

Sistemi Operativi

AAF - Secondo anno - 3CFU

A.A. 2023/2024

Corso di Laurea in Matematica

La Gestione della Memoria - Parte 1

Annalisa Massini

Dipartimento di Informatica
Sapienza Università di Roma

Argomenti trattati

1 Gestione della memoria

- Requisiti di base
- Partizionamento della memoria

Perché gestire la memoria (nel SO)

- Le moderne applicazioni richiedono sempre maggiore **memoria**, che oggi troviamo a costo sempre più basso
- La **gestione della memoria** deve garantire che ci sia sempre un **numero ragionevole di processi pronti all'esecuzione**, così da non lasciare inoperoso il processore

Perché gestire la memoria (nel SO)

- Gestire la memoria include lo swap di blocchi di dati da memoria principale a memoria secondaria
- Lo scambio tra memoria principale e memoria secondaria è lento (più lento del processore), essendo la memoria secondaria vista come un dispositivo di I/O
- Il SO deve pianificare lo swap in modo intelligente, così da massimizzare l'efficienza del processore

Requisiti per la gestione della memoria

Gli aspetti principali nella gestione della memoria sono:

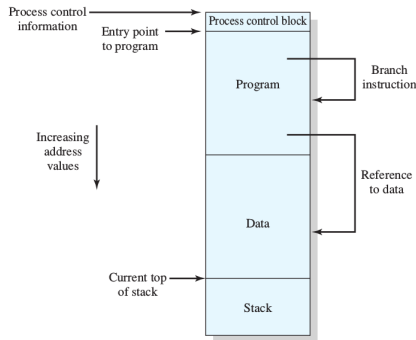
- **Rilocazione**
 - serve aiuto hardware
- **Protezione**
 - serve aiuto hardware
- **Condivisione**
- **Organizzazione logica**
- **Organizzazione fisica**

Requisiti: Rilocalizzazione

- Il *programmatore* non sa (e non ha bisogno di sapere) in quale zona della memoria verrà caricato il programma
 - A un certo punto il programma potrebbe essere **swappato su disco** e al ritorno in memoria principale potrebbe essere in un'**altra posizione**
 - Il programma potrebbe essere in **porzioni di memoria non contigue**, oppure con alcune parti in RAM e altre su disco
 - NB: in questo contesto, per *programmatore* si intende chi usa l'assembler o il compilatore
- I riferimenti alla memoria devono corrispondere ad un indirizzo fisico *vero*, che viene ottenuto quando il programma viene portato in memoria principale
 - preprocessing
 - oppure a run-time con supporto hardware

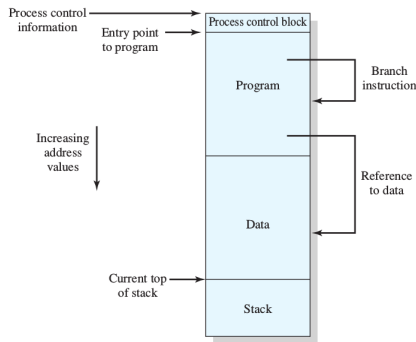
Rilocazione: gli indirizzi nei programmi

- In figura l'immagine di un processo
- Assumiamo occupi una porzione di memoria contigua
- Il sistema operativo deve conoscere:
 - la locazione delle informazioni di controllo del processo
 - la locazione dell'istruzione di inizio del programma per avviare il processo
 - la locazione della stack



Rilocazione: gli indirizzi nei programmi

- Il sistema operativo ha questi indirizzi quando trasferisce il programma in memoria
- Il processore usa gli indirizzi all'interno del programma:
 - per le istruzioni di salto (branch)
 - per reperire i dati presenti nelle istruzioni (tramite indirizzi)
- Quindi, processore (HW) e SO (SW) collaborano per tradurre i riferimenti alla memoria in indirizzi fisici della memoria principale



Rilocazione: gli indirizzi nei programmi

- Ogni volta che un processo viene riportato in memoria, potrebbe essere in un porzione diversa di memoria
- Nel frattempo, potrebbero essere arrivati altri processi e prenderne il posto
- Quindi, ad ogni ricaricamento in RAM, occorre individuare gli indirizzi presenti nel codice sorgente del processo e determinare i valori effettivi
- Essendo troppo oneroso per il SO, si ha il supporto con soluzioni hardware

Indirizzi

Distinguiamo **indirizzi**

Logici: il riferimento in memoria è indipendente dall'attuale posizionamento del programma in memoria

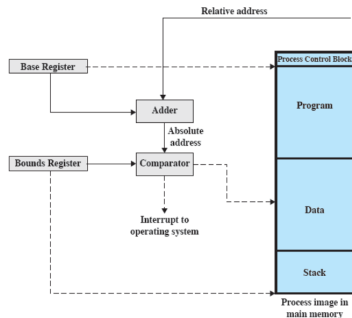
Relativi: il riferimento è espresso come uno spiazzamento rispetto ad un qualche punto noto

- caso particolare degli indirizzi logici

Fisici o Assoluti: il riferimento effettivo alla memoria

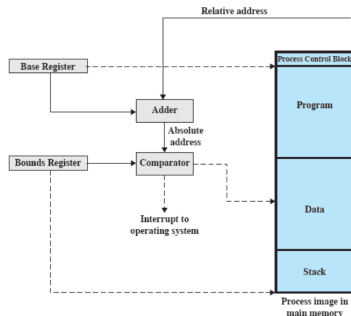
Rilocazione: gli indirizzi nei programmi

- Si usano:
 - Base register (registro base)
 - indirizzo di partenza del processo
 - Bounds register (registro limite)
 - indirizzo di fine del processo
- Vengono settati quando il processo viene posizionato in memoria
 - mantenuti nel PCB del processo
 - vanno calcolati, non semplicemente ripristinati



Rilocazione: gli Indirizzi nei Programmi

- Il valore del **registro base** viene aggiunto al valore dell'**indirizzo relativo** per ottenere l'indirizzo assoluto
- Il risultato è confrontato con il registro limite
- Se va oltre, viene generato un interrupt per il sistema operativo



Requisiti: Protezione

- I processi non devono poter accedere a locazioni di memoria di un altro processo, a meno che non siano autorizzati
- A causa della rilocalizzazione, non si può controllare a tempo di compilazione
- Quindi bisogna farlo a tempo di esecuzione
- E quindi serve supporto hardware

Requisiti: Condivisione

- Deve essere possibile permettere a più processi di accedere alla stessa zona di memoria
 - ovviamente, solo se è effettivamente utile allo scopo perseguito dai processi
- Caso tipico: più processi vengono creati eseguendo più volte lo stesso sorgente
 - finché questi processi restano in esecuzione, è più efficiente che condividano il codice sorgente, visto che è lo stesso
- Ci sono anche casi in cui processi diversi vengono programmati esplicitamente per accedere a sezioni di memoria comuni
 - usando chiamate di sistema

Requisiti: Organizzazione Logica

- A livello hardware, la memoria è organizzata in modo **lineare**
 - sia RAM che disco
- A livello software, i programmi sono scritti in moduli
 - i moduli possono essere scritti e compilati separatamente
 - a ciascun modulo possono essere dati diversi permessi (sola lettura, sola esecuzione)
 - i moduli possono essere condivisi tra i processi
- Per facilitare la realizzazione dei punti precedenti, il SO usa la tecnica di gestione di memoria basata sulla **segmentazione**

Requisiti: Organizzazione Fisica

- La gestione del flusso tra memoria principale (piccola, veloce e volatile) e memoria secondaria (grande, lenta e permanente) NON può essere lasciata al programmatore:
 - il programmatore non sa quanta memoria avrà a disposizione
 - la memoria potrebbe non essere sufficiente a contenere il programma ed i suoi dati
 - la tecnica dell'*overlaying* (sovrapposizione) permette a più moduli di essere posizionati nella stessa zona di memoria (in tempi diversi...), ma è difficile da programmare
- Quindi, serve il supporto del SO

Considerazioni

- La principale operazione nella gestione della memoria è **portare i programmi in memoria principale** affinché possano essere eseguiti dal processore
- Nei moderni sistemi multiprogrammati, ciò si ottiene grazie al sofisticato meccanismo di **memoria virtuale**
- La memoria virtuale si basa sulle due tecniche (una o entrambe) di **paginazione** e **segmentazione**
- Prima di esse vediamo una tecnica più semplice (nelle sue diverse varianti) che facilita la comprensione della memoria virtuale: il **partizionamento**

Partizionamento

- Il **partizionamento** è stato uno dei primi metodi per la gestione della memoria
- Non è più molto usato
- I due tipi di partizionamento sono:
 - **Partizionamento fisso**
 - Uniforme
 - Variabile
 - **Partizionamento dinamico**

Partizionamento fisso uniforme

- La memoria è suddivisa in partizioni di ugual lunghezza
 - se un processo ha una dimensione minore o uguale della misura di una partizione, allora può essere caricato in una partizione libera
- Il sistema operativo può togliere un processo da una partizione a caricarne un altro (swap)
 - ad esempio, se tutte le partizioni sono occupate e nessuno dei processi attualmente in memoria è in stato *ready*



Partizionamento fisso uniforme: problemi

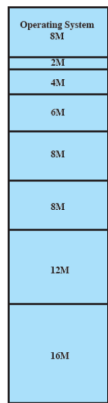
I problemi del partizionamento fisso uniforme sono:

- Un programma potrebbe non entrare in una partizione
 - sta(va) al programmatore dividere il suo programma e usare l'*overlay*
- Uso inefficiente della memoria
 - ogni programma, anche il più piccolo, occupa un'intera partizione
 - problema della *frammentazione interna*



Partizionamento fisso variabile

- La memoria è suddivisa in partizioni di dimensioni diverse
- Questa soluzione mitiga entrambi i problemi
 - ma non li risolve completamente
- Per i programmi più piccoli, ci sono le partizioni più piccole
- Sfruttando le diverse dimensioni delle partizioni si può ridurre l'uso dell'**overlay**
- È sempre partizionamento fisso: quindi le partizioni sono decise all'inizio e non cambiano nel tempo



Algoritmo di posizionamento

- Partizioni di uguale lunghezza
 - se ci sono partizioni libere, ogni processo può andare in qualunque partizione
 - l'algoritmo banale è andare in ordine
 - se non ci sono partizioni libere, serve lo *swap* tra processi e la decisione riguarda lo *scheduling*
- Partizioni di diversa lunghezza
 - un processo va nella partizione più piccola che può contenerlo
 - questo minimizza la quantità di spazio sprecato
 - gestione a coda:
 - una coda per ogni partizione, ma potrebbero rimanere inutilizzate le partizioni più grandi
 - oppure una coda unica per tutte le partizioni

Partizionamento fisso: problemi irrisolti

L'introduzione del partizionamento fisso variabile non risolve tutti i problemi:

- C'è un numero massimo di processi in memoria principale
 - corrispondente al numero di partizioni deciso inizialmente
- Se ci sono molti processi piccoli, la memoria verrà usata in modo inefficiente
 - sia con le partizioni di lunghezza uguale che con quelle variabili

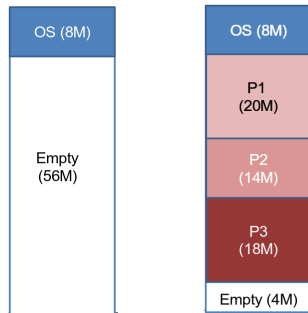
Partizionamento dinamico

- Alcuni problemi del partizionamento fisso vengono superati con il **partizionamento dinamico**
- Si tratta comunque di una tecnica soppiantata da tecniche più sofisticate
- Con il **partizionamento dinamico**:
 - Le partizioni variano sia in misura che in quantità
 - Per ciascun processo viene allocata esattamente la quantità di memoria che serve

Partizionamento dinamico

Esempio

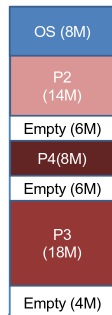
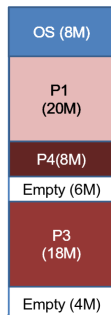
- All'inizio la memoria principale è vuota, eccetto per il SO
- Vengono poi caricati - a partire da dove finisce il SO - i primi tre processi:
 - P1 di 20M, P2 di 14M, P3 di 18M
- Resta libera una piccola porzione di memoria (4M)
- Arriva il processo P4 di 8M, ma lo spazio rimasto non è sufficiente



Partizionamento dinamico

Esempio

- Ad un certo punto nessuno dei processi in memoria è *ready*
- Il SO esegue uno swap portando P2 in memoria secondaria, guadagnando spazio per P4
- Si crea così un altro buco di 6M
- Se poi si riporta P2 in memoria principale facendo *swap* con P1 si ha un altro buco da 6M



Partizionamento dinamico

- Andando avanti la memoria può presentare sempre più buchi e l'utilizzazione della memoria diventa sempre meno efficiente
- Si ha il fenomeno di **frammentazione esterna**: la memoria che non è usata per nessun processo viene frammentata
- Si può risolvere con la **compattazione**
 - il SO sposta i processi in modo che siano contigui
 - tecnica con un elevato overhead

Partizionamento dinamico

- Al problema della frammentazione si può anche ovviare usando algoritmi di posizionamento o rimpiazzamento più sofisticati
- Se ci sono più blocchi liberi, il SO deve decidere a quale blocco libero assegnare un processo
- Si usano essenzialmente tre algoritmi di posizionamento:
 - **best-fit**
 - **first-fit**
 - **next-fit**

Partizionamento dinamico

- Algoritmo **best-fit** (miglior blocco tra quelli adatti)
 - sceglie il blocco la cui misura è la più vicina (in eccesso) a quella del processo da posizionare
 - nonostante il nome, è quello con risultati peggiori
 - lascia frammenti molto piccoli
 - costringe a fare spesso la compattazione

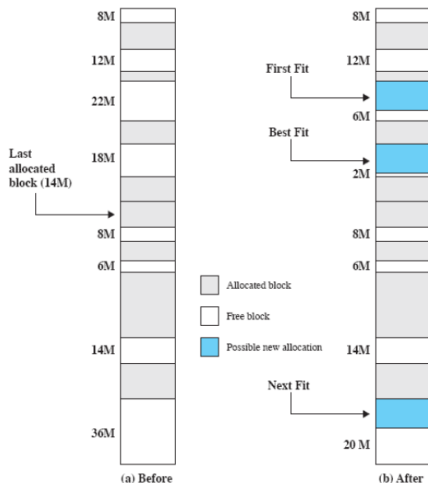
Partizionamento dinamico

- Algoritmo **first-fit** (il primo blocco tra quelli adatti)
 - si scorre la memoria dall'inizio
 - si sceglie il primo blocco di memoria abbastanza grande
 - molto veloce
 - conti fatti, è(ra) il migliore
 - tende a riempire solo la prima parte della memoria

Partizionamento dinamico: Esempi di allocazione

La memoria **prima e dopo** l'allocazione di un blocco da 16M con i tre algoritmi

- **best-fit**
- **first-fit**
- **next-fit**



Partizionamento dinamico

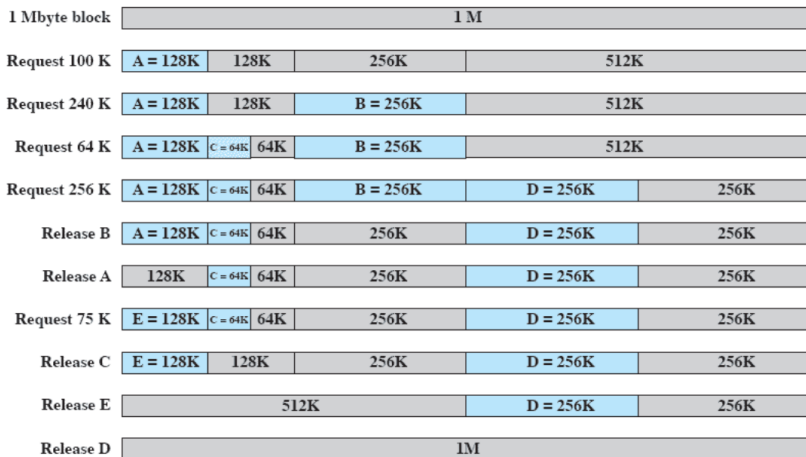
Esercizio

- Al tempo t i seguenti blocchi di memoria **occupati** e liberi sono intercalati nel seguente ordine: **4M** - 8M - **8M** - 12M - **20M** - 6M - **16M** - 8M - **10M** - 16M - **2M** - 22M
- L'ultimo blocco allocato è quello che occupa 10M
- Si ricevono le seguenti richieste di allocazione di blocchi di memoria:
 - Al tempo $t+1$ richiesta per 10M
 - Al tempo $t+2$ richiesta per 6M
 - Al tempo $t+3$ richiesta per 20M
- Mostrare l'allocazione di memoria dopo le tre richieste, con i tre diversi algoritmi **best-fit**, **first-fit**, **next-fit**

Buddy System (Sistema del Compagno)

- Compromesso tra partizionamento fisso e dinamico
- Siano:
 - 2^U la dimensione del blocco più grande di memoria (all'inizio tutta la memoria disponibile) della memoria
 - 2^L la dimensione del blocco più piccolo di memoria
 - s la dimensione del processo da mettere in RAM
- Si dimezza lo spazio fino a quando si trova un X t.c.
 $2^{X-1} < s \leq 2^X$, con $L \leq X \leq U$
 - una delle 2 porzioni viene usata per il processo
 - L serve per dare un lower bound e non creare partizioni troppo piccole
- Occorre tenere traccia delle porzioni già occupate
- Quando un processo finisce, se il *buddy* è libero si può fare una fusione

Esempio di Buddy System



Esempio di Buddy System: Rappresentazione ad Albero

