p_i ha bisogno di risorse e queste non sono disponibili, può:
 1. Aspettare che tutti i p_j terminino, liberando le risorse di cui ha bisogno;
 2. Eseguire quando i p_j terminano e le risorse sono libere, allocando e quindi liberando nuovamente le sue risorse;
 3. Quando p_i termina liberando le sue risorse, p_{i+1} può eseguire, e così via.

Possiamo corredare i fatti di base di 15.0.1 con alcuni altri fatti, riguardanti lo stato safe:

- Se il sistema è in stato safe, non c'è possibilità di deadlock;
- Se il sistema non è in stato safe, c'è possibilità, ma non la sicurezza, di avere deadlock.

La deadlock *avoidance* corrisponde esattamente a evitare che il sistema arrivi a stati unsafe.

Vediamo un caso di esempio tipico secondo il modello dello stato sicuro. Poniamo di avere 2 processi (p_0 e p_1), interessati ad ottenere in maniera concorrente 2 risorse (r_0 e r_1):

- Processo p_0 :

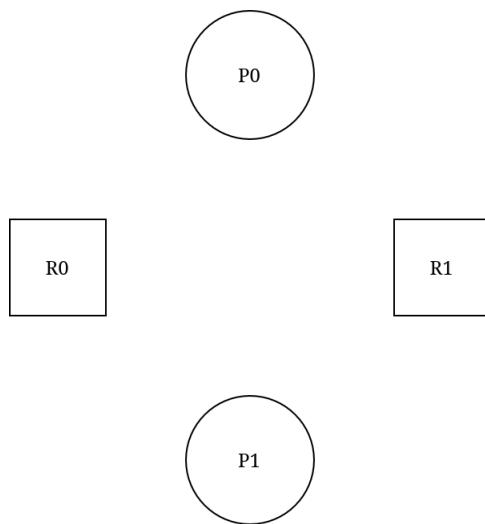
```
1 // p0, p1
2 wait(m0);
3 wait(m1);
4
5 // elaborazione su r0 e r1
6
7 signal(m1);
8 signal(m0);
```

- Processo p_1 :

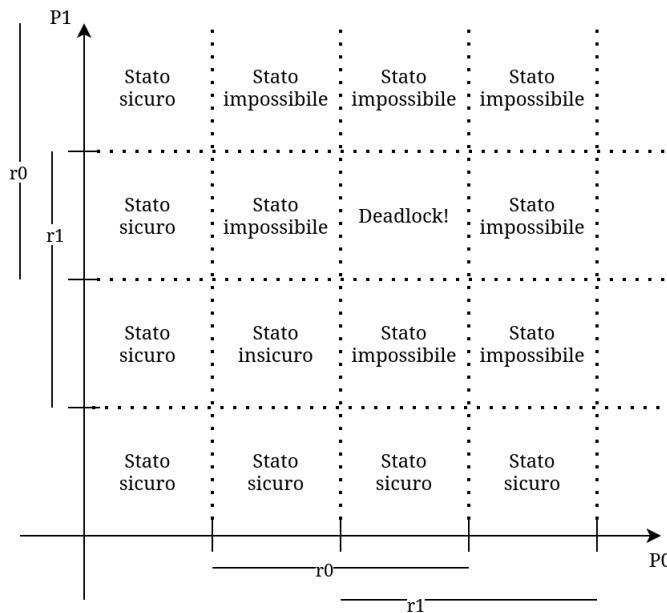
```
1 // p0, p1
2 wait(m1);
3 wait(m0);
4
5 // elaborazione su r0 e r1
6
7 signal(m0);
8 signal(m1);
```

Notiamo che i processi accedono ai mutex delle risorse in ordine inverso.

Il grafo di allocazione è il solito che abbiamo già visto:

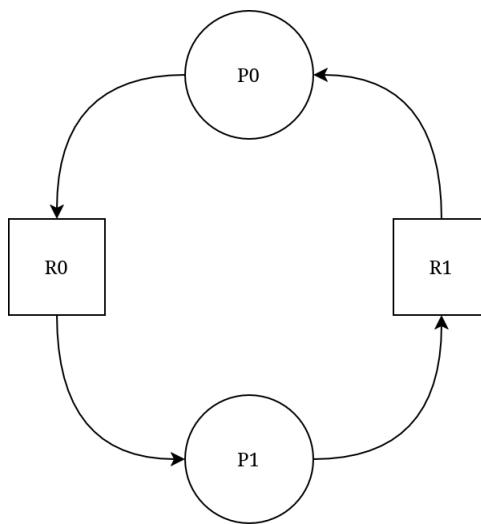


Possiamo modellizzare la stessa situazione, in evoluzione nel tempo, attraverso un grafico di questo tipo:



Dove notiamo l'evoluzione temporale (sui 2 assi) dei 2 processi, in funzione delle risorse acquisite, e gli stati corrispondenti che incontriamo. Vediamo come appena i 2 processi acquisiscono rispettivamente la risorsa r_0 e r_1 , andiamo a trovarci nello stato insicuro, a cui segue un deadlock appena provano ad accedere alla risorsa seguente.

Nel grafo di allocazione questo è il deadlock:



15.1.2 Grafici di allocazione estesi

Estendiamo i grafi di allocazione, che abbiamo finora trattato in maniera abbastanza informale.

Avevamo che i cerchi rappresentano processi, e i quadrati risorse. Riguardo agli archi che li collegano, abbiamo:

- Un arco di *claim* da un processo a una risorsa simboleggia che la risorsa è desiderata dal processo, e potrebbe essere richiesta. Lo rappresentiamo come una freccia tratteggiata;
- Un arco di *claim* si trasforma in arco di *request* quando il processo effettivamente richiede la risorsa. Lo rappresentiamo come una freccia piena;
- Un arco di *assignment* (o allocazione) è diretto da una risorsa a un processo (il contrario degli archi di *claim* e *request*), e simboleggia che la risorsa è effettivamente posseduta dal processo. Lo rappresentiamo sempre come una freccia piena.

A questo punto il significato di un algoritmo di deadlock avoidance dovrebbe essere chiaro: se le transizioni in fase di allocazione di risorse consistono nella trasformazione di arco di *request* a un arco di *assignment*, allora dobbiamo assicurarcì che le richieste vengano soddisfatte solo quando l'arco di *assignment* formato non porta alla formazione di cicli nel grafo di allocazione.

15.2 Algoritmo del banchiere

Vediamo quindi l'algoritmo più celebre di deadlock avoidance: il cosiddetto **algoritmo del banchiere**.

Le ipotesi dell'algoritmo sono:

- Esistono istanze multiple di risorse (nel caso banale, una sola istanza);
- Ogni processo deve fare *claim* a priori delle risorse massime che potrebbe usare;
- Ogni processo deve essere disposto ad aspettare per le sue risorse;
- Quando un processo ottiene tutte le risorse necessarie alla sua esecuzione, deve eseguire e restituirle in un lasso finito di tempo.

Definiamo le condizioni di stato dell'algoritmo. Sia n il numero di processi e m il numero di tipi di risorse.

- Le risorse **disponibili** saranno un vettore di interi di lunghezza m . Se la risorsa all'indice j vale k , significa che ci sono k istanze della risorsa corrispondente disponibili;
- Per ogni processo dobbiamo sapere la quantità **massima** di risorse che vorrà allocare. Rappresentiamo questo con una matrice bidimensionale $n \times m$. Se questa all'indice (i, j) vale k , allora il processo i potrà richiedere al massimo k risorse di tipo j ;
- Manteniamo una matrice simile, di risorse **allocate**. Se questa all'indice (i, j) vale k , allora il processo i avrà allocato k risorse di tipo j ;
- Infine, manteniamo un'altra matrice simile, di risorse **desiderate**. Se questa all'indice (i, j) vale k , allora il processo i avrà bisogno (oltre a quelle che già ha), di k risorse di tipo j .

Definiamo quindi le matrici e i vettori come *Avail*, *Max*, *Alloc* e *Need*. Viene da sé che:

$$\text{Need}[i, j] = \text{Max}[i, j] - \text{Alloc}[i, j]$$

Vediamo allora un primo algoritmo, di *sicurezza*, che controlla se il sistema si trova in uno stato sicuro.

1. Siano $Work$ e $Finish$ vettori di lunghezza rispettivamente m e n . Inizializza:

- $Work = Avail$
- $Finish[i] = \text{false}$ per $i = 0, 1, \dots, n - 1$

2. Trova un i tale che:

- $Finish[i] = \text{false}$
- $Need[i] \leq Work$

se non esiste nessun i che soddisfa le condizioni, vai al passo 4;

3. Poni:

- $Work = Work + Alloc$
- $Finish[i] = \text{true}$

quindi vai al passo 2;

4. Se $Finish == \text{true}$ per ogni i , allora il sistema è in stato sicuro.

Vediamo allora l'algoritmo del banchiere vero e proprio. Sia $Request$ il vettore richiesta per il processo all'indice i . Se $Request[j] = k$ allora il processo i vuole k istanze della risorsa di tipo j .

1. Controlla che $Request[i] \leq Need[i]$. Se il controllo passa, vai al passo 2. Altrimenti solleva un errore, in quanto il processo ha richiesto più risorse di quante ha dichiarato di volere;
2. Se $Request[i] \leq Avail$, vai al passo 3. Altrimenti aspetta, in quanto alcune risorse non sono disponibili e devono essere liberate dai processi precedenti;
3. Simula l'allocazione delle risorse al processo i modificando lo stato come segue:

- $Avail = Avail - Request$
- $Alloc[i] = Alloc[i] + Request$
- $Need[i] = Need[i] - Request$

A questo punto esegui l'algoritmo di sicurezza.

- Se lo stato è sicuro, non cambiare nulla ed alloca le risorse al processo i ;
- Se lo stato non è sicuro, riporta lo stato a prima della simulazione e aspetta.

16 Lezione del 11-11-25

16.1 Prevenzione dinamica, deadlock detection

Visto l'algoritmo del banchiere, celebre algoritmo di deadlock *avoidance*, veniamo allo studio della deadlock detection.

Nell'esempio più semplice di deadlock *detection*, dove esiste una singola istanza di ogni tipo di risorsa, vogliamo mantenere un grafo di allocazione delle risorse. Periodicamente, invocheremo un algoritmo che analizza il grafo per trovare un ciclo. Se esiste un ciclo, esiste un deadlock (dai fatti in 15.0.1). Ricordiamo adesso che trovare cicli in un grafo di n nodi richiede un'algoritmo $O(n^2)$.

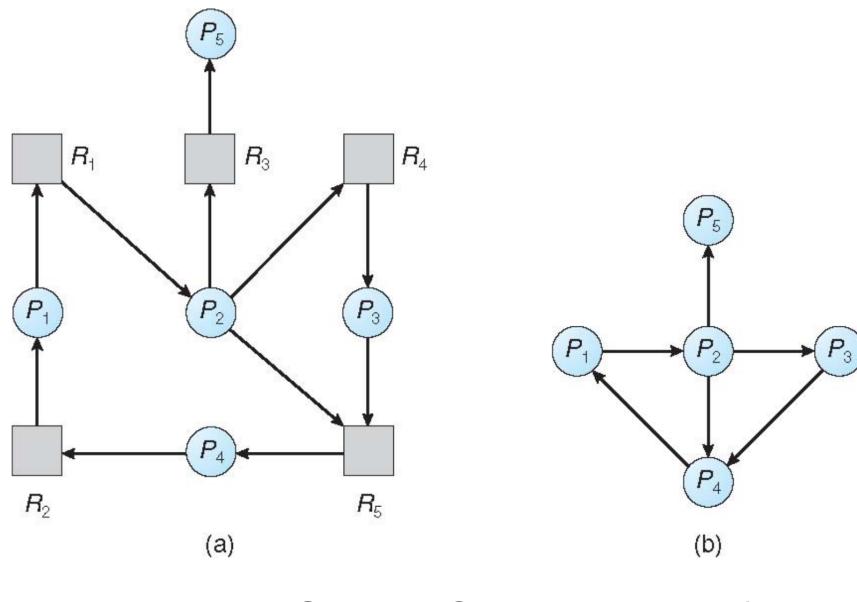
In ogni caso, quando si rileva un deadlock sarà opportuno invocare un qualche altro algoritmo, detto di deadlock *recovery*, per risolvere il deadlock.

16.1.1 Grafo di attesa

Il grafo che manteniamo in un algoritmo di deadlock detection non è propriamente un grafo di allocazione, ma un cosiddetto grafo di **attesa** (o grafo *wait-for*).

In questo caso l'informazione che vogliamo allocare non è l'attesa sulle *risorse*, ma sui processi che competono per risorse. La conversione da grafo di allocazione a grafo di attesa è banale: basta sostituire le triple freccia-risorsa-freccia con singole frecce dirette nello stesso senso.

Questo è meglio spiegato dal seguente grafico, che mostra un grafo di allocazione a sinistra, e a destra il corrispondente grafo di attesa:



Resource-Allocation Graph Corresponding wait-for graph

Il grafo di attesa è meno espressivo del grafo di allocazione, ma contiene comunque tutte le informazioni necessarie a fare deadlock detection. Infatti, se il grafo di allocazione contiene cicli, il grafo di attesa conserva tali cicli.

16.1.2 Algoritmo di deadlock detection

Potremmo interrogarci su quando è conveniente invocare un determinato algoritmo di deadlock detection.

Abbiamo introdotto in 16.1 che questo avrà necessariamente complessità $O(n^2)$ (in quanto deve scansionare un grafo per cicli).

- Chiaramente, eseguirlo troppo spesso (come ad esempio ad ogni allocazione di risorsa, cioè nei momenti in cui il grafo di attesa cambia e si verifica la possibilità di deadlock) porterebbe ad un'overhead troppo consistente;
- Un approccio simile al precedente è quello di effettuare deadlock detection ogni volta che non si riesce a soddisfare una richiesta di allocazione di risorse. Notiamo però che questo causa esecuzioni inutili quando il sistema ha, effettivamente, finito le risorse;

- Un buon approccio è quello di valutare l'utilizzo della CPU, cioè quanto tempo della CPU viene effettivamente utilizzato, e quanto sprecato in attesa di risorse. A questo punto, decidiamo di mettere in esecuzione l'algoritmo di deadlock detection quando il tempo di utilizzo è minore di una certa soglia (magari il 40 – 60%).
- Infine, l'approccio meno drastico e con minore overhead è quello di eseguire l'algoritmo periodicamente, a prescindere dall'utilizzo CPU o dalle risorse allocate.

16.1.3 Detection in più istanze

Abbiamo fatto, in 16.1, una semplificazione per il nostro algoritmo di deadlock detection: quella di assumere che tutte le risorse esistano in singola istanza. Avevamo visto da 15.0.1 che questo significa che un ciclo è causa sufficiente per un deadlock. Vediamo come si possono implementare algoritmi di deadlock che funzionano in presenza di istanze multiple di risorsa, cioè quando questa è solo causa necessaria.

Vogliamo comportarci come nell'algoritmo del banchiere, e quindi mantenere, sia n il numero di processi e m il numero di tipi di risorse:

- Le risorse **disponibili** saranno un vettore di interi di lunghezza m . Se la risorsa all'indice j vale k , significa che ci sono k istanze della risorsa corrispondente disponibili;
- Manteniamo una matrice di risorse **allocate**. Se questa all'indice (i, j) vale k , allora il processo i avrà allocato k risorse di tipo j ;
- Infine, manteniamo un'altra matrice simile, di risorse **richieste**. Se questa all'indice (i, j) vale k , allora il processo i sta richiedendo (oltre a quelle che già ha), k risorse di tipo j .

Definiamo quindi le matrici e i vettori come *Avail*, *Alloc* e *Request*.

Notiamo che queste sono in qualche modo corrispondenti alle variabili di stato viste in 15.2 (dove si è discusso l'algoritmo del banchiere), se non per il fatto che l'equazione che legava *Need* a *Max* e *Alloc* non è più presente. Questo è chiaro dal fatto che non abbiamo nessuna indicazione delle risorse massime richieste da un processo, e che questo potrebbe richiederne ancora, teoricamente all'infinito.

In verità, la matrice *Need* non è propriamente duale alla *Request*: se la *Need* mantiene uno stato noto a priori e che varia in fase di allocazione di risorse da parte di un processo, la *Request* varia nel tempo sulla base del comportamento del processo (varia sostanzialmente in fase di chiamata di primitive di allocazione).

L'algoritmo a questo punto è molto simile all'algoritmo di sicurezza del banchiere:

1. Siano *Work* e *Finish* vettori di lunghezza rispettivamente m e n . Inizializza:

- $Work = Avail$
- $Finish[i] = \text{false}$ se $Alloc[i] \neq 0$, altrimenti true , per $i = 0, 1, \dots, n - 1$. Questa condizione non equivale strettamente al fatto che il processo è terminato. Piuttosto, significa che il processo non ha risorse allocate, per cui non può in nessun modo essere parte di un ciclo di deadlock.

2. Trova un i tale che:

- $Finish[i] = \text{false}$

- $\text{Request}[i] \leq \text{Work}$

se non esiste nessun i che soddisfa le condizioni, vai al passo 4;

3. Poni:

- $\text{Work} = \text{Work} + \text{Alloc}$
- $\text{Finish}[i] = \text{true}$

quindi vai al passo 2;

4. Se $\text{Finish} == \text{false}$ per qualche i , allora il processo i è in deadlock.

16.2 Prevenzione dinamica, deadlock recovery

Interroghiamoci quindi su cosa fare in fase di rilevamento di deadlock.

Se l'algoritmo di detection è invocato su base arbitraria, potrebbero esserci molti cicli nel grafo di attesa, e quindi potremmo non essere in grado di capire quali dei processi in deadlock hanno in qualche modo "provocato" il deadlock.

L'operazione fondamentale di recupero dal deadlock è quella di **rollback** del processo, cioè riportare il processo ad uno stato precedente a quando il deadlock si è verificato. Un'altra opzione, molto più brutale, è quella di abortire il processo coinvolto.

Possiamo comunque interrogarci su *quale* processo (o processi) fare rollback o abort.

- La soluzione drastica è quella di influenzare tutti i processi coinvolti nel deadlock. Questo chiaramente risolve i problemi ma è distruttivo per il sistema;
- Una soluzione più intelligente potrebbe essere quella di influenzare un processo per ogni ciclo disgiunto trovato nel grafo di attesa.

In questo caso dobbiamo però simulare lo stato ottenuto facendo rollback (o abort) di ogni processo, in modo da assicurarci che la *vittima* selezionata sia effettivamente quella che causa il deadlock.

Un buon approccio a questo metodo è quello di definire una certa *metrica* per la scelta dei processi vittima. Ad esempio, potremmo iniziare a selezionare vittime classificando per:

- La priorità del processo;
- Il tempo per cui il processo ha eseguito, e il tempo rimanente fino alla terminazione (vedere 6.2.5);
- Le risorse che il processo ha già allocato;
- Le risorse che servono al processo per terminare;
- Il numero di processi da terminare (preferire di terminare meno processi possibili);
- Se il processo è interattivo o batch.

Notiamo tutte queste metodologie rischiano sempre la *starvation*: un processo potrebbe essere sempre vittima di abort o rollback, e quindi non riuscire mai a terminare.

16.3 Gestione della memoria

Veniamo quindi alla gestione della seconda risorsa più importante dopo la CPU, cioè la **memoria** principale a disposizione del calcolatore.

L'idea è di offrire a tutti i processi il loro spazio di indirizzamento locale, quindi organizzare una risorsa fisica in più risorse logiche. Vedremo poi come potrebbe essere opportuno definire meccanismi più sofisticati, come divisione di memoria fra S/O e processi, e condivisione di memoria (appunto, *memoria condivisa*) fra più processi.

Gestione della memoria e gestione della CPU hanno dei parallelismi:

- Nella CPU offriamo più CPU virtuali mantenendo il contesto dei processi nel **PCB** (*Process Control Block*).

Nella memoria vorremo offrire più memorie virtuali, allocando altre apposite strutture dati (*descrittori*) che mantengano il contesto relativo al singolo processo;

- In particolare, potremmo voler implementare meccanismi di *swap-in* e *swap-out* che permettono di spostare i contenuti della memoria principale nella memoria secondaria (disco rigido o simili), quando la prima risulta satura. Questo processo è effettivamente parallelo al cambio di contesto CPU.

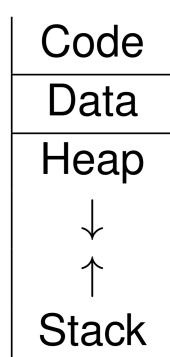
Esistono però anche differenze fra gestione della memoria e gestione della CPU:

- La gestione della CPU è atomica, cioè tutta la CPU viene allocata ad un singolo processo (a meno di approcci multiprocessore, per cui abbiamo ampiamente discusso meccanismi di sincronizzazione).

Nella gestione della memoria, però, è fondamentale che più parti della memoria vengano allocate a processi diversi, per cui è necessario fin da subito prevedere meccanismi di *protezione* della memoria di processi dall'accesso di altri processi.

16.3.1 Immagine di un processo

Abbiamo detto che ogni processo vorrà vedere il suo *spazio di indirizzamento* privato e locale. Vediamo nel dettaglio come questo spazio potrebbe essere organizzato:



- Vogliamo mantenere un primo segmento dedicato al **codice**, cioè al programma vero e proprio in esecuzione nel processo. Si assume che questa si troverà all'inizio dello spazio di indirizzamento, e conterrà l'entry point del programma. Questa sezione conterrà inoltre i **dati in sola lettura** del programma (in quanto si assume che sia il codice che questi dovranno essere in sola lettura, come poi vedremo);

- Vogliamo quindi mantenere un segmento per i **dati** statici del programma, incluse le variabili inizializzate (**dati in lettura/scrittura**) e inizializzate a 0 (**BSS**, *Block Started by Symbol*);
- Vorremo poi mantenere un segmento per l'**heap**, cioè la regione di memoria allocata in memoria dinamica;
- Infine vorremo un segmento per lo **stack**, che come sappiamo si sviluppa dall'alto verso il basso (e quindi si sviluppa a partire dal fondo allo spazio di indirizzamento).

Potrebbe essere utile ripercorrere i passaggi che la nostra toolchain compie in fase di compilazione di un programma, per generare appunto un'immagine di queste sezioni da poi caricare in memoria.

- Il *compilatore* si occupa di prendere il codice sorgente del programma e tradurlo in *oggetti*. Dobbiamo ricordare che gli oggetti non contengono indirizzi veri e propri nello spazio di indirizzamento processo, ma indirizzi *simbolici*, cioè non ancora risolti e che puntano a segmenti magari adesso nemmeno definiti;
- Il *linker* (*collegatore*) si occupa di prendere gli oggetti da noi generati, e quelli già presenti nel sistema sotto forma di *library*, e appunto collegarli in un file eseguibile. Solo in questo passaggio i segmenti vengono preparati, e gli indirizzi simbolici vengono tradotti in indirizzi reali nello spazio di indirizzamento di processo;
- Infine, un *loader* (*caricatore*) si occuperà di prendere i segmenti preparati dal linker e caricarli in memoria (magari impostando come lettura/scrittura gli opportuni segmenti, e azzerando i segmenti di bss, ecc...). A questo punto basterà spostare il PC all'entry point del programma ed eseguire.

16.4 Rilocazione statica/dinamica

Vediamo quindi a come si può svolgere la gestione della memoria vera e propria, cioè come il caricatore può comportarsi quando gli viene fornita l'immagine di un processo.

- L'approccio più semplice è quello della rilocazione **statica**. In questo caso si prevede un caricatore *rilocante*, cioè che si occupa di prendere l'immagine del programma e allocarla in memoria a partire da un certo indirizzo, detto **base**. Chiaramente dovrà preoccuparsi di calcolare il contenuto del program counter (aggiungendo all'entry point la base), e degli indirizzi fisici del programma (ancora, aggiungendo a tali indirizzi la base).

Il problema di tale approccio è che è limitante per quanto riguarda la rilocazione successiva dell'immagine di processo (magari a causa di uno swap-out e successivo swap-in): in questo caso saremo costretti a fare qualcosa di indicibile come ricalcolare gli indirizzi in memoria a tempo di esecuzione (praticamente impossibile), o ricaricare il processo esattamente nello stesso posto di prima (cosa che rende abbastanza inutile prevedere lo swap-out in primo luogo);

- Decidiamo quindi usare un approccio a rilocazione **dinamica**. In questo caso dobbiamo prevedere un nuovo componente hardware, detto **MMU** (*Memory Management Unit*). La MMU ha il compito di implementare una qualche funzione di

traduzione da indirizzo *virtuale* a indirizzo *fisico*:

$$\text{MMU_translate}(\text{virt}) = \text{phys}$$

In questo modo il processore può continuare a pensare ai suoi indirizzi virtuali (che sono ad esso locali), e il compito di traduzione effettiva da questi a indirizzi fisici in memoria è delegato alla MMU, così che i processi non debbano preoccuparsi di dove si trovano effettivamente i loro dati, ma possono specificare indirizzi relativi al loro spazio di indirizzamento locale.

L'implementazione più semplice della MMU è quella parallela all'esempio della rilocazione statica appena fatta. Prevediamo infatti la presenza di un registro **base** e un registro **limite**: quello che la MMU farà sarà controllare che i registri (virtuali) forniti dal processore non cadano fuori dal limite, e quindi sommarvi il registro base. Il S/O avrà il compito di impostare tali registri per ogni processo (in fase di cambio contesto), e così avremo il comportamento della rilocazione statica senza mai dover agire sugli indirizzi fisici nel codice del programma.

Tralasciamo per adesso le complicazioni date dal fatto che il sistema operativo stesso deve accedere alla memoria attraverso la MMU, e quindi subendo traduzioni di indirizzi (solitamente si prevede una finestra, la finestra **FM**, che permette l'accesso diretto a tutta o una parte rilocabile della memoria fisica con traduzioni identità o identità con offset).

16.4.1 Memoria unica/segmentata

Un'altra distinzione ortogonale nella gestione della memoria è data da come si gestisce lo spazio di memoria fisico.

- La memoria è gestita come **unica** (o *flat*) quando si fornisce ad ogni processo una sezione contigua (o non contigua, ma comunque vista come uno o più blocchi contigui) di memoria, poi suddivisa nei vari segmenti (codice, dati, ecc...). Questo è l'approccio usato ad esempio dai sistemi a *paginazione*;
- Un approccio una volta diffuso e oggi caduto in disuso (in particolare, introdotto nell'Intel 286 e sostanzialmente rimosso a partire nell'architettura x86_64 di AMD) è quello della memoria **segmentata**. In questo caso ogni segmento viene gestito a livello hardware, per cui gli indirizzi diventano composti da due interi: il *segmento* e l'*offset* all'interno del segmento. Chiamiamo sistemi che usano questo approccio a *segmentazione*.

In quanto a pro e contro di questi approcci, abbiamo che:

- La memoria **unica** elimina i problemi di frammentazione *esterna* (ad ogni processo si alloca la memoria di cui necessita), ma introduce problemi di frammentazione *interna* (visto che solitamente la memoria si richiede in blocchi, potrebbe essere un problema riempire tali blocchi);
- La memoria **segmentata**, di contro, elimina i problemi di frammentazione *interna* (ogni segmento viene richiesto esattamente della dimensione necessaria), ma introduce problemi di frammentazione *esterna* (bisogna capire come usare lo spazio disponibile per allocare i segmenti).

Un'approccio da usare in questi casi è quello del *compattamento* (o **deframmentazione**): periodicamente, si può analizzare la memoria in modo da spostare in un unico blocco contiguo i segmenti, in modo da massimizzare lo spazio disponibile per le immagini di nuovi processi.

16.4.2 Memoria contigua/non contigua

Vediamo un'altra distinzione ortogonale sulla gestione della memoria, in particolare relativa all'allocazione della memoria *fisica*:

- L'approccio più semplice che possiamo immaginare è quello di allocare la memoria in maniera **contigua**, cioè assicurare che ogni processo ottenga, nel suo spazio virtuale, un blocco contiguo (e opportunamente spostato di un certo offset) in memoria fisica;
- Risulta però molto più comodo concedere l'allocazione **non contigua** della memoria fisica, agendo sulla funzione di traduzione dell'MMU. Questo permette di ridurre a zero la frammentazione esterna, in quanto non c'è mai la necessità di mantenere, in primo luogo, separate e contigue le regioni allocate ai processi.

Questo, però, introduce spesso un quanto minimo di spazio che possiamo allocare (come avevamo già introdotto), ed è l'approccio usato nella *paginazione*. Vediamo che la paginazione ha un overhead non indifferente: la realizzazione di una funzione di traduzione indirizzi che permetta l'allocazione flat e non contigua in memoria fisica richiede infatti tabelle anche consistenti in memoria, che possono essere allocate efficientemente solo usando schemi di allocazione particolari.

16.4.3 Dimensioni della memoria

Infine, vorremo distinguere sulla **dimensione** della memoria virtuale che vogliamo fornire al processo rispetto alla memoria fisica disponibile al sistema:

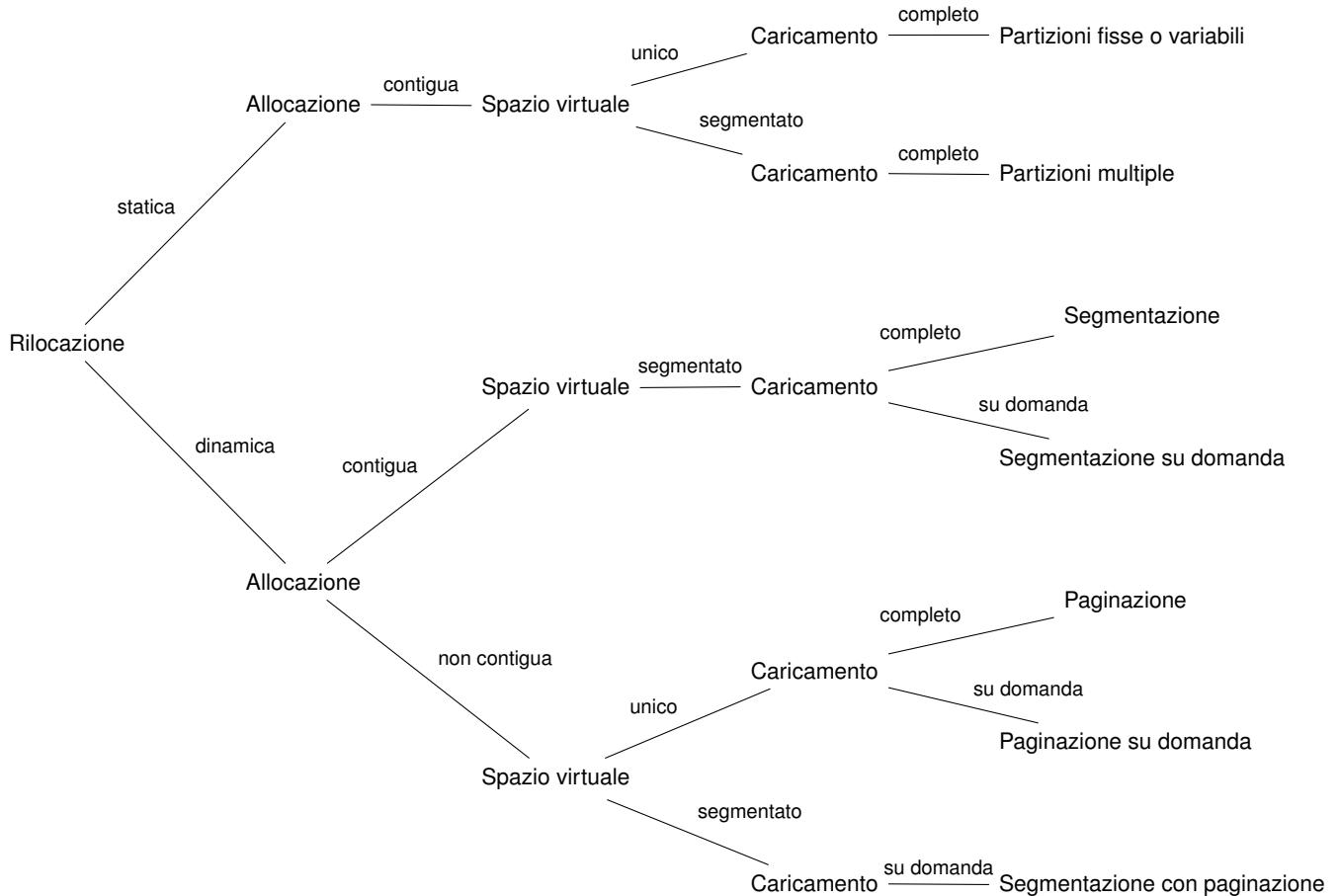
- Se la memoria virtuale è **minore o uguale** alla memoria fisica, saremo in condizioni di *caricamento unico*: potremo caricare l'intera immagine di processo in memoria fisica ed andare avanti;
- Altrimenti, se la memoria virtuale è **maggior**e della memoria fisica, dovremmo implementare meccanismi di *paginazione su domanda*, cioè allocare e deallocare regioni di memoria al processo su base dinamica, cioè quando questo le richiede.

16.4.4 Riassunto sulla gestione della memoria

Riassumiamo quindi i tipi di approcci possibili alla gestione della memoria, sulla base delle caratteristiche ortogonali viste finora:

- Rilocazione statica/dinamica;
- Allocazione contigua/non contigua;
- Spazio virtuale unico/segmentato;
- Caricamento completo/su domanda (abbiamo detto corrispondono a *dimensione della memoria virtuale* minore o uguale/maggior della memoria fisica.)

Vediamo quindi gli approcci possibili sulla base di tali caratteristiche:



17 Lezione del 13-11-25

Cominciamo a vedere le tecniche di gestione della memoria viste nella scorsa lezione, in ordine.

17.1 Partizioni fisse

Questo è un approccio a rilocazione *statica*, allocazione *contigua*, spazio virtuale *unico* e caricamento *completo*.

Ciò che facciamo è suddividere lo spazio di indirizzamento in **partizioni fisse**, dedicate ciascuna ad un processo (una specifica partizione andrà dedicata al S/O). Diciamo, quindi, di allocare D_0 Mbyte al sistema operativo, e poi secondo la dimensione della memoria $D_1 \dots D_n$ partizioni di memoria ai processi $1 \dots n$.

Il problema in cui incorreremo sarà la **frammentazione interna**: all'interno della partizione di ogni processo non è detto che il processo utilizzerà sempre la totalità della partizione. Diciamo infatti che il processo i usi N_i Mbyte all'interno della sua partizione: avremo che $D_i - N_i$ sarà spazio sprecato (frammentazione interna) all'interno della partizione.

Tenendo conto di tutti i processi, si avrà che:

$$F_i = \sum_1^n (D_i - N_i)$$

con F_i appunto la frammentazione interna, quindi la memoria sprecata, complessivamente nel sistema.

Dal punto di vista pratico, il modo più ragionevole per implementare tale soluzione è sfruttare una struttura dati a lista di descrittori di partizione.

- A questo punto si possono effettuare ricerche su tale liste secondo un approccio **first-fit**, cioè caricare il processo nella prima partizione capace di contenere la sua immagine (assumiamo infatti che le partizioni abbiano dimensione variabile). Nel caso in cui tale partizione non esista, chiaramente, siamo in overflow di memoria e il processo non può essere caricato.

Per questo approccio, va bene che le partizioni in lista siano ordinate per indirizzo (cioè appaiano nella lista nell'ordine in cui vengono disposte in memoria);

- Un approccio più conservativo è quello del **best-fit**: in questo caso si cerca di usare la partizione più piccola possibile che può contenere il processo.

In questo caso usare l'ordinamento per indirizzo può risultare inefficiente: bisogna controllare per ogni processo ogni partizione. Può essere quindi più ragionevole ordinare le partizioni per dimensione, a partire dalla più piccola: in questo caso il semplice inserimento first-fit sarà automaticamente di tipo best-fit (la prima partizione che contiene è automaticamente la più piccola che contiene).

Avevamo introdotto in 6.1.1 lo scheduling di **medio termine**. Nell'approccio a partizione fisso questo consiste ad adottare più code di processi, una per ogni partizione, e decidere a quale processo dedicare una determinata partizione.

17.2 Partizioni variabili

Evolviamo l'approccio a partizioni fisse. Nelle **partizioni variabili**, infatti, vogliamo permettere alle partizioni non solo di avere memoria diversa, ma di poter variare la loro dimensione nel tempo. Questo significa che l'approccio è comunque a rilocazione *statica*, allocazione *contigua*, spazio virtuale *unico* e caricamento *completo*.

Diciamo quindi che inizialmente il sistema contiene due partizioni:

- D_0 , dedicata come prima al sistema operativo;
- D_1 , dedicata a qualsiasi processo.

Quando un processo, diciamo P_1 , entra in esecuzione, dividiamo D_1 in due partizioni: una resta D_1 , e viene ridimensionata esattamente alla memoria N_1 richiesta dal processo, mentre l'altra viene denominata D_2 e lasciata libera. Questo processo chiaramente si può ripetere per ogni nuovo processo in arrivo, eliminando sostanzialmente il problema della frammentazione interna.

Quello che chiaramente andiamo ad introdurre è però la **frammentazione esterna**: quando un processo i termina, e libera la sua partizione D_i , ciò che accade è che nello spazio di indirizzamento rimane un "buco" di dimensione N_i . Dopo un certo tempo, all'interno del sistema si vano quindi a formare buchi, che impediscono a successivi processi di essere caricati in maniera contigua in memoria.

Potrebbero infatti verificarsi dove la memoria ha complessivamente abbastanza spazio per contenere un nuovo processo, ma la dimensione di nessuno dei singoli buchi è tale da consentirlo nella pratica.

Anche nelle partizioni variabili è utile discutere sull'approccio alla scelta della partizione dove allocare nuovi processi.

- L'approccio **first-fit** resta il più veloce, in quanto non abbiamo bisogno di riordinare la lista quando si alloca memoria (e quindi si dividono le partizioni).
- Nel caso si scelga di usare un approccio **best-fit**, si riduce chiaramente la frammentazione, a costo di dover riordinare la lista in fase di divisione di partizioni.

L'approccio alternativo è quello di scansionare l'intera lista di partizioni per ogni nuova inserzione, che però avevamo detto parlando delle partizioni fisse è meno efficiente.

Notiamo inoltre che un vantaggio non indifferente che possiamo avere è la possibilità di fare **fusione** delle partizioni al momento di liberazione di spazio dedicato a un processo. Infatti, se la partizione liberata è adiacente ad una partizione libera, si possono combinare le due partizioni in un'unica partizione di dimensione maggiore.

17.2.1 Riassunto sui criteri di allocazione

Possiamo quindi riassumere velocemente le tecniche di allocazione viste nel partizionamento fisso e variabile:

- **First-fit**: è il più veloce, non ha pretese particolari sulle modalità di scansione della lista di partizioni o sul suo ordinamento;
- **Best-fit**: permette di ridurre la frammentazione (*interna* per partizioni fisse ed *esterna* per partizioni multiple), ma è più costoso in fase di creazione di nuovi processi in quanto ha delle pretese sulle modalità di inserzione.

17.3 Segmentazione

L'approccio a **segmentazione** è a rilocazione *dinamica*, allocazione *contigua*, spazio virtuale *segmentato* e caricamento *completo*.

Ciò che vogliamo fare è dividere lo spazio virtuale stesso in più *segmenti*. Ciò significherà che lo spazio di indirizzamento, visto dai processi, diverrà *bidimensionale*: gli indirizzi saranno infatti formati da coppie:

$$x = \langle \text{segmento}, \text{offset} \rangle$$

dove si specifica il segmento di riferimento e l'*offset* all'interno di tale segmento.

Dal punto di vista implementativo, avremmo bisogno di una **tavella dei segmenti**. Ogni segmento sarà composto da 2 valori:

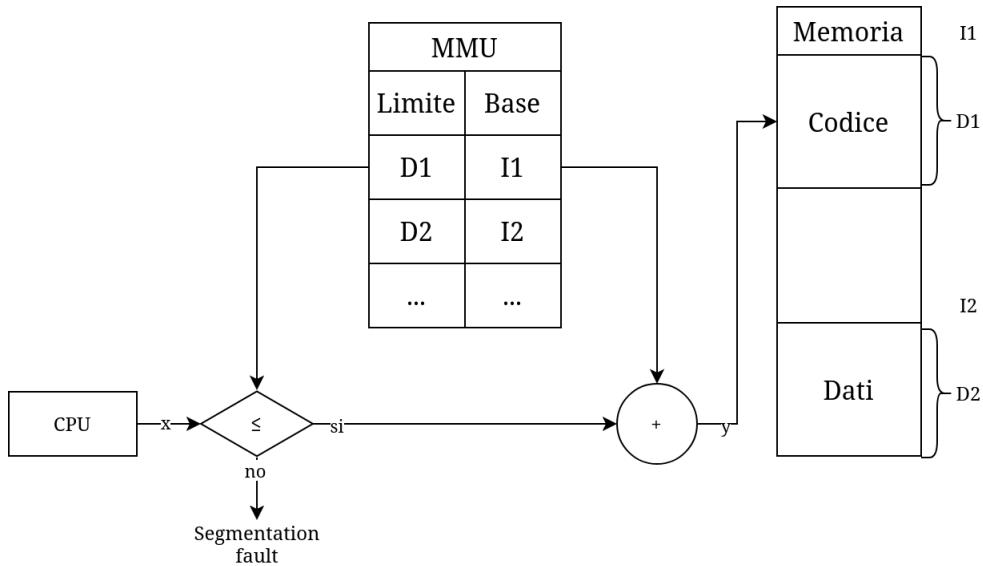
- L'indirizzo al partire dal quale il segmento è allocato in memoria (detto **base**);
- La dimensione del segmento, detta **limite**.

Ogni segmento sarà quindi individuato nella regione di memoria [base, base + limite].

Dotare la macchina di una tale tabella richiede l'introduzione di 2 nuovi registri di controllo, il registro **STBR** (*Segment Table Base Register*), ed il registro **STLR** (*Segment Table Length Register*), contenenti rispettivamente l'indirizzo a partire dal cui si memorizza la tabella dei segmenti e la sua dimensione.

Il supporto hardware alla segmentazione sarà quindi quello di una **MMU** (*Memory Management Unit*), introdotto in 16.4, che ha accesso ai registri STBR e STLR e alla memoria.

Il suo funzionamento può grossomodo essere schematizzato come segue:



dove l'eccezione di *Segmentation fault* viene introdotta appositamente per rilevare accessi al di fuori dello spazio dedicato ai segmenti a tempo di esecuzione. Di base, avremo bisogno di alcuni segmenti di default per ogni programma, fra cui individuiamo:

- Segmento **codice**, che contiene il programma stesso (e magari dati in sola lettura);
- Segmento **dati**, che contiene i dati su cui il programma fa elaborazione;
- Segmento **pila**, che ormai sappiamo è fondamentale all'esecuzione di codice scritto in linguaggio di alto livello.

Questi segmenti corrispondono essenzialmente con i segmenti di immagine di processo che abbiamo introdotto 16.3.1.

Notiamo però che l'uso di un approccio a segmentazione richiede che il processore sia al corrente di quali segmenti riferire in diverse fasi di operazione. Ad esempio:

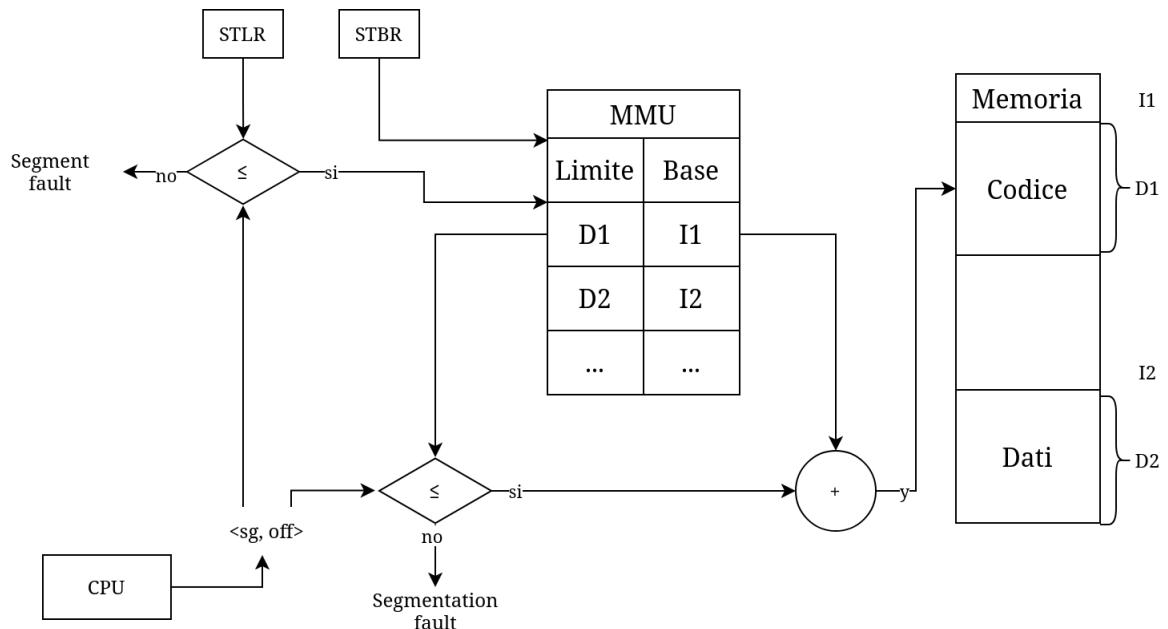
- Nella fase di **fetch**, vogliamo riferirci al segmento *codice* per la lettura di istruzioni;
- Nella fase di **esecuzione**, vogliamo riferirci al segmento *dati* per le operazioni in memoria;
- Nel caso di esecuzione di operazioni in **pila** (ad esempio le **PUSH** e **POP**), vogliamo riferirci al segmento **pila** (per ovvi motivi).

Abbiano quindi che la segmentazione necessita di una modifica del programma: bisogna scrivere programmi che siano a conoscenza di questa funzionalità, e si riferiscono quindi ad indirizzi ben formati (da coppie segmento/offset). L'unica eccezione è chiaramente quella di programmi che richiedono solo un segmento.

Questo perché abbiamo effettivamente fatto una semplificazione non indifferente rispetto ai sistemi reali. Questi infatti:

- Riferiscono la tabella dei segmenti attraverso i registri STLR e STBR: non dobbiamo dimenticarci che la tabella dei segmenti stessa è in memoria fisica;
- Permettono un numero arbitrario di segmenti: si richiede infatti alla CPU di offrire due costanti per indirizzo, cioè come avevamo detto *segmento* e *offset*.

L'MMU aggiornata si può quindi schematizzare come segue:



dove si introduce una nuova eccezione, la *Segment fault*, che non si riferisce ad accessi invalidi all'interno di segmenti ma accesso a segmenti in primo luogo inesistenti.

17.3.1 Descrittori di segmento

Un vantaggio della segmentazione è che possiamo definire nella tabella dei segmenti, oltre ai semplici valori *base* e *limite* di segmento, anche altre informazioni (dette di **controllo**), all'interno del relativo **descrittore di segmento**.

Potremmo infatti prevedere più informazioni in formato tabulare: Dal punto di vi-

Base	Limite	Controllo
Indirizzo del segmento	Dimensione del segmento	Accesso in lettura? scrittura? ecc...

sta implementativo, queste informazioni verranno rappresentate attraverso apposite maschere di bit.

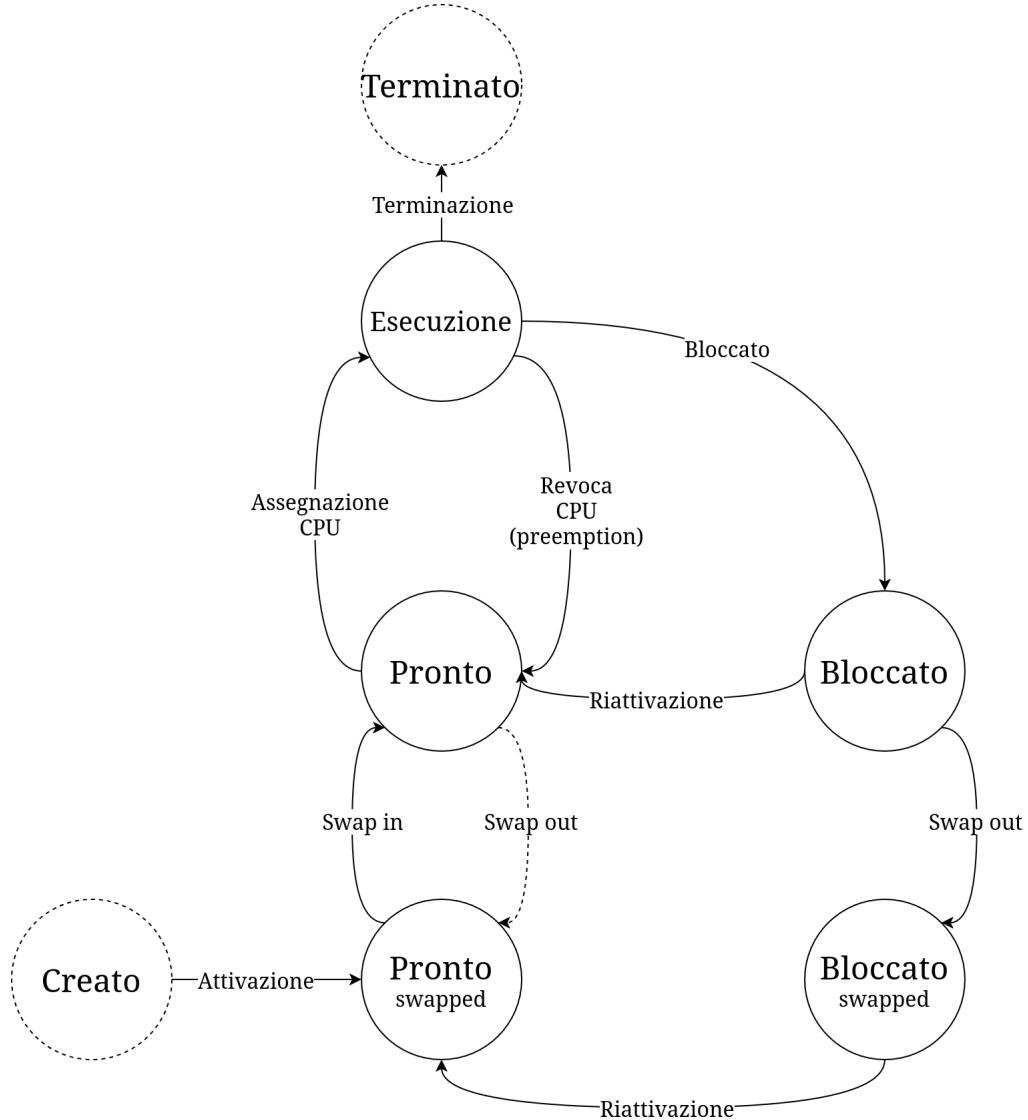
Un esempio tipico è quello di prevedere bit di accesso in lettura (R) o scrittura (W) per segmento. Questo torna utile, prendendo in esame i segmenti visti nella scorsa sezione, come segue:

- Vogliamo che il segmento **codice** sia in sola lettura (quindi R alto e W basso), in quanto come abbiamo introdotto in 5.0.2 il codice può essere condiviso fra processi e non vogliamo che un processo modifichi il codice condiviso con altri processi;
- Vogliamo invece che i segmenti **dati** e **pila** siano in lettura e scrittura (quindi sia R che W alti), per ovvi motivi (il nostro programma dovrà lavorare su qualcosa, anche scrivendo).

17.3.2 Swap di segmenti

La memoria segmentata supporta un altro meccanismo, che è quello dello **swapping** di segmenti dalla memoria centrale alla memoria secondaria.

Aggiorniamo quindi il grafico introdotto in 5.0.1 relativo allo stato dei processi come segue:



Prevediamo quindi due nuovi stati, *Pronto swapped* e *Bloccato swapped*, su cui si transisce rispettivamente dagli stati pronto e bloccato attraverso le primitive `swap_in()` e `swap_out()`.

Facciamo alcune note su queste transizioni:

- Dallo stato *pronto swapped* vogliamo transire allo stato *pronto*, attraverso la `swap_in()`.

Potremmo prevedere anche la `swap_out()` da *pronto* a *pronto swapped*, ma questo non è particolarmente utile: quando un processo è pronto, ci aspettiamo di aver fatto del lavoro per renderlo tale. A questo punto, fare swap out invaliderebbe tale lavoro, richiedendo una successiva `swap_in()` prima di poter mettere il processo in esecuzione;

- Dallo stato *bloccato*, invece, è più che ragionevole fare `swap_out()` nello stato *bloccato swapped*. Non si prevede il contrario (prima si diventa pronti e poi si viene swappati in memoria principale).

17.4 Segmentazione su domanda

Veniamo quindi all'evoluzione naturale della segmentazione, la **segmentazione su domanda**. Questa rappresenta un'approccio a rilocazione *dinamica*, allocazione *contigua*, spazio virtuale *segmentato* e caricamento *su domanda*.

Quello che vogliamo effettivamente fare è fornire lo swap out dei segmenti, e fare il successivo swap in solo quando quei segmenti ci vengono effettivamente richiesti.

Per fare ciò ci dotiamo di nuovi bit all'interno del campo di controllo del descrittore di segmento:

- Bit **U**, detto di *uso*, aggiornato quando si usa (legge o scrive) il segmento;
- Bit **M**, detto di *modifica*, aggiornato quando si fa un'operazione di scrittura sul segmento;
- Bit **P**, detto di *presenza*, indica se il segmento è effettivamente caricato in memoria o se ne è fatto swap out.

L'operazione che vogliamo effettuare è quindi quello di sfruttare il bit P, in fase di accesso al segmento, per lanciare un'eccezione di *segment fault* nel caso tale segmento non sia caricato. In tal caso si cattura l'eccezione e si procede a chiamare la `swap_in()`.

I bit U e M, ricordiamo, forniscono invece delle euristiche utili al S/O per effettuare lo swap out e lo swap in in maniera più efficiente (non copiare segmenti che non sono stati modificati, non deallocare segmenti che vengono usati spesso, ecc...).

17.5 Paginazione

Veniamo quindi alla **paginazione**. Questa rappresenta un'approccio a rilocazione *dinamica*, allocazione *non contigua*, spazio virtuale *unico* e caricamento *completo* (nella versione senza caricamento su domanda).

Abbiamo quindi che lo spazio virtuale viene suddiviso in blocchi di indirizzi virtuali di dimensione fissa (le cosiddette **pagine**). Lo spazio fisico viene suddiviso in blocchi della stessa dimensione delle pagine, detti **frame**.

Ogni pagina corrisponde ad un frame, e pagine consecutive possono essere allocate in frame non necessariamente consecutivi: questo significa che offriamo ai processi spazi di indirizzamento virtuali effettivamente contigui. Vediamo ad esempio come un sistema può avere una divisione delle pagine apparentemente contigua, quando i numeri dei rispettivi frame sono invece non contigui:

Pagina	Sezione	Frame
0	Null	//
1	Sistema	1
2	Text P_1	2
3	Data P_1	3
...	...	//
7	Stack P_1	4

dove nell'esempio si è iniziato a dare qualche informazione semantica su *cosa* contengono le varie pagine.

L'implementazione di un tale sistema richiede di effettuare la traduzione da indirizzo virtuale a corrispondente indirizzo fisico attraverso **tabelle delle pagine** che rappresentano l'associazione fra pagine e frame.

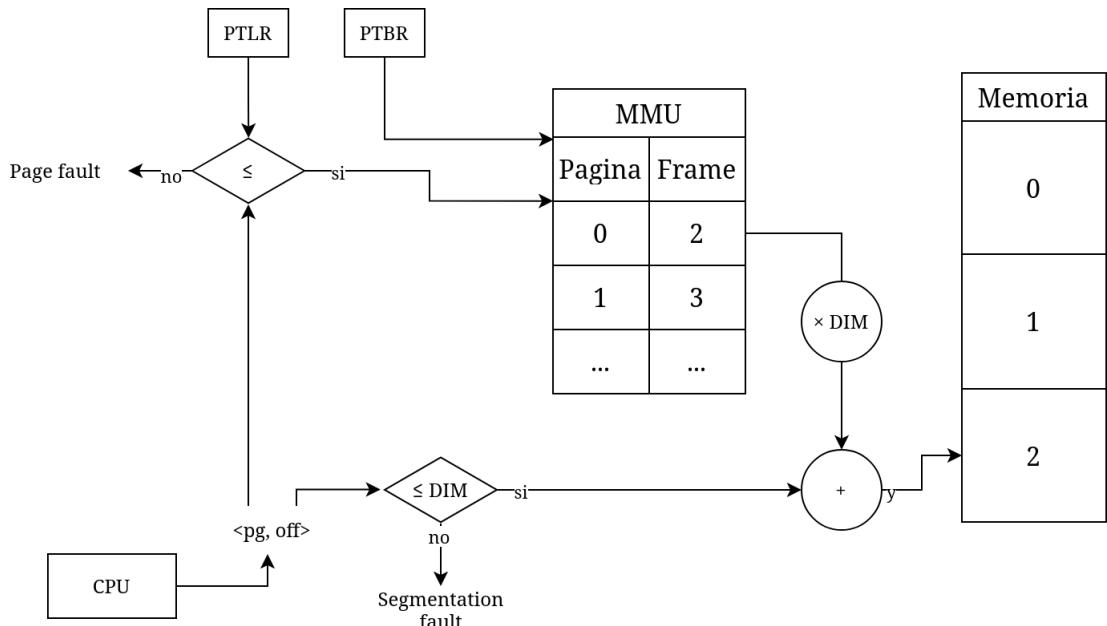
In linea di principio, anche l'MMU in paginazione dovrebbe richiedere dalla CPU una coppia per ogni indirizzo, composta come:

$$x = \langle \text{pagina}, \text{offset} \rangle$$

dove si specifica la pagina di riferimento e l'*offset* all'interno di tale pagina. La differenza sostanziale dalla segmentazione è però che tutte le pagine sono uguali, per cui le operazioni che dobbiamo fare per passare da indirizzi virtuali a indirizzi fisici sono molto più semplici.

Dettagliamo quindi il funzionamento di tale MMU in paginazione. Come per la segmentazione, necessitiamo dell'introduzione di 2 nuovi registri di controllo, il registro **PTBR** (*Page Table Base Register*), ed il registro **PTLR** (*Page Table Length Register*), contenenti rispettivamente l'indirizzo a partire dal cui si memorizza la tabella delle pagine e la sua dimensione. Viene da sé che la MMU avrà accesso ai registri PTBR e PTLR e alla memoria.

Il suo funzionamento potrà quindi essere schematizzato, senza particolari semplificazioni, come segue:



dove si riportano direttamente 2 eccezioni:

- L'eccezione di *Segmentation fault* viene conservata dalla segmentazione, e usata per rilevare accessi al di fuori dello spazio dedicato alla singola pagina a tempo di esecuzione;
- L'eccezione di *Page fault* viene introdotta per segnalare accessi a pagine inesistenti (non presenti nella tabella delle pagine, vedremo poi con bit P di presenza basso).

Vediamo quindi come il fatto che le pagine hanno dimensione fissa ci semplifica il modo in cui:

1. Controlliamo le *segmentation fault*, in quanto ci basterà stabilire se l'offset di pagina è superiore alla dimensione di pagina (solitamente nell'ordine di 1, 2 o 4 KiB);
2. Calcoliamo gli indirizzi fisici, in quanto ci basterà moltiplicare il numero di frame per la dimensione della pagina, e sommare l'offset.

18 Lezione del 18-11-25

Continuiamo a parlare della paginazione.

18.0.1 Descrittori di pagina

Come per i segmenti, possiamo associare alle pagine **descrittori di pagina** che contengono informazioni riguardo all'accesso alla pagina.

Le informazioni tipiche sono, ancora una volta, i bit di accesso in lettura (R) o scrittura (W).

18.1 Paginazione su domanda

Altre informazioni contenute nei descrittori di pagina sono quelle legate alla **paginazione su domanda**, praticamente analoghe a quanto discusso in 17.4:

- Bit **U**, detto di *uso*, aggiornato quando si usa (legge o scrive) la pagina;
- Bit **M**, detto di *modifica*, aggiornato quando si fa un'operazione di scrittura sulla pagina;
- Bit **P**, detto di *presenza*, indica se la pagina è effettivamente caricata in memoria o se ne è fatto swap out.

Come per la segmentazione, il bit P associerà alla pagina un certo significato:

- Bit *P* a 1: la pagina è valida e caricata in memoria;
- Bit *P* a 0: la pagina può non essere valida, oppure può essere valida ma non caricata in memoria.

In questo caso utilizziamo il bit P per lanciare condizionalmente un'eccezione di *page fault* in caso di accesso a pagine con bit P a 0. Capire se la pagina va caricata o l'accesso è effettivamente invalido è quindi compito dell'handler predisposto dal S/O a tale eccezione.

Ricordiamo che la paginazione su domanda è un'approccio a rilocazione *dinamica*, allocazione *non contigua*, spazio virtuale *unico* e caricamento *su domanda*.

18.1.1 Gestione del page fault

Abbiamo detto che il page fault è un'eccezione, generata dal sistema, e completamente incodizionata dal programma.

In altre parole, il processo è ignaro dell'esistenza dei page fault e qualsiasi problema si verifichi deve essere risolto dal S/O. Tutto ciò che potremmo osservare dal lato processo è un aumento del tempo di *turnaround*, cioè dell'intervallo dal tempo di revoca al successivo assegnamento CPU, impiegato chiaramente a caricare la pagina richiesta.

Andiamo quindi a dettagliare come si gestisce effettivamente il page fault:

1. Eseguendo il codice del processo (p_0) in esecuzione, la CPU genera un indirizzo virtuale che corrisponde ad una pagina non caricata in memoria;
2. La MMU riceve tale indirizzo, esplora la *tabella delle pagine del processo* fino all'indirizzo, controlla il bit P è lo trova uguale a 0. Viene lanciata un'eccezione di page fault;
3. Il supporto hardware alle eccezioni (cioè lo stesso a supporto delle interruzioni) carica il puntatore alla routine di gestione del page fault e la mette in esecuzione.
Ora, in memoria assumiamo esista una tabella, detta *tabella delle pagine fisiche*, che associa alle pagine fisiche un'informazione riguardante se quella pagina è libera o meno;
4. Nel caso una pagina fisica (pf) sia trovata libera, si fa una copia (con DMA o altri metodi) dalla *memoria di swap* alla *memoria centrale* della pagina richiesta. L'indirizzo della pagina in memoria di swap può essere, ad esempio, memorizzato nella sezione dedicata all'indirizzo fisico del descrittore di pagina non caricata.
Quindi si aggiorna l'indirizzo fisico del descrittore di pagina con quello di pf, e si mette il bit di presenza a 1;
5. A questo punto i primi 2 passaggi si ripetono, con la differenza che la MMU trova la pagina desiderata con bit P uguale a 1, la tabella delle pagine fisiche contiene la pagina effettivamente richiesta, e il processo può proseguire.

18.1.2 Rimpiazzamento di pagine

La paginazione su domanda richiede spesso che, per liberare spazio in memoria centrale per la nuova pagina, si deallochi una vecchia pagina. Questa viene detta *vittima*, e l'intero processo viene detto **rimpiazzamento**.

Dettagliamo anche questo processo, immaginando di avere un altro processo (p_1) in esecuzione in parallelo a quello del primo esempio:

1. Quando diventa necessario caricare la pagina del processo p_0 , supponiamo che p_1 sia in possesso dell'indirizzo fisico pf (cioè che pf sia indirizzo fisico in uno dei descrittori di pagina di p_1);
2. In questo caso si mette in esecuzione un dato algoritmo di *rimpiazzamento*, e si dealloca la pagina di p_1 nella memoria di swap.
Come abbiamo detto, per ricordare dove abbiamo messo la pagina memorizziamo nella sezione dedicata all'indirizzo fisico, l'indirizzo in memoria di swap dove abbiamo memorizzato la pagina;
3. Alla fine di questo processo, il processo p_1 si trova nella stessa situazione in cui si trovava p_0 prima di accedere alla pagina: questa esiste nella tabelle delle pagine di processo, ma non punta ad un'entrata valida della tabella delle pagine fisiche (bensì ad una pagina di cui si è fatto swap).

L'andamento del rimpiazzamento delle pagine porta naturalmente alla formazione di un *working set* associato ad ogni processo, cioè l'insieme di pagine su cui quel processo fa accesso con frequenza. Il working set inizialmente si espande, ma poi tende a rallentare la sua crescita.

Gli algoritmi di rimpiazzamento che consideriamo saranno quindi:

- L'algoritmo *ottimo* sarebbe quello che rimpiazza le pagine che non verranno più riferite, o almeno, che verranno riferite più tardi nel tempo. Questo però è impossibile da ottenere, in quanto implicherebbe di conoscere il futuro;
- Possiamo quindi adottare un algoritmo **FIFO** (*First In First Out*): la vittima scelta è la pagina che è da più tempo in memoria. Questo è probabilmente l'algoritmo più semplice che potremmo usare.

Un problema dell'approccio FIFO è che esiste un sottoinsieme del working set che probabilmente resterà quasi sempre rilevante all'esecuzione del programma, e che quindi non dovrà caricare. Questo sottoinsieme sarà ad esempio quello che contiene le strutture dati di base del programma, le costanti e le variabili statiche;

- Un'altro algoritmo, che risolve quest'ultimo problema, è il **LRU** (*Least Recently Used*): la vittima è la pagina meno recentemente utilizzata.

Questo chiaramente richiede un qualche tipo di statistica sull'uso delle pagine (a questo tornano utili i bit U e M che avevamo predisposto). Un'alternativa è quella di memorizzare, anziché il bit U, un *timestamp* nel descrittore di pagina. A questo punto basterà semplicemente aggiornare la pagina con timestamp più remoto. Questo però ha chiaramente delle problematiche in termini di overhead (non solo lo spazio che va allocato nel descrittore per il timestamp, ma il tempo impiegato ad aggiornare il timestamp ad ogni operazione sulla pagina).

Può essere utile parlare di *rimpiazzamenti locali*: quando un processo provoca un page fault, si sceglie come vittima una pagina presente nel working set di tale processo. Questo impedisce le interferenze con altri processi, e assegna la penalità associata al page fault al processo stesso che ha generato il page fault.

Se l'algoritmo di scelta delle vittime non è ottimale il sistema può andare in *trashing*. Si dice che il sistema è in trashing quando la percentuale di page fault sugli accessi in memoria supera una certa soglia. La condizione (ideale, fortunatamente non reale) di *trashing completo* è quella in cui ogni accesso in memoria provoca un page fault.

18.1.3 Rimpiazzamento second-chance

Un'altro algoritmo di selezione delle vittime è il cosiddetto algoritmo di rimpiazzamento *second-chance*, o algoritmo dell'*orologio*.

Supponiamo di fare rimpiazzamento locale, e cioè di considerare come possibili vittime solo le pagine nel working set del processo che ha provocato il page fault. Prevediamo quindi un puntatore alla vittima, che inizialmente corrisponde al puntatore alla pagina da più tempo in memoria dell'approccio FIFO. La differenza col FIFO è però data dal fatto che, prima di deallocare la pagina, si controlla il suo bit U. Nel caso questo sia impostato, si annulla e si procede con le pagine successive: questo processo si ripete finché non si trova una pagina con bit U a 0 (e questa verrà deallocated).

Il comportamento ottenuto è quindi che si cerca di evitare di deallocate pagine che sono state usate: si torna sulla prima pagina considerata solo nel caso in cui anche tutte le altre pagine sono già state usate.

18.2 Segmentazione paginata

Concludiamo l'argomento della gestione di memoria parlando della **segmentazione paginata** (o *segmentazione con paginazione*): questo è un approccio a rilocazione *dinamica*, allocazione *non contigua*, spazio virtuale *segmentato* e caricamento *su domanda*.

L'idea della segmentazione paginata è che la CPU produce sempre indirizzi:

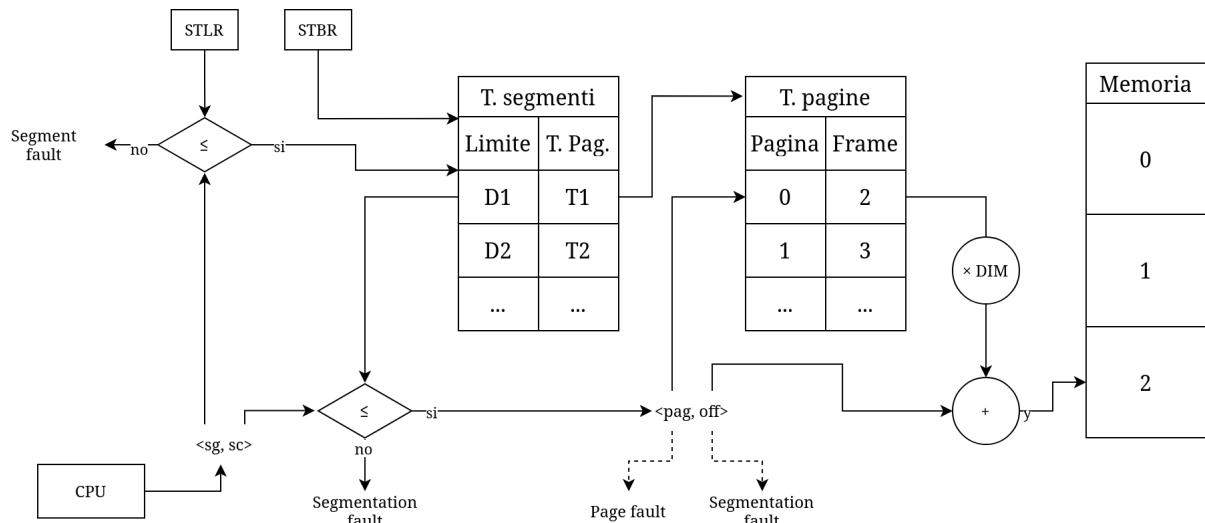
$$x = \langle \text{segmento}, \text{scostamento} \rangle$$

La differenza è però che lo *scostamento* (non ancora offset) viene diviso in due sezioni:

$$\text{offset} = \text{pag} \mid \text{offset}$$

cioè viene usato per indicizzare una pagina e l'offset all'interno di tale pagina. Quello a cui punterà il segmento sarà quindi un *descrittore di segmento*, contenente a sua volta un puntatore alla *tabella delle pagine* associata a tale segmento. Lo scostamento verrà poi usato, come nei normali approcci a paginazione, per indicizzare su tale tabella di pagine.

Questo può essere meglio schematizzato come segue:



dove l'eccezione di *page fault*, che può evidentemente essere lanciata, è stata riportata ma non dettagliata nel funzionamento: chiaramente bisognerà mantenere da qualche parte (nella tabella dei segmenti o come costante del S/O) un numero massimo di pagine per segmento, e controllare la pagina richiesta dallo scostamento con tale numero. La stessa cosa vale per la *segmentation fault* causata dalla paginazione: qui sappiamo, come già visto in 17.5, di avere dimensione di pagina fissa.

Notiamo quindi come questo approccio può portare a 3 tipi di eccezione diversi:

- *Segment fault*: lanciata quando si richiedono segmenti inesistenti (fuori dai bound o non presenti);
- *Segmentation fault*: lanciata quando si effettuano accessi invalidi all'interno del segmento o della pagina;
- *Page fault*: lanciata quando si accede ad una pagina di segmento inesistente (fuori dai bound o non presente).

Questo è l'approccio tipicamente usato da UNIX in segmentazione. Nelle prossime sezioni, infatti, approfondiremo l'approccio che questo S/O usa alla gestione della memoria.

18.3 Gestione della memoria in UNIX

In UNIX la tabella delle pagine fisiche viene detta **core map**. Questa ha quindi il compito di mantenere i frame fisici di pagina, e informazioni riguardo a quale pagina contengono, o se questi sono liberi.

Vengono quindi stabilite 3 variabili per i limiti delle pagine libere:

- *lotsfree*: numero minimo di frame liberi per evitare sostituzione;
- *minfree*: numero minimo di frame liberi necessari per evitare lo swapping dei processi;
- *desfree*: numero minimo di frame desiderabile per un buon funzionamento del sistema.

Chiaramente vale la diseguaglianza:

$$\text{minfree} < \text{desfree} < \text{lotsfree}$$

La sostituzione delle pagine viene quindi effettuata come segue:

- Un processo di sostituzione di pagine, detto **pagedaemon**, esegue periodicamente e sostituisce le pagine solo se:

$$\text{num-frame-liberi} < \text{lotsfree}$$

- Il processo **swapper**, che si occupa invece di effettuare swap in e swap out di processi, fa lo swap out di interi processi se sono soddisfatte le seguenti condizioni:

$$\text{num-frame-liberi} < \text{min-free} \wedge \text{num-medio-frame-liberi} < \text{des-free}$$

La seconda condizione si basa su una statistica fatta sul numero medio di frame liberi. Questo ci permette di effettuare swapping solamente nel caso in cui si hanno picchi di occupazione dei frame fisici che si prolungano per un certo periodo di tempo, mantenendo il numero di pagine libere sotto la media desiderata.

18.4 Gestione delle periferiche

Veniamo allora ad un altro concetto fondamentale dei S/O: la **gestione delle periferiche**. Viene da sé che l'ambiente delle periferiche hardware è estremamente variegato, e l'obiettivo sarà quindi quello di fornire alle applicazioni la possibilità di accedervi in maniera unificata.

Di base, prevediamo un *controllore* per ogni *periferica*: i controllori sono quelli che vengono effettivamente montati sul bus, e vengono visti dal processore come *interfacce*, a cui si accede secondo una modalità *a porte* ben definita (con la differenza di alcune interfacce che potrebbero essere montate *in memoria*).

Come primo esempio dell'eterogeneità delle periferiche disponibili ai moderni calcolatori, anche prima delle differenze tecniche nel modo in cui vi si accede, possiamo notare la grande differenza in termini di *velocità di trasferimento*: passiamo da dispositivi come tastiere e mouse vecchio stile (velocità nell'ordine del Kb/sec) a moderni dispositivi di rete Ethernet (125 Mb/sec).

18.4.1 Sottosistema di I/O

In particolare, prevederemo un **sottosistema di I/O**, che avrà il compito di nascondere i dettagli hardware dei controllori dei dispositivi.

Dal sottosistema di I/O ci aspettiamo il compimento dei seguenti compiti:

- Definizione di uno *spazio dei nomi* con cui identificare in maniera univoca i dispositivi (spazi di indirizzamento PCI, ecc... non verranno visti nel dettaglio in questo corso);
- Gestione dei *malfunzionamenti* dei dispositivi, senza che debbano occuparsene le applicazioni;
- Garanzia della *sincronizzazione* tra l'attività di un dispositivo e del processo che lo ha attivato: il programmatore dovrà essere libero di realizzare la logica del programma senza preoccuparsi della sincronizzazione esplicita con i dispositivi;
- *Bufferizzazione*, cioè disaccoppiamento temporale e spaziale tra processi e periferiche (i processi forniscono *buffer* in memoria condivisa o privata, che vengono quindi riempiti dalla periferica, attraverso il S/O, in differita).

18.4.2 Organizzazione logica dell'I/O

Dal punto di vista **logico** il sottosistema di I/O è estremamente stratificato.

- A livello *utente*, avremo che i **processi applicativi** (cioè le applicazioni) si interfacciano con **library** che espongono funzionalità offerte da *interfacce applicative* fornite dal S/O;
- A livello *S/O*, si offre l'**interfaccia applicativa (I/O API)** vera e propria per la gestione delle periferiche.
 - Questa si interfaccia con una certa quantità di strutture e routine gestite dal S/O che esistono a priori dalle periferiche. Tali strutture formano la cosiddetta **interfaccia device independent** del sottosistema di I/O. Per tornare all'esempio di Unix, si ha che l'associazione periferica \leftrightarrow file è parte dell'interfaccia device independent. Sono dello stesso tipo anche tutte le varie primitive `open()`, `close()`, `read()`, `write()`, ecc...
 - Al di sotto dell'interfaccia device independent si trova chiaramente una parte di interfaccia **device dependent**. Questa è composta da strutture e routine gestite dal S/O che esistono direttamente in funzione delle periferiche montate nel sistema.

Parte dell'interfaccia device dependent sono i *driver*, cioè sostanzialmente dai gestori delle *interruzioni* lanciate dai dispositivi. Chiaramente i driver sono strettamente legati ai dispositivi che gestiscono, in quanto devono conoscere i loro dettagli di funzionamento.

Sempre riconducendosi all'esempio di Unix, abbiamo che anche questo livello implementa le sue `read()`, `write()`, con la particolarità che queste si preoccupano del funzionamento effettivo della periferica. In questo, la chiamata della `read()` di livello indipendente si tradurrà in una chiamata alla `read()` di livello dependent, e lo stesso per la `write()`, ecc...

- Alla fine della gerarchia c'è il livello *hardware*, formato dall'**interfaccia di accesso ai controllori**, e quindi coi **controllori** veri e propri.

18.4.3 Bufferizzazione

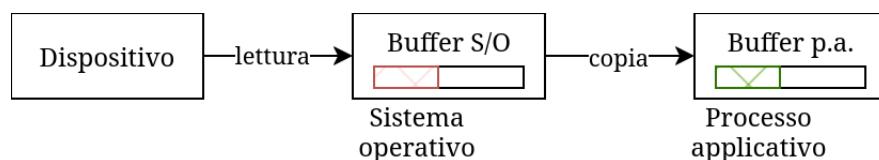
Vediamo nel dettaglio l'attività di **bufferizzazione** svolta dal sottosistema di I/O.

Quello che vogliamo fare è, ad esempio per un'operazione di lettura, permettere ai processi che chiamano primitive di ingresso di fornire un *buffer*. Questi verranno quindi bloccati per la durata del trasferimento, e il S/O si impegnerà a riempire tale buffer con quanto ottenuto dalla periferica.

Il disaccoppiamento che cerchiamo è sia *spaziale* che *temporale*:

- **Spaziale**, in quanto non chiediamo al programmatore di conoscere le dimensioni di buffer ideali per la periferica, ma ci prendiamo la briga di gestire eventuali buffer intermedi e di riempire il buffer da egli fornito;
- **Temporale**, in quanto il buffer viene riempito in differita, dopo che viene fornito dal programmatore.

Questo approccio richiede la presenza di 2 buffer:



- Un buffer di **S/O**, la cui dimensione è effettivamente dettata dal dispositivo stesso, e che il sistema operativo usa per gestire immediatamente le operazioni di scrittura dalla periferica verso il bus;
- Un buffer di **processo applicativo**, dichiarato dal programmatore, all'interno del quale ci aspettiamo di trovare i dati (nel nostro esempio dopo una `read()`). Questo viene riempito attraverso operazioni di *copia* dal buffer di S/O.

18.4.4 Funzioni indipendenti

Vediamo quindi le funzioni predisposte *a priori* dei dispositivi, cioè quelle presenti al cosiddetto livello **device independent**:

- **Spazio dei nomi** dedicato ai dispositivi: questo deve permettere un qualche tipo di associazione:

`< nome simbolico > ==> &id_desc`

dove `&id_desc` è un puntatore a una qualche struttura dati `id_desc`, detta *descrittore di dispositivo*, che rappresenta informazioni riguardo a tale dispositivo.

Ricordiamo che questo in Unix è effettuato attraverso la metafora dispositivo \leftrightarrow file, e quindi le funzionalità di *naming* dei dispositivi sono implementate direttamente all'interno del file system. Possiamo infatti anticipare che il file system Unix è gestito attraverso descrittori di file detti *inode*, e che alcuni *inode* particolari hanno il compito di rappresentare proprio i dispositivi.

- Gestione dei **malfunzionamenti**;
- Gestione degli accessi **concorrenti** allo stesso dispositivo (questa si realizza lato S/O usando le primitive di sincronizzazione già ampiamente studiate).

18.4.5 Funzioni dipendenti

Vediamo quindi le funzioni *dipendenti* dai dispositivi, cioè quelle del cosiddetto livello **device dependent**.

Qui il nostro compito è quello di definire primitive come la `read()`

```
1 read(descrittore, buffer, dim_buffer);
```

che offrono un'interfaccia univoca all'accesso ai dispositivi.

Il supporto S/O delle funzioni dipendenti è dato, come abbiamo anticipato, dai **driver**. Il driver è infatti quella collezione di strutture e routine (principalmente gestori di interruzione) predisposti all'interoperazione fra S/O e dispositivo, ed è l'unico componente software che si preoccupa effettivamente delle modalità di funzionamento dell'hardware del dispositivo.

Abbiamo già detto che il driver si interfaccia in particolare con il *controllore* di dispositivo, e non con il dispositivo stesso. Un controllore generico è formato dalle seguenti componenti:

- Registri di **controllo**, su cui la CPU può scrivere, che dettano al dispositivo quali operazioni compiere;
- Registri di **stato**, da cui la CPU può leggere, che forniscono informazioni riguardo allo stato del dispositivo;
- Registri di **buffer** (o *dati*), su cui la CPU può leggere, scrivere o entrambi a seconda del tipo di dispositivo. Questi si occupano della lettura e scrittura effettiva di dati da e sul dispositivo.

Il funzionamento è quindi il seguente:

1. La CPU scrive comandi sui registri di controllo del controllore, legge lo stato dai registri di stato del controllore, e scambia informazioni attraverso i registri dati del controllore;
2. Il controllore invia *segnali* al dispositivo, e legge *dati* dal dispositivo. Segnali e dati scambiati col dispositivo sono direttamente influenzati da quanto il controllore ha ricevuto comunicando con la CPU.

In questo, il controllore si comporta effettivamente da *buffer* fra CPU e dispositivo.

I controllori hanno poi quasi sempre la possibilità di generare **interruzioni** per la CPU (attraverso un componente intermedio detto *controllore di interruzioni*, che si occupa di organizzare gerarchicamente e bufferizzare le interruzioni per la CPU). Le interruzioni sono ormai fondamentali alla gestione dei dispositivi: l'esempio tipico è quello di generare un'interruzione quando si ha un aggiornamento dei bit di stato (ad esempio per segnalare nuovi dati da leggere o la terminazione di un'operazione in scrittura).

19 Lezione del 02-12-25

Riprendiamo dopo tempo immemore l'argomento delle periferiche.

19.1 Processi esterni

Il comportamento del controllore di un dispositivo è assimilabile ad un processo, che chiamiamo **processo esterno**. Il controllo del comportamento del processo esterno esterno è influenzato dai bit del registro di controllo, che vengono aggiornati da un corrispondente *processo interno*.

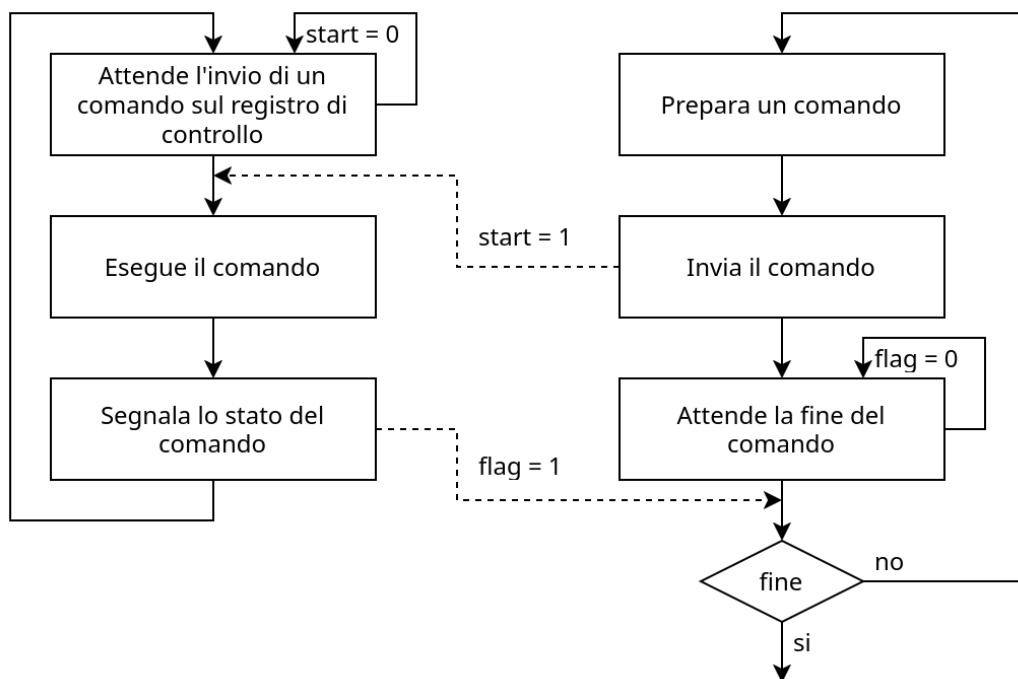
Ipotizziamo un processo esterno che si attiva quando viene alzato un dato bit di controllo, detto *bit di start*, e aggiorna un dato bit di stato, detto *bit di flag*. Il suo funzionamento potrebbe quindi essere il seguente:

1. Il PE attende l'invio di un comando sul registro di controllo;
2. Quando il bit di start transisce a 1, esegue il comando richiesto;
3. Successivamente all'esecuzione del comando aggiorna il bit di flag e torna allo stato 1.

Il corrispondente processo interno si comporta quindi come segue:

- Il PI prepara un comando;
- Invia il comando impostando il bit di start a 1;
- Attende la fine del comando aspettando che il bit di flag transisca a 1;
- Finisce o eventualmente ripete tornando allo stato 1.

Il comportamento appena descritto può essere riassunto dal seguente schema, dove si mettono in evidenza (con linee tratteggiate), i cambiamenti di stato che rappresentano una comunicazione fra PE e PI:

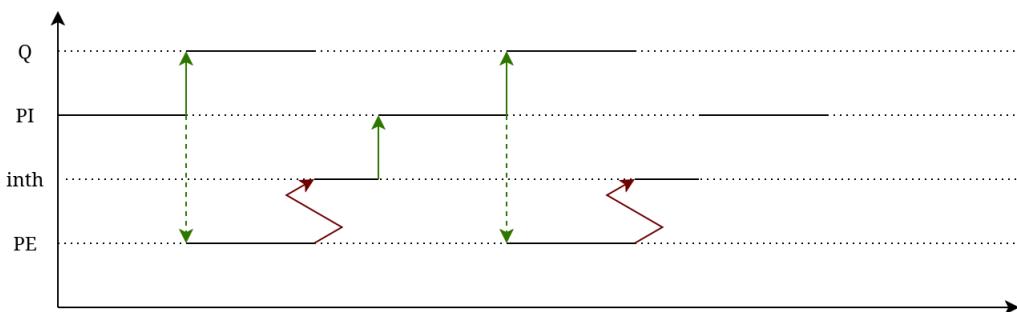


19.1.1 Gestione dei processi esterni

Assumiamo che i processi esterni modellizzano il comportamento di periferiche che eseguono il loro codice, o comunque portano avanti le loro operazioni, in parallelo al sistema. Sarà quindi vero che i PE eseguono solitamente in parallelo ad altri processi (anche solo processi utente).

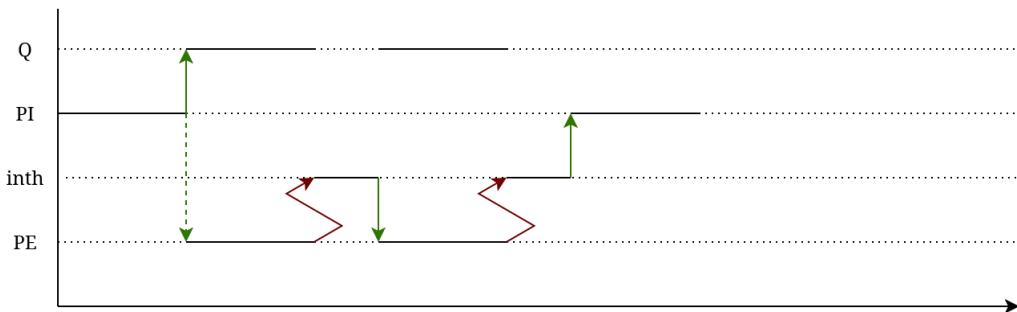
La transizione da processo utente a PI, quando un PE si mette in attesa (magari alzando un bit di flag), viene fatta sfruttando il meccanismo dell'*interruzione*. In particolare, prevediamo di associare un'interruzione alla modifica dei bit di flag da parte del dispositivo, e di predisporre nel sistema un handler per tale interruzione che rimetta in esecuzione il PI (che chiamiamo *inth*).

Possiamo visualizzare tale comportamento su un grafico:



Notiamo che l'esecuzione continua dell'inth, e della transizione dall'inth al PI, porta ad avere diversi cambi di contesto per ogni byte (o quanto di informazione su cui è tattato il buffer del dispositivo), cosa che chiaramente porta ad un overhead non indifferente. Possiamo pensare di migliorare la situazione includendo all'interno del inth stesso routine per la gestione di buffer da più di un byte (cioè fare in modo che sia l'inth a tenere conto dei byte letti / scritti finora, e a leggere dal / fornire al dispositivo il prossimo se necessario).

In tal caso il grafico ha il seguente aspetto:



19.1.2 Descrittori di dispositivo

All'interno del sistema, i dispositivi andranno raccontati da appositi **descrittori**. Questi descrittori vivranno all'interno dei driver (cioè assieme alla routine pensate per manipolarli, e per usarli per modificare lo stato dei dispositivi).

Routine di lettura scrittura fornite al programmatore sotto forma di API come `read()` e `write()`, nonché l'inth stesso, dovranno interagire con questo descrittore per portare avanti operazioni di trasferimento su e da dispositivo.

Vediamo quindi la struttura, a grandi linee, del descrittore di dispositivo:

- **Indirizzi dei registri**, relativi per ogni funzione alla solita tripla:
 - Indirizzo del registro di *controllo* (chiamato `ctl`);
 - Indirizzo del registro di *stato* (chiamato `sts`);
 - Indirizzo del registro di *dati* (chiamato `dat`).
- La configurazione di tali registri viene fatta in fase di bootstrap del sistema (ricordiamo ad esempio che il bus PCI prevede uno spazio di indirizzamento specifico, quello di *configurazione*, per il rilevamento dei dispositivi e la configurazione dei loro registri);
- Dati relativi alla **sincronizzazione** sul dispositivo (o meglio sul processo esterno), fra cui:
 - Un **semaforo** che indica se un dato è disponibile, detto appunto `dato_disponibile` (su cui chiameremo le classiche `signal()` e `wait()`);
 - Un **contatore** che indica il numero di dati da trasferire, che chiamiamo semplicemente `contatore`.
 - Vorremo poi un **puntatore al buffer** in memoria su o da cui stiamo facendo trasferimento, che chiamiamo semplicemente `puntatore`.
- Infine, terremo traccia di un qualche flag rappresentante l'**esito del trasferimento**, che chiamiamo `esito`.

19.1.3 Vista di un device driver

Vediamo quindi l'implementazione di un semplice device driver, assumendo un dispositivo in sola lettura (si implementa solo la `read()`).

Notiamo che ci aspettiamo di implementare il secondo approccio visto in 19.1.1, cioè quello dove è l'inth a occuparsi di tenere conto dei byte letti / scritti finora, e a leggere dal / fornire al dispositivo il prossimo se necessario.

Infine, per i valori che restituiscono errore (come `esito` del descrittore di dispositivo, o i registri di stato del dispositivo), assumiamo la costante `ERR_CODE ≠ 0` come codice di errore generico.

```

1 // descrittore del dispositivo
2 struct des {
3     int ctl;
4     int sts;
5     int dat;
6
7     sem dato_disponibile;
8     int contatore;
9     char* puntatore;
10
11    int esito;
12 }
13
14 // primitiva read, legge @cont byte in @pbuff dal disp. @fd
15 int read(int fd, char* pbuf, int cont) {
16     // ottiene il descrittore di dispositivo
17     // (assumiamo un'array indicizzata)
18     des* disp = &descr[fd];
19
20     // imposta il buffer

```

```

21     disp->contatore = cont;
22     disp->puntatore = pbuf;
23
24     // attiva dispositivo
25     out(disp->ctl, 1);
26
27     // aspetta per il termine dell'operazione
28     wait(disp->dato_disponibile);
29
30     // raccoglie l'esito
31     if(disp->esito == ERR_CODE) return -1;
32
33     // restituisce il numero di byte ancora da leggere
34     return(cont - disp->contatore);
35 }
36
37 // interrupt handler
38 void inth() {
39     // assumiamo lettura
40     // inoltre, assumiamo che des* disp sia noto
41
42     // legge registro di stato
43     char sts;
44     in(disp->sts, sts);
45     if(sts == ERR_CODE) {
46         < gestione di errore (device-specific) >
47
48         if(< errore non recuperabile >) {
49             // se non si puo recuperare, chiudi la comunicazione qui
50             disp->esito = ERR_CODE;
51             signal(disp->dato_disponibile);
52         }
53     }
54
55     // legge byte da dispositivo
56     char b;
57     in(disp->dat, b);
58
59     // copia nel buffer utente e aggiorna descrittore
60     *(disp->puntatore++) = b;
61     disp->contatore--;
62
63     // ha terminato?
64     if(disp->contatore != 0) {
65         // se no, chiede altro byte
66         out(disp->ctl, 1);
67     } else {
68         // se si, segnala
69         disp->esito = 0; // corretta terminazione
70         signal(disp->dato_disponibile);
71     }
72 }
```

19.2 Dispositivo timer

Vediamo nel dettaglio una periferiche reale, cioè il **timer**, che dovrà essere un *generatore di eventi programmabile*, su base temporale.

Il timer può essere molto utile, ad esempio solo per realizzare una primitiva di `sleep()` all'interno del sistema. Ci occupiamo adesso di definire il comportamento del controllore e realizzarne un driver.

19.2.1 Descrittore del timer

Vediamo quindi il descrittore del dispositivo timer.

Notiamo che questo dispositivo sarà capace di mantenere più timer *virtuali* contemporaneamente. Ciò sarà implementato gestendo più timer virtuali nel descrittore e aggiornandoli cumulativamente all'arrivo di eventi (interruzione) da parte del singolo timer fisico installato nel sistema.

- **Indirizzi dei registri**, relativi per ogni funzione alla solita tripla:
 - Indirizzo del registro di *controllo* (chiamato `ctl`);
 - Indirizzo del registro di *stato* (chiamato `sts`);
 - Indirizzo del registro di *dati* (chiamato `dat`).
- Dati relativi alla **sincronizzazione** sul timer, fra cui:
 - Un'array di **semafori** che indicano se i timer hanno terminato, detti `fine_attesa[N]`;
 - Un'array di interi che rappresenta i ritardi di ogni timer, detta `ritardi[N]`.
- Infine, il classico flag di `esito`.

19.2.2 Driver del timer

Vediamo allora un semplice driver per il timer appena visto:

```

1 // descrittore del timer
2 struct des {
3     int ctl;
4     int sts;
5     int dat;
6
7     sem fine_attesa[N];
8     int ritardi[N];
9
10    int esito
11 }
12
13 des tim; // istanza globale
14
15 // primitiva delay, aspetta per @ritardo
16 void delay(int ritardo) {
17     // usiamo il proc. corrente per indicizzare il timer
18     int proc = < processo corrente >;
19
20     // configura il timer
21     tim.ritardo[proc] = ritardo;
22
23     // aspetta il timer
24     wait(tim.fine_attesa[proc]);
25 }
```

```

27 // interrupt handler
28 void inth() {
29     for(int i = 0; i < N; i++) {
30         if(descr.ritardo[i] != 0) {
31             descr.ritardo[i]--;
32             if(descr.ritardo[i] == 0)
33                 signal(descr.fine_attesa[i]);
34         }
35     }
36 }
```

19.2.3 Dispositivi a blocchi

Iniziamo a parlare dei *dispositivi a blocchi*, e in particolare dei **dischi**. Questi sono dispositivi che permettono di memorizzare, seppur in maniera più lenta rispetto alla RAM, vaste quantità di dati per l'archiviazione a lungo termine.

Storicamente (ed ancora oggi) i dischi venivano realizzati come dischi magnetici veri e propri (**HDD**, da *Hard Disk Drive*), mentre oggi si stanno diffondendo sempre più dischi allo stato solido (**SSD**, da *Solid State Drive*). Una discussione più approfondita delle specifiche hardware si può trovare in <https://raw.githubusercontent.com/seggiani-luca/appunti-ce/638d3abf2e1d473632b575401582203c3b113c82/master/master.pdf>.

Ciò che basta sapere in questo contesto è che un disco è formato da più **tracce** disposte radialmente, ed ogni traccia è divisa in **settori**. Spesso, inoltre, più dischi sono sovrapposti fra di loro a formare **cilindri**.

19.2.4 Scheduling di dischi

I dischi sono dispositivi ad accesso *sequenziale*. Questo significa che è necessario un'algoritmo di **scheduling** degli accessi a disco, che minimizzi il movimento della testina di lettura e quindi i tempi medi di accesso.

Ne vediamo alcuni:

- **FCFS (First Come First Served)**: è l'algoritmo più semplice, dove gestiamo le richieste di accesso nell'ordine in cui arrivano. Lato software è estremamente semplice e veloce, ma lato hardware richiede potenzialmente il numero massimo di spostamenti della testina;
- **SSSF (Shortest Seek Time First)**: sceglie l'accesso più vicino alla posizione corrente della testina. In questo caso riduciamo il numero di movimenti della testina, ma l'overhead non è più trascurabile:
 - Ogni volta che gestiamo una richiesta dobbiamo scorrere tutta la coda delle richieste per individuare quella più vicina;
 - C'è il rischio di starvation (se entrano spesso richieste con seek time minore).
- **SCAN** (o *algoritmo dell'ascensore*): è ispirato dal funzionamento degli ascensori. Si decide una direzione di andamento della testina, e si inizia a gestire le richieste seguendo tale direzione. Una volta arrivati ad un estremo (alla richiesta di indice più basso o più alto) si cambia direzione.

20 Lezione del 03-12-25

20.1 File system

Nella maggior parte dei S/O general purpose odierni è definito un componente destinato alla gestione dei **file system**. Un *file system* è un sistema che governa l'organizzazione e l'accesso ai *file*, spesso allocati su dispositivi a blocchi come i *dischi*.

Con un file system andiamo quindi a realizzare tutta una serie di concetti astratti, fra cui:

- Il **file**, unità logica di memorizzazione dati;
- La **directory** (o *direttorio*), insieme di file o altre directory;
- La **partizione**, un insieme di file associati ad un particolare dispositivo fisico (o una sua porzione).

File e directory rappresentano i nodi di una struttura ad *albero*. Le caratteristiche di file, directory e partizione sono del tutto indipendenti dalla natura e dal tipo di dispositivo fisico utilizzato. Sono, appunto, *astrazioni*.

20.1.1 Organizzazione logica del file system

Il file system è, come tutti i moduli del sistema, una struttura gerarchica:

- Il livello più alto è quello *logico*, dove esiste solamente l'astrazione di file e directory. Questo è il livello che viene offerto alle *applicazioni*;
- Segue il livello di *accesso*, che governa le modalità in cui si accede ai file (sequenziale, diretta, ecc...), e i vari meccanismi di protezione che possono essere implementati;
- Segue il livello di *organizzazione fisica*, che riguarda l'allocazione dei file nei blocchi fisici. Vediamo infatto ogni *disco virtuale* (più informazioni sotto) come un vettore di blocchi fisici, che vengono quindi distribuiti fra file;
- In fondo c'è quindi l'astrazione del *dispositivo virtuale*, costruita al di sopra dell'hardware (la memoria secondaria, cioè i dispositivi a blocchi), e visto come già introdotto come un vettore lineare di blocchi fisici.

20.1.2 File

Un **file** è un insieme di informazioni, rappresentate secondo un insieme di *record logici* (bit, byte, parole, ecc...). In UNIX il record è 1 byte.

Ogni file è ulteriormente caratterizzato da un insieme di *attributi*, cioè metadati che contengono il file:

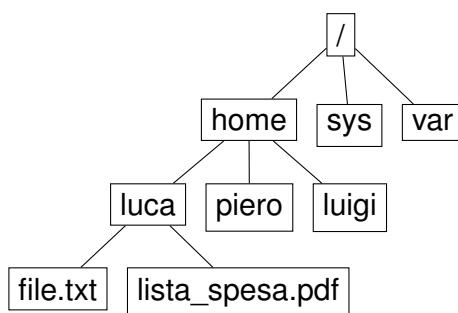
- **Nome** del file;
- **Tipo** del file (si distingue fra file eseguibili, batch, di testo, ecc...);
- **Indirizzo** del file nella memoria secondaria;
- **Dimensione** del file, cioè il numero di record da cui è composto in memoria secondaria;

- **Timestamp** di creazione del file, e di ultima modifica.

Nei sistemi operativi multiutente, inoltre, si vuole includere informazioni riguardo all'utente **proprietario** del file, e i **permessi** degli altri utenti riguardo alla manipolazione del file.

20.1.3 Alberi di file

Abbiamo tutti nota l'organizzazione ad albero dei moderni file system. A scopo di ripasso, notiamo che si definisce un primo directory detto *radice*, rappresentato in sintassi come `/`. Per individuare un oggetto (un'altra directory o un file) a partire dalla radice si continuano a frapporre nomi di oggetti fra `/`. Ad esempio, `/home/luca/file.txt` cerca il file `file.txt`, nella directory `luca`, a sua volta nella directory `home`, a sua volta nella directory radice:



Notiamo che in verità la presenza di *link* all'interno del file system, cioè riferimenti allo stesso file fisico da più locazioni, rende l'astrazione migliore non più l'albero ma il **DAG** (*Directed Acyclic Graph*), cioè il grafo aciclico diretto (che rispecchia un albero ma permette la connessione fra nodi con radici mutualmente diverse).

20.1.4 Operazioni su file system

Su un file system dobbiamo permettere, di base, un insieme minimo di operazioni:

- **Creazione/cancellazione directory**: modificano la struttura logica del file system, aggiungendo/eliminando rami al grafo che rappresenta il file system;
- **Aggiunta/cancellazione file**: inseriscono nuovi dati all'interno del file system;
- **Listing**: generano listati dei contenuti delle directory;
- **Attraversamento directory**: permettono il passaggio da una directory all'altra, e quindi la navigazione del file system.

20.1.5 Descrittore di file

Per realizzare l'astrazione del file dobbiamo implementare un qualche tipo di struttura dati per la sua rappresentazione, cioè un **descrittore di file**.

Questo conterrà gli attributi già notati in 20.1.2. I descrittori di file devono essere memorizzati in modo persistente, e quindi vengono allocati in apposite strutture in memoria secondaria. In particolare, ricordiamo la terminologia UNIX di *i-node*, *i-list* e *i-number*.

Per la rappresentazione delle directory, che assumiamo come categorie particolari di file, dobbiamo sicuramente mantenere collegamenti ai descrittori di tutti i file che questa contiene.

Notiamo a questo punto la differenza nella gestione del filesystem fra i due sistemi operativi principali:

- In Microsoft Windows si adotta un approccio *distribuito*, cioè le informazioni sui file sono contenute in strutture dati rappresentanti le loro directory, in maniera locale;
- In UNIX si adotta invece un approccio *centralizzato*, cioè dove le informazioni su file e directory sono contenute in tabelle centralizzate gestite dal S/O.

:wa

20.1.6 Accesso ai file

Il compito del S/O è quello di consentire l'accesso *on-line* ai file. Le operazioni permesse saranno quelle di **accesso** ai file, cioè:

- **Lettura** di record logici dal file;
- **Scrittura** su file, cioè inserimento di nuovi record logici all'interno del file.

Ognuna di queste operazioni richiederebbe la localizzazione di informazioni sul disco, fra cui ad esempio:

- Gli indirizzi dei record logici a cui accedere;
- Gli altri attributi del file;
- I record logici.

Per migliorare l'efficienza, il S/O mantiene in memoria centrale una struttura dati che registra i file attualmente in uso. Per ogni file aperto vogliamo mantenere il puntatore al file in memoria centrale (più informazioni sotto), il descrittore del file, e la sua posizione nel disco. I file aperti verranno quindi *mappati* in memoria centrale, cioè temporaneamente copiati, durante l'accesso, per aumentare la velocità.

Le operazioni necessarie saranno:

- In fase di **apertura** del file, introduzione di un nuovo elemento nella tabella dei file aperti e eventuale mapping in memoria (se non era già stato fatto) del file;
- In fase di **chiusura** del file, salvaggio del file in memoria secondaria e eliminazione dell'elemento corrispondente della tabella dei file aperti.

20.1.7 Metodi di accesso

L'accesso ai file può avvenire secondo varie modalità:

- **Accesso sequenziale**

In questo caso il file è inteso come una sequenza $[R_1, R_2, \dots, R_N]$ di record logici. Per accedere al record R_i , bisogna necessariamente accedere prima ai precedenti R_1, \dots, R_{i-1} record.

In questo caso possiamo prevedere operazioni come:

- `readn(f, &v)`, che permette la lettura del prossimo record logico (col riferimento `$v`) del file `f`;
- `writen(f, v)`, che permette la scrittura del prossimo record logico (ottenuto col riferimento `v`) nel file `f`.

Ad ogni modo, il nodo centrale dell'accesso sequenziale è che ognuna di queste operazioni posiziona il puntatore del file al record successivo a quello letto;

- **Accesso diretto**

In questo caso il file è inteso come un insieme $\{R_1, R_2, \dots, R_N\}$ di record logici. Noto l'indice i , si può accedere direttamente all' i -esimo record.

In questo caso possiamo prevedere operazioni come:

- `readd(f, i, &v)`, che permette la lettura del i -esimo record logico (col riferimento `$v`) del file `f`;
- `writed(f, i, v)`, che permette la scrittura dell' i -esimo record logico (ottenuto col riferimento `v`) nel file `f`. Vediamo quindi come il punto centrale dell'accesso diretto è la possibilità per il programmatore di poter specificare un indice specifico a cui scrivere nel file, senza aver bisogno di scannerizzarlo in una direzione o l'altra.

- **Accesso a indice**

Con l'accesso a indice andiamo ad interporre fra l'accesso al file e il file stesso una struttura a *indice*, che permette l'accesso alle informazioni nel file sfruttando *chiavi*.

- `readk(f, key, &v)`, che permette la lettura del record logico indicizzato dalla chiave `key` (col riferimento `$v`) del file `f`;
- `writek(f, i, v)`, che permette la scrittura del record logico indicizzato dalla chiave `key` (ottenuto col riferimento `v`) nel file `f`.

In questo caso è chiaro che l'accesso al file avviene solo dopo un'operazione ricerca sull'indice, che dovrà essere memorizzato in un altro file, o comunque in una locazione accessibile al filesystem.

La modalità di accesso è indipendente dal tipo di dispositivo utilizzato, o dalle tecniche di allocazione dei blocchi in memoria secondaria. Chiaramente, tali soluzioni determinano le loro modalità di accesso, ma vorremo che il S/O faccia da livello di compatibilità per supportare qualsiasi metodo di accesso.

20.1.8 Organizzazione fisica del file system

Abbiamo già detto che ogni dispositivo di memorizzazione secondaria verrà partizionato in *blocchi*, che possiamo anche dire *record fisici*. In particolare:

- Un **blocco** è l'unità minima di trasferimento nelle operazioni di I/O da e verso il dispositivo *fisico*. La sua dimensione è per questo costante;
- Un **record fisico** è invece l'unità di trasferimenti minima nelle operazioni di accesso file, viste dai processi (e quindi dalle *applicazioni*).

La corrispondenza fra blocchi è record logici è che un singolo blocco può contenere più record logici. Per questo motivo si ha che la dimensione di un blocco è maggiore di quella di un record logico.

20.1.9 Allocazione contigua

Iniziamo quindi a vedere come possiamo mappare i record logici dei file nei vari record fisici. Il caso più semplice è quello dell'**allocazione contigua**, dove ogni file è mappato su insieme di blocchi fisicamente contigui.

I vantaggi di questo approccio sono:

- Velocità nella ricerca di un blocco: per trovare il blocco contenente l' i -esimo byte di un file allocato a partire dal blocco B basta prendere:

$$i_B = B + \frac{i}{N_{\text{byte}}}$$

dove N_{byte} è il numero di byte per blocco.

- La possibilità di fornire accesso sequenziale e diretto in maniera molto semplice, che va in qualche modo di pari passo con la caratteristica di accesso rapido appena nominata.

Gli svantaggi sono invece:

- La **frammentazione esterna**: man mano che il disco si riempie, rimangono zone contigue sempre più piccole e quindi inutilizzabili. A questo punto si rende necessario operare il *compattamento* del file system;
- Cercare spazio libero per un nuovo file ha un certo costo. Inoltre bisogna fare le dovute considerazioni riguardo agli approcci *first-fit* o *best-fit*;
- Se la dimensione del file cambia, si hanno dei seri problemi generati dalla riallocazione del file, dovesse questo andare ad impattare per allocazione continua regioni già occupate da altri file.

21 Lezione del 04-12-25

Continuiamo a trattare le tecniche di mappatura dei record fisici in record logici.

21.0.1 Allocazione a lista concatenata

In questo caso vogliamo organizzare i blocchi sui quali viene mappato ogni file secondo una struttura a **lista concatenata**.

I vantaggi di questo approccio sono:

- Non esiste frammentazione esterna, in quanto ogni record fisico libero può essere usato ed inserito come successivo a qualsiasi altro record logico;
- Il costo di allocazione è quindi minore;
- L'accesso sequenziale è a basso costo (le liste concatenate sono ottime proprio per gli accessi sequenziali).

Gli svantaggi sono invece:

- La facilità di introdurre errori nel caso di danneggiamenti di link fra nodi della lista;

- Lo spazio aggiuntivo occupato dai puntatori ai prossimi elementi di ogni elemento della lista;
- L'accesso diretto risulta oneroso: per accedere all' i -esimo byte abbiamo bisogno di:

$$t_B = \frac{i}{N_{\text{byte}}}$$

iterazioni, cioè dobbiamo scansionare i primi t_B blocchi della list;

- Dobbiamo inoltre notare il costo della ricerca di un blocco puntato, che non è più contiguo ai blocchi precedenti, ma potrebbe trovarsi in regioni arbitrarie della memoria.

Il sistema operativo Microsoft Windows implementa l'allocazione a lista concatenata attraverso le cosiddette **FAT** (*File Allocation Table*). Queste non sono altro che tabelle che associano ad ogni blocco fisico, il blocco fisico successivo.

21.0.2 Allocazione a lista doppiamente concatenata

Possiamo facilmente estendere l'approccio precedente rendendo la lista **doppiamente concatenata**.

I vantaggi sono immediati:

- Possiamo permettere la scansione in due direzioni di ogni file;
- Il sistema è reso più robusto dalla presenza di più link (si possono recuperare errori di perdita di un link).

Gli svantaggi sono invece dati principalmente dall'aumento dello spazio necessario per blocco, in quanto raddoppierà il numero di puntatori da mantenere (da 1 a 2).

21.0.3 Allocazione a indice

Nell'**allocazione a indice**, a ogni file viene associato un blocco (detto *indice*) in cui sono contenuti tutti gli indirizzi dei blocchi su cui è allocato il file.

I vantaggi sono gli stessi dell'allocazione a lista, con l'aggiunta di:

- Possibilità di fare accesso diretto senza scansioni (sfruttando l'indice);
- Maggiore velocità di accesso rispetto alle liste.

Lo svantaggio principale di questo approccio è la sua difficile **scalabilità**, data dalla crescita delle dimensioni dell'indice.

Assunta N_{byte} come la dimensione del blocco in byte, S come la dimensione del disco, abbiamo che ogni entrata dell'indice dovrà essere almeno:

$$S_r = \frac{S}{N_{\text{byte}}}$$

per cui il numero di blocchi che potremo indicizzare sarà:

$$N_{\text{blocchi}} = \frac{N_{\text{byte}}}{S_r}$$

e la dimensione massima del file indicizzabile:

$$S_{\max} = N_{\text{blocchi}} \times N_{\text{byte}}$$

Una soluzione può essere di concatenare più indici, dedicando ad ogni indice un nuovo blocco.

Vediamo che in UNIX si usa un approccio simile all'allocazione a indice.

21.1 Filesystem UNIX

Abbiamo introdotto il fatto che in UNIX un file è rappresentato da un descrittore detto **i-node**.

Questo è un descrittore che contiene i seguenti attributi:

- Il **tipo** del file, scelto fra *file ordinario*, *directory* o *file speciale*;
- Il **proprietario** del file (utente e gruppo, user-id e group-id);
- I 12 bit di **protezione** (i 9 bit dei permessi, SUID, SGID e STICKY bit);
- Le **date** di creazione e modifica del file;
- La **dimensione** del file;
- Il numero di **link** al file;
- Il cosiddetto **vettore di indirizzamento** (costituito da 13 a 15 indirizzi di blocchi). Questo consente l'indirizzamento dei blocchi di dati sui quali è allocato il file secondo una struttura ad indice.

Approfondiamo il vettore di indirizzamento. Assunto che la dimensione di un blocco è 512 byte, e gli indirizzi si trovano su 32 byte, si ha che ogni blocco può contenere 128 indirizzi di blocco.

Adottiamo quindi un'approccio a più livelli di indirezione sulla base della dimensione del file:

- I primi 10 blocchi di dati sono accessibili direttamente ($10 \times 512 \text{ byte} = 5 \text{ KiB}$);
- I prossimi 128 blocchi sono accessibili con *indirezione singola*, accedendo al puntatore 11 ($128 \times 512 \text{ byte} = 64 \text{ KiB}$);
- I prossimi 128×128 blocchi sono accessibili con *indirezione doppia*, accedendo al puntatore 12 ($128 \times 128 \times 512 \text{ byte} = 8 \text{ MiB}$);
- I prossimi $128 \times 128 \times 128$ blocchi sono accessibili con *indirezione tripla*, accedendo al puntatore 13 ($128 \times 128 \times 128 \times 512 \text{ byte} = 1 \text{ GiB}$);

Questo approccio si ripete così fino al 15-esimo indice. In questo modo si riesce a raggiungere dimensioni massime del file dell'ordine del GB, mantenendo però le strutture a indice piccole e con poca indirezione per file di piccole dimensioni.

21.1.1 Organizzazione logica del file system UNIX

La filosofia del file system UNIX è che *tutto è un file*, per cui un file può rappresentare un file effettivo in memoria, una directory, o un *file speciale*, che può rappresentare un dispositivo accessibile in lettura/scrittura, come un socket, come un costrutto di sistema (come ad esempio le pipe).

Abbiamo visto come il descrittore di file è l'i-node. Possiamo anticipare che gli i-node sono allocati in una tabella centralizzata detta i-list, dove ogni i-node è identificato da un identificatore univoco detto i-number.

Una volta stabilita l'identificazione attraverso i-number, si possono rappresentare le directory semplicemente come liste di i-number associate a nomi per ogni file.

21.1.2 Organizzazione fisica del filesystem UNIX

Abbiamo quindi visto che in UNIX si adotta un tipo di allocazione ibrida (allocazione ad indici a più livelli). I blocchi fisici hanno dimensione tra i 512-4096 byte. La superficie del disco virtuale è quindi divisa in quattro partizioni:

- Il **boot block**: contiene informazioni fondamentali per il bootstrap (appendice sul disco delle informazioni o programmi che devono essere eseguiti al momento del bootstrap, inizializzazione del sistema);
- Il **super block**: contiene informazioni sull'organizzazione del file system (in particolare i limiti delle quattro regioni che stiamo descrivendo, il puntatore alla lista dei blocchi liberi e il puntatore a una lista degli i-node liberi)
- L'**i-list**: tabella contenente tutti gli i-node di file, directory e dispositivi. Ogni file ha associato uno o più nomi simbolici, uno e un solo descrittore (già visto) detto i-node (raggiungibile a partire da un intero detto i-number, che è l'indice dell'elemento posto nell'array i-list);
- La partizione **data blocks**: contenente i blocchi utilizzati per allocare file.

21.1.3 Accesso ai file in UNIX

L'**accesso ai file** si fa con le solite primitive `read()` e `write()`, notando che si possono avere diversi tipi di accesso anche per la solita primitiva (`write()` in `truncate` o `append`, ecc...).

L'accesso è *sequenziale*, e non si ha *strutturazione* (i file sono solamente sequenze di record, cioè byte). La posizione corrente dell'accesso viene mantenuta da un puntatore detto *I/O pointer*.

Vediamo quindi come il programmatore accede al filesystem. A ogni processo è associata una tabella dei **file aperti**. Ogni elemento della tabella rappresenta un file (che ricordiamo può anche essere un dispositivo, cioè un file speciale). Il *file descriptor* (o **filde**) di tale file non è altro che l'indice del file nella tabella dei file aperti di un processo. Esistono 3 filde di default:

- **stdin** (indice 0): lo stream di ingresso di default;
- **stdout** (indice 1): lo stream di uscita di default;
- **stderr** (indice 2): lo stream di errore di default.

Essendo un'informazione utile quando il processo è caricato, la tabella dei file aperti del processo è allocata nella sua *user structure*.

Lato kernel, manteniamo 2 tabelle per l'accesso ai file:

- La tabella dei **file attivi**, che per ogni file aperto contiene una copia dell'i-node e il numero di riferimenti a tale file (per permettere la rimozione automatica quando si disattiva l'ultimo riferimento);
- La tabella dei **file aperti di sistema**, che ha un elemento per ogni operazione di apertura non ancora chiusa, contenente:
 - L'I/O pointer, cioè la posizione corrente all'interno del file;
 - Il puntatore all'i-node del file nella tabella dei *file attivi*.

La divisione in due tabelle è resa necessaria dal fatto che più processi possono aprire lo stesso file: l'informazione condivisa fra i processi (cioè l'inode) è contenuta nella prima tabella, mentre l'informazione specifica ad ogni processo (cioè l'I/O pointer) è contenuta nella seconda tabella.

21.1.4 Primitive di accesso ai file

Vediamo quindi le primitive di accesso ai file nel dettaglio:

- `int open(char nomefile[], int flag, int? mode);`

Questa primitiva si occupa di aprire un file, cioè crearne l'entrata nella tabelle dei file aperti e restituirne il filde.

- `nomefile` è il nome del file, relativo alla working directory corrente, o assoluto;
- `flag` esprime il metodo di accesso. Ad esempio, si può specificare `O_RDONLY` per l'accesso in sola lettura, o `O_WRONLY` per l'accesso in sola scrittura;
- `mode` è un parametro richiesto soltanto se l'operazione di apertura richiede la creazione di un file (con `O_CREAT`). In tal caso, specifica i bit di protezione.

Il valore restituito è il filde del file, oppure -1 in caso di errore;

- `int close(int fd);`

Questa primitiva è la duale alla `open()`, e si occupa di chiudere il file aperto indicizzato da un certo filde.

- `fd` è, appunto, il fide del file da chiudere.

Restituisce 0 se l'operazione va a buon fine, e -1 in caso di errori;

- `int read(int fd, char* buf, int n);`

Si occupa di leggere `n` byte nel buffer `buf` da un certo filde.

- `fd` è il filde del file da cui leggere;
- `buf` è l'area in cui trasferire i byte letti;
- `n` è il numero di byte da leggere.

In caso di successo, restituisce un intero positivo $\leq n$ che rappresenta il numero di byte letti. Altrimenti restituisce -1;

- `int write(int fd, char* buf, int n);`
Si occupa di scrivere n byte dal buffer `buf` su un certo file.

- `fd` è il file descriptor del file da cui leggere;
- `buf` è l'area da cui ottenere i byte da scrivere;
- `n` è il numero di byte da scrivere.

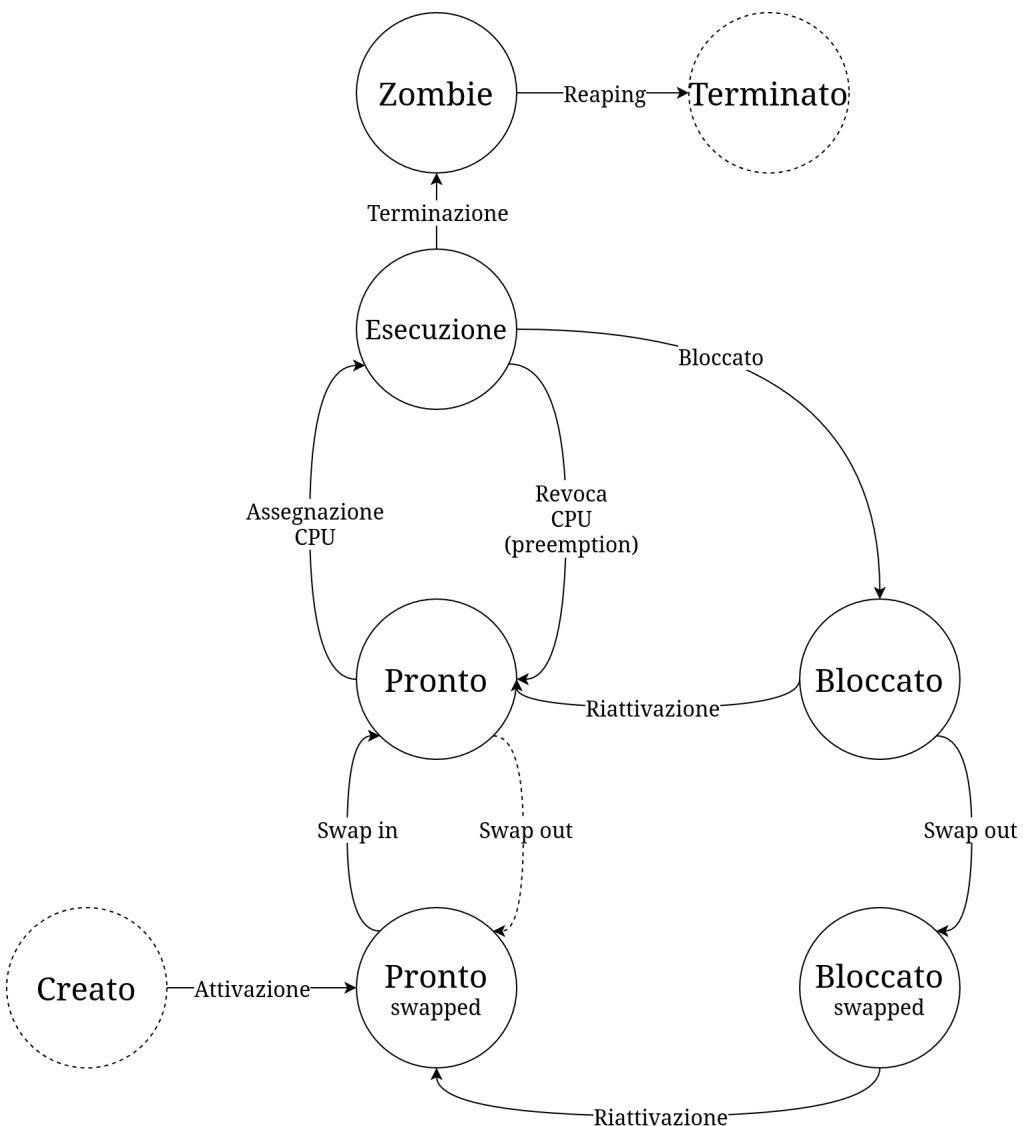
In caso di successo, restituisce un intero positivo $\leq n$ che rappresenta il numero di byte scritti. Altrimenti restituisce -1.

21.2 Processi in UNIX

Restando sull'argomento UNIX, vediamo nel dettaglio come questo sistema operativo gestisce i processi.

21.2.1 Stato dei processi UNIX

UNIX supporta lo *swapping* di processo, per cui il diagramma di stato del processo è simile a quello riportato in 17.3.2:



L'aggiunta a questo diagramma è quella dello stato **zombie**, in cui un processo va a trovarsi prima di essere effettivamente terminato. Un processo passa allo stato zombie quando viene effettivamente terminato, prima che il suo processo padre ne raccolga lo stato (il cosiddetto *reaping*).

21.2.2 Descrittori di processo UNIX

Avevamo definito in 5.0.2 come un processo è rappresentato da un descrittore detto **PCB** (*Process Control Block*). Questo descrittore è sostanzialmente quello adottato da UNIX.

In verità, tale struttura si divide in più sottostrutture sulla base delle aree di competenza dei dati da rappresentare. In particolare, vogliamo distinguere su 2 caratteristiche ortogonali:

- Dove i dati devono essere **accessibili**, cioè parte *kernel* e parte *utente*;
- Se i dati possono essere soggetti a **swapping**, cioè parte *swappable* e parte *non swappable* o *residente*.

Vediamo quindi come dividiamo il PCB:

- **Process structure**: contiene le informazioni necessarie al sistema per la gestione del processo (a prescindere dallo stato del processo).

In particolare, contiene:

- Il **PID** (*Process IDentifier*);
- Lo **stato** del processo;
- I puntatori alle aree **dati** e **stack**;
- Il riferimento alla **text structure**, su cui abbiamo più informazioni sotto;
- Le informazioni sullo **scheduling** da operare sul processo;
- Un riferimento al processo **padre** (cioè il suo PID),
- Informazioni relative alla gestione dei **segnali** UNIX (segnali inviati ma non ancora gestiti, maschere di segnale, ecc...);
- Puntatori a processi successivi nelle **code** di scheduling;
- Un puntatore alla **user structure** (la prossima che vediamo).

- **User structure**: contiene le informazioni necessarie solo se il processo è *residente* in memoria centrale (non si è fatto swap).

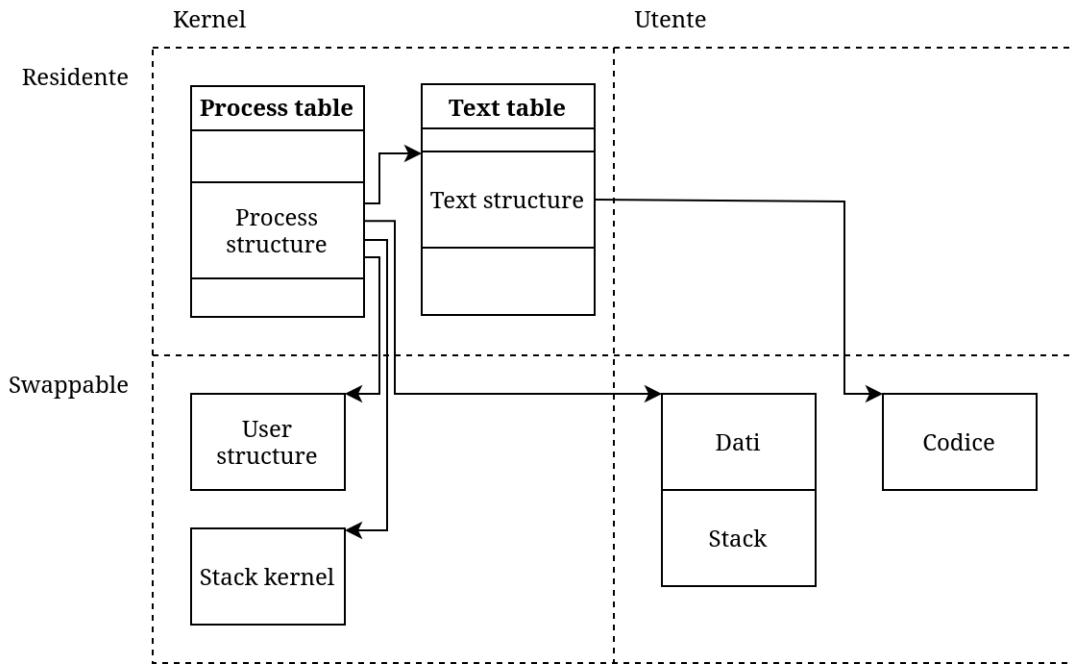
In particolare, contiene:

- La copia dei **registri** CPU;
- Come abbiamo già visto, le informazioni sulle risorse allocate, come ad esempio la tabella dei **file aperti**;
- Altre informazioni riguardanti la gestione dei **segnali** (handler segnali, ecc...);
- L'**ambiente** del processo, e quindi il direttorio corrente, gli argomenti, l'utente e il gruppo che lo hanno lanciato, le variabili di sistema (fra cui il PATH), ecc....

- **Text structure**: contiene informazioni riguardo a dove il codice di un processo si trova. Questa tabella viene implementata in quanto UNIX sfrutta il cosiddetto codice *rientrante*: più processi possono riferirsi allo stesso codice, e in tal caso il codice viene caricato una sola volta e riferito attraverso questa tabella.

Oltre a queste strutture, chiaramente, vorremo mantenere le aree *dati* e *stack* del processo, nonché ovviamente l'area *text* dove il codice puntato dalla text structure verrà effettivamente contenuto.

Vediamo quindi un diagramma che mostra come tutte le strutture riguardanti un processo appena visto sono disposte, sulla base delle 2 caratteristiche ortogonali viste prima:



21.3 Protezione e sicurezza

Veniamo quindi alla discussione della **protezione** e della **sicurezza** nei S/O.

- La **protezione** riguarda l'insieme di attività che si occupano di garantire all'interno di un sistema il controllo dell'accesso alle risorse logiche e fisiche;
- La **sicurezza** riguarda invece la prevenzione di comportamenti dannosi da parte di programmi o utenti all'interno del S/O.

21.3.1 Protezione

Il controllo degli accessi, prerogativa della **protezione**, è suddivisibile in 3 livelli concettuali:

- **Modelli:** rappresentano le astrazioni che il nostro sistema realizza.

In particolare, un *modello di protezione* definisce *soggetti*, e gli *oggetti* ai quali i soggetti hanno accesso e diritto di accesso, ovvero su cui possono svolgere operazioni. In questo, gli *oggetti* sono la parte passiva del sistema (risorse fisiche e logiche), mentre i *soggetti* sono la parte attiva (processi che agiscono per conto di utenti per accedere ad oggetti). Notiamo che un soggetto può avere diritti di accesso sia per gli oggetti che per altri soggetti (un soggetto può controllarne un altro).

Come abbiamo introdotto in 5.0.3, un soggetto in UNIX è rappresentato dalla tripla:

$$S = \langle \text{PID}, \text{UID}, \text{GID} \rangle$$

composta da **PID** (*Process IDentifier*, già visto), e **UID** e **GID** (*User IDentifier* e *Group IDentifier*), che rappresentano il proprietario del processo.

Un soggetto può ulteriormente essere considerato come una coppia (*processo, dominio*), dove il dominio è l'*ambiente* di protezione nel quale il processo sta eseguendo (l'insieme dei suoi diritti di accesso per ogni oggetto). Si possono avere domini *disgunti* o domini con diritti di accesso *comuni*.

Un dominio di protezione è unico per un soggetto, mentre un processo può cambiare dominio durante la sua esecuzione. In altre parole, il soggetto S_i può rappresentare il processo P in un certo dominio, e S_j può rappresentare lo stesso processo in un altro dominio. L'associazione fra processo e dominio può essere:

- *Statica*: cioè l'insieme delle risorse disponibili ad un processo rimane fisso durante il suo tempo di vita;
 - *Dinamica*: l'associazione fra processo e dominio varia durante l'esecuzione del processo. In questo caso chiaramente dobbiamo prevedere meccanismi per consentire il passaggio da un dominio all'altro.
- **Politiche**: sono gli insiemi di regole attraverso le quali i soggetti possono accedere agli oggetti.

Possono essere di più tipi:

- **DAC** (*Discretionary Access Control*): il creatore di un oggetto controlla i diritti di accesso per quell'oggetto. Ad esempio, è il tipo di politiche usate in UNIX;
- **MAC** (*Mandatory Access Control*): i diritti di accesso vengono gestiti centralmente. Viene usato in sistemi di alta sicurezza (enti governativi, difesa, ospedali, ecc...);
- **RBAC** (*Role-Based Access Control*): ad un ruolo sono assegnati specifici diritti di accesso sulle risorse. Gli utenti possono appartenere a diversi ruoli.

Qualunque sia tipo di politica, si nota un concetto base, cioè quello del *privilegio minimo*: ad ogni soggetto dovrebbero essere garantiti solo i diritti di accesso strettamente necessari alla sua esecuzione;

- **Meccanismi**: sono gli strumenti messi a disposizione dal sistema di protezione per imporre una determinata politica.

La separazione fra politiche e meccanismi è la seguente:

- La politica definisce *cosa* va fatto;
- Il meccanismo definisce *come* viene fatto.

Ci interessa assicurare la *flessibilità* del sistema di protezione: i meccanismi di protezione devono essere sufficientemente generali per consentire l'applicazione di diverse politiche di protezione.

21.3.2 Implementazione in UNIX

Abbiamo già detto che in UNIX il *soggetto* è rappresentato dalla solita tripla $S = \langle \text{PID}, \text{UID}, \text{GID} \rangle$. Di questo UID e GID definiscono il *dominio* di protezione.

L'associazione soggetto/dominio è dinamica, in quanto è previsto il cambio da *user mode* a *system mode*. Inoltre, il programmatore ha a disposizione le primitive di tipo `exec()` su processi con bit SUID o SGID attivati (che quindi vengono elevati ai diritti di accesso del loro proprietario o gruppo proprietario).