

La gestione della memoria

Abbiamo già visto come le memorie siano una componente fondamentale di ogni calcolatore digitale. Una analisi più approfondita è tesa a caratterizzare le varie tipologie di memoria, in base ai presupposti fisici e tecnologici e all'utilizzo di ciascuna di esse.

Dal punto di vista dei supporti fisici e della tecnologia con cui sono realizzate, le memorie si dividono usualmente in tre categorie: elettroniche (con tecnologia VLSI), magnetiche ed ottiche. Questa classificazione fornisce la chiave di lettura fondamentale per comprendere le qualità dei vari tipi di memoria. I parametri che in tal senso vengono considerati sono: modo di accesso, tempo di accesso, volatilità e costo per singolo bit immagazzinato.

Il modo di accesso è determinato dalla tecnica utilizzata per accedere al dato. Da questo punto di vista si parla di memorie ad accesso casuale, tipicamente le memorie in cui l'accesso al dato è "diretto" e dunque il tempo impiegato è lo stesso indipendentemente dalla sua posizione in memoria (ad esempio i CD audio), e memorie ad accesso seriale, in cui il tempo di accesso al dato dipende dalla sua posizione in memoria, in quanto occorre attraversare tutte le locazioni che precedono la locazione interessata (ad esempio le musicassette audio, caso tipico di nastro magnetico).

Il tempo di accesso rappresenta il tempo che intercorre tra la richiesta di un dato e la sua effettiva disponibilità al richiedente, che nel nostro caso è la CPU. Questo "ritardo" (che può dipendere anche dalla modalità di accesso) caratterizza la velocità di una memoria. I tipi di memoria con performances migliori si dice che sono ad un livello "più alto" di quelle con performances peggiori. Per volatilità si intende la possibilità o meno di conservare i dati memorizzati indipendentemente dalla fruibilità di corrente elettrica, in altre parole una memoria è volatile se le informazioni sono perse dalla memoria dopo lo spegnimento del calcolatore.

Anche il costo (per singolo bit di memoria) rappresenta un parametro critico per la caratterizzazione delle memorie e per la "composizione" effettiva di un elaboratore. In fondo la filosofia è sempre la stessa: ottenere il massimo risultato con il minimo sforzo.

I sistemi di memoria possono essere suddivisi, in modo del tutto generale, in base alla loro funzione e alla loro dislocazione nel calcolatore: memoria centrale, ulteriormente suddivisibile in interna o esterna al processore, e memoria di massa o secondaria.

La memoria centrale che genericamente comprende i registri interni alla CPU, la memoria cache, la memoria ROM e la memoria RAM viene spesso identificata con la sola memoria RAM, che di fatto è la memoria principale del calcolatore. Queste memorie, cui la CPU accede direttamente, sono veloci e costose e servono a memorizzare temporaneamente i dati e le istruzioni da elaborare. Le dimensioni vanno dalle decine di bytes fino alle decine di Mbytes.

La memoria di massa comprende supporti generalmente interni al calcolatore, come l'hard disk, e supporti generalmente esterni al calcolatore e quindi trasportabili, come i floppy disk, i Compact Disk e i Digital Versatile Disk (DVD). Tutte queste memorie, non di immediato interesse della CPU, sono lente e poco costose e servono a memorizzare indefinitamente le informazioni. Le dimensioni vanno da pochi Megabytes fino a decine di Gigabytes.

La memoria virtuale

Il sistema di memorie viene progettato e gestito in modo da ottenere:

- capacità di memorizzazione adeguata, in quanto l'elaborazione delle informazioni sui moderni sistemi di calcolo richiede grandi quantità di memoria;
- prestazioni accettabili, per non ritardare le esecuzioni dei calcoli, già in quantità considerevoli;
- costi ridotti.

Per questi motivi viene utilizzata una gerarchia che cerca di ottimizzare costi (minimi), capacità (massime) e tempi di accesso (minimi).

In altre parole se le memorie M_1 e M_2 hanno caratteristiche differenti (denotiamo con c_i il costo per singolo bit, t_i il tempo di accesso alle locazioni, S_i le dimensioni fisiche), ad esempio se la memoria M_1 ha un costo superiore per bit ($c_1 > c_2$) ed un tempo di accesso inferiore ($t_1 < t_2$), allora le dimensioni di M_1 saranno inferiori alle dimensioni di M_2 ($S_1 < S_2$).

Un esempio concreto di gerarchia di memoria è il sistema di memoria di uno studente che prevede, nell'ordine: la propria memoria, la borsa di scuola, lo scaffale di casa, la libreria o biblioteca, i depositi della casa editrice.

Nei personal computer moderni la gerarchia prevede il seguente ordine: Registri della CPU, memoria Cache, memoria RAM, dischi magnetici, dischi ottici (vedere fig. a lato)

Inoltre la gestione deve essere automatica e trasparente all'utente.

I dati utilizzati più raramente sono posti in memorie con tempi di accesso più elevato; i dati utilizzati più frequentemente in memorie con prestazioni più elevate.

L'obiettivo nella gestione della gerarchia è di mimare il sistema completo delle memorie con: i tempi di accesso della più veloce, le dimensioni della maggiore, i costi della più economica.



Il meccanismo di gestione della gerarchia di memoria si dice **Memoria Virtuale**. Vediamo il suo funzionamento con un esempio.

Supponiamo di avere le memorie M_1 e M_2 con le seguenti caratteristiche: la memoria M_1 ha un costo superiore per bit ($c_1 > c_2$), un tempo di accesso inferiore ($t_1 < t_2$), e capacità di memorizzazione o dimensioni fisiche inferiori ($S_1 < S_2$). Vogliamo far dialogare la CPU con la memoria M_1 , che è più veloce, ma su questa abbiamo il problema dello spazio che è inferiore, e molto spesso è addirittura insufficiente a contenere un solo processo in esecuzione. Allora si alloca il processo da eseguire nella memoria M_2 e lo si divide in lotti di istruzioni di grandezza tale da poter essere contenuti nella memoria M_1 . In questo modo la CPU preleva le informazioni dalla memoria M_1 , ciò finché trova tutte le informazioni di cui necessita. Quando è richiesto un dato non appartenente al lotto caricato nella memoria M_1 (page fault), si esegue il caricamento nella memoria M_1 di un nuovo lotto, ovvero quello contenente il dato interessato, al posto del precedente (page swaps) o spostando altri lotti (di altri processi) non immediatamente necessari. Così il meccanismo prosegue dettando i nuovi eventuali page swaps per caricare nella memoria più veloce le varie parti del processo in esecuzione. Anche se il programma non entra nella memoria M_1 è come se vi risiedesse, qualità che fa sembrare all'utente di avere a disposizione più memoria ad un certo livello usando memoria di livello inferiore.

E' da osservare che questo meccanismo, oltre a sfruttare efficacemente la gerarchia, rende il programma in esecuzione indipendente dalle dimensioni della memoria della macchina. Inoltre, poiché la gestione della memoria è implementata in maniera automatica, il programmatore viene liberato dalla gestione dello spazio di memoria fisico.

La validità del meccanismo della memoria virtuale risiede nel fatto che l'unità di controllo non deve continuamente effettuare page swaps. Se così non fosse il tempo risparmiato nell'accesso ai dati non compenserebbe il ritardo causato dai continui cambiamenti di pagina. La convenienza di questo meccanismo, dunque, si basa su un principio empiricamente vero, detto **principio di località**:

“se al tempo t si accede all'indirizzo X , è molto probabile che nell'intervallo di tempo fra t e $t + \Delta t$ si richieda l'indirizzo $X + \Delta X$ (dove Δt = breve lasso temporale, ΔX = piccola distanza spaziale)”.

In altre parole, in brevi intervalli di tempo, gli indirizzi di memoria utilizzati da un programma sono confinati in regioni limitate, cioè tendono a “concentrarsi” attorno a pochi indirizzi (tranne casi di salto).

La memoria centrale

Con questa definizione si usano distinguere quelle memorie utilizzate direttamente dalla CPU per le sue elaborazioni.

Memoria ROM

La memoria di tipo **ROM** (*Read Only Memory*), di sola lettura, è adatta per i programmi di inizializzazione del computer, infatti in essa i dati sono memorizzati in modo permanente.

Quando il calcolatore viene acceso, la RAM è vuota. Tuttavia la CPU deve poter trovare da qualche parte una sequenza di istruzioni, ovvero un programma iniziale, da eseguire. Le routine di base riguardano sostanzialmente la diagnostica in fase di avvio del computer, il caricamento del sistema operativo da disco, un'interfaccia di basso livello con le periferiche.

La ROM assolve a questa funzione attraverso il BIOS (Basic Input/Output System) cioè una serie di programmi utili per l'accesso ai dati che consente al computer di disporre, all'accensione, degli strumenti di base per poter funzionare. Il BIOS contiene un piccolo programma che consente di trasferire dalla memoria di massa a quella centrale il sistema operativo e altri dati essenziali (fase di caricamento o di boot).

Registri della CPU

Sono memorie elettroniche interne alla CPU, della grandezza di poche decine di bytes, e servono a memorizzare temporaneamente, per brevi lassi di tempo, le istruzioni da eseguire, i dati da processare, le risultanze delle elaborazioni effettuate. Sono memorie di tipo flip-flop e dunque molto veloci (l'accesso al dato è dell'ordine del miliardesimo di secondo). Sono memorie di tipo volatile.

Memoria RAM

In un personal computer la memoria principale viene chiamata RAM (Random Access Memory), maggiore è la sua quantità e migliori saranno le prestazioni del computer dove essa opera.

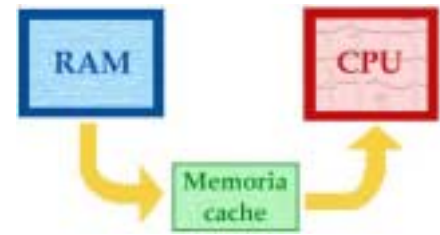
La memoria RAM, anch'essa di tipo elettronico si può così riassumere:

- è formata da una matrice di celle identificate da un indirizzo (*address*) il cui valore va da 0 ad n (in notazione binaria);
- il contenuto delle celle si chiama *word* (parola), l'unità logica elaborabile; le parole possono essere di 8, 16, 32, 64 bit (dipendente dalla lunghezza del Registro Indirizzi della CPU)
- se il numero di bit utilizzati per descrivere l'indirizzo è k , lo *spazio di indirizzamento* (ovvero il numero di locazioni di memoria) è uguale a 2^k . Ad esempio, il processore Pentium ha indirizzi lunghi 32+16 bit e può indirizzare 2^{48} celle, che vengono gestiti come $2^{16} = 65.536$ segmenti di 4 Gbyte (4 Gbyte = 2^{32} bytes).
- la velocità di accesso è dell'ordine dei 10-20 nanosecondi
- l'accesso alle locazioni è uniforme (RAM = Random Access Memory): memoria ad accesso diretto, il tempo di ritrovamento del dato non dipende dalla sua posizione
- è una memoria di tipo volatile, perde il suo contenuto allo spegnimento del computer
- l'operazione di lettura non è distruttiva, mentre l'operazione di scrittura è distruttiva, nel senso che viene perso il contenuto precedente.

Memoria Cache

Alcuni computer sono dotati di questo speciale modulo di memoria "ausiliaria" detta cache (deposito). Le *Cache memory* sono memorie elettroniche veloci basate su flip-flop e dunque volatili. Esse rappresentano aree di memoria del computer in cui sono memorizzati temporaneamente dei dati precedentemente utilizzati, interposte tra una memoria a bassa velocità ed un sistema ad alta velocità (tipicamente una CPU).

Lo scopo della cache memory è di “mascherare” la lentezza della memoria vera e propria, andando a compensare la differenza di velocità tra le elaborazioni della CPU e la memoria RAM. Quando un’applicazione richiede informazioni, controlla prima l’area cache per vedere se quei dati sono già disponibili; in caso positivo li preleva velocizzando il caricamento ed evitando al processore di perdere molti cicli di clock.



Il suo basso tempo di accesso la rende più costosa e viene quindi utilizzata con parsimonia per immagazzinare le variabili utilizzate più spesso e quelle adoperate più di recente (LRU = Least recently used). Quando le variabili sono molte ed occupano una gran quantità di memoria tale da sfiorare la capacità della cache si utilizza il sistema della memoria virtuale. Quando si verifica una cache miss (dato richiesto non presente nella cache), tutta la cache line (blocco di dimensione fissata che contiene il dato richiesto) viene trasferita dalla memoria centrale alla cache. Per il principio di località alcune delle parole caricate, ma non richieste, saranno presto utilizzate. A volte è integrata nella stessa CPU, si parla allora di “cache di primo livello”, che può eventualmente essere affiancata da una “cache di secondo livello” (esterna alla CPU). La cache di primo livello è accessibile in modo quasi istantaneo e risulta quindi la più veloce, la cache di secondo livello risulta 4 o 5 volte più lenta ma la differenza di velocità con la RAM è sempre altissima: la RAM, infatti, è più lenta di 20 o 30 volte. Generalmente si usano memorie cache di alcuni KB (poche centinaia di KB), ma maggiore è la dimensione della memoria cache, migliori saranno le prestazioni della CPU.

La memoria secondaria: i supporti magnetici

le altre memorie sono usate per memorizzare programmi e dati in modo permanente.

Una memoria secondaria ha le seguenti caratteristiche fondamentali, che la differenziano dalla memoria centrale: è di tipo permanente (non volatile), ha grandi capacità di memorizzazione (anche di diversi ordini di grandezza), ha bassi costi (il costo per bit è minore di diversi ordini di grandezza), ha minore velocità di accesso (i tempi sono maggiori di diversi ordini di grandezza). In pratica la memoria secondaria memorizza tutti i programmi e i dati del calcolatore, mentre la memoria centrale memorizza solo i programmi in esecuzione e i dati ad essi necessari. In ogni caso i dati di una memoria secondaria, per essere elaborati dal processore, debbono transitare nella memoria centrale.

Il primo tipo di memoria di massa sono state le schede perforate: erano cartoncini che venivano forati e rappresentavano un certo numero di byte ma il loro uso è stato lentamente annullato perché era disagevole e poco affidabile a causa dell’ingombro e della facilità con cui potevano danneggiarsi.

Il primo tipo di memoria di massa magnetica sono stati i nastri magnetici, caratterizzati da accesso sequenziale. Questo tipo di memoria è stato rimpiazzato da memoria a lettura random per i quali oggi si utilizzano supporti magnetici di forma diversa, tipicamente dischi.

Sommando il movimento di rotazione del disco e il movimento radiale della testina (addetta alla lettura/scrittura) è possibile raggiungere rapidamente qualunque punto del disco. In pratica il tempo massimo per raggiungere una certa posizione è pari al tempo richiesto per una rotazione o perché la testina attraversi completamente il cerchio (il maggiore dei due).

Le unità a dischi magnetici costituiscono oggi la più diffusa unità di memoria periferica. Sono principalmente di due tipi: fissi o hard disk e portatili o floppy disk (ma esistono anche altri formati proprietari, come lo Zip).

Hard disk

Un disco fisso è composto da uno o più piatti di alluminio rotanti (sullo stesso asse di rotazione), ricoperti di materiale magnetico (la memorizzazione avviene sotto forma di stati di polarizzazione) e da testine di lettura e scrittura.

La suddivisione delle locazioni all'interno di un disco magnetico si può così sintetizzare:

- ciascun piatto è composto da due *superfici* (le facce del disco) e ciascuna faccia è suddivisa in tracce (circolari) e *settori* (a spicchio); l'insieme delle tracce in una data posizione radiale è detto *cilindro*, mentre gruppi contigui di settori vengono chiamati *cluster*;
- i settori sono le unità logiche di memorizzazione i cui indirizzi sono dati dalle seguenti informazioni: cilindro, superficie, settore;
- la formattazione del disco consiste nella creazione di questi riferimenti, che dipendono da sistema a sistema.

Il disco rigido è una parte importante dei PC in quanto conserva il sistema operativo, le applicazioni software, i dati dell'Utente.

La memorizzazione dei file sul disco avviene tenendo nota della sua posizione attraverso una tabella di allocazione dei files, ad esempio sotto Windows è la FAT (File Allocation Table), ovvero una specie di mappa del disco che contiene i riferimenti ai cluster in cui iniziano i files. E' da osservare che quando si esegue la formattazione di un disco non nuovo si cancellano i riferimenti contenuti in questa tabella, in modo da rendere indisponibili i dati.

Una porzione dell'Hard Disk, inoltre, viene utilizzata dai sistemi operativi per realizzare la memoria virtuale (come memoria di appoggio della RAM).

Oggidi esistono due differenti tecnologie di gestione degli hard disk: la EIDE (Enhanced Integrated Drive Electronics) e la SCSI (Small Computer System Interface).

Floppy disk

Il Floppy disk (disco flessibile) è un dispositivo di memoria magnetico trasportabile. E' formato da un disco magnetico con il formato standard di 3,5", ad alta densità e a doppia faccia. E' formattato con 80 tracce concentriche e 18 settori (da 512 byte) per traccia con una capacità totale di 1.44 MB (da ciò ne consegue che sulle tracce interne le informazioni sono più concentrate rispetto alle tracce più esterne).

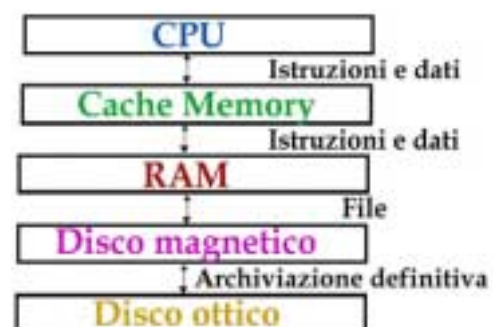
Nastri magnetici

Della stessa tecnologia degli hard disk e dei floppy sono i nastri magnetici che sono removibili e trasportabili a piacimento. La differenza sostanziale tra nastro magnetico e hard disk (o floppy) è che il primo è una memoria ad accesso sequenziale (mediamente molto più lenta) ma cambiano anche capacità di memorizzazione, intesa sia come densità di scrittura che come dimensioni del supporto, e motivazioni all'utilizzo (generalmente usati per copie di sicurezza o *backup*).

La memoria di massa: i dischi ottici

Sono i dischi che offrono la massima densità di memorizzazione (byte/cm^2) e per questo vengono utilizzati come supporti ad alta capacità di memorizzazione trasportabili.

Esistono diverse tipologie di dischi ottici, oggi suddivise in due grandi categorie: i Compact Disk e i Digital Versatile Disk.



Compact disk

I compact disk sono dispositivi periferici che impiegano la tecnologia ottica sviluppata da Philips e Sony nel 1983.

I dischi CD-ROM sono memorie a sola lettura, possono contenere qualsiasi tipo di dato binario, sono stampati da una matrice di metallo e non consentono di essere modificati o cancellati dall'utente. Inizialmente con una capacità fino a 650 MByte, attualmente i CD-ROM possono contenere circa 737 MBytes di informazioni. Sebbene l'accesso a questi dischi sia casuale, le informazioni non sono memorizzate su più tracce come negli hard disk, ma su un'unica traccia a spirale, il che comporta più lunghi tempi d'accesso. Esistono differenti tipi di compact disk:

- CD-DA (Compact Disc Digital Audio): sono dei CD musicali con una capacità massima di registrazione di 74 minuti di suono stereo digitale;
- CD-R (Compact Disk Recordable): sono CD scrivibili una sola volta;
- CD-RW (CD-ROM Rewritable): sono compact disks riscrivibili (un certo numero di volte).

Digital Versatile disk

La differenza sostanziale tra CD e DVD sta nel fatto che nei DVD-ROM il disco è composto da più strati e il raggio laser può essere selettivamente focalizzato su uno di questi strati: quindi è come avere più CD "impilati" uno sull'altro (memorie olografiche). Inoltre in questi ultimi le dimensioni delle celle sono notevolmente più piccole (circa 1/7 di quelle dei CD).

Anche tra i DVD esistono due categorie: i DVD-ROM (a sola lettura) e i DVD-RW (riscrivibili).

Lo schema che segue ha lo scopo di sintetizzare le caratteristiche principali dei vari supporti fisici ed i relativi attributi:

<i>Tecnologia utilizzata</i>	<i>Nome dispositivo</i>	<i>Velocità di accesso</i>	<i>Costo per milione di byte</i>	<i>Dimensioni</i>	<i>Volatili</i>
Elettronica a transistor	Registri CPU	$< 10 \times 10^{-9}$ sec (miliardesimi)	100 €	< 1 Kb	SI
	Cache	$< 20 \times 10^{-9}$ sec	100 €	0 - 512/1024 Kb	SI
Elettronica a condensatori	RAM	$< 1000 \times 10^{-9}$ sec	10 €	4 - 256/512 Mb	SI
Magnetica	Hard Disk	$< 10 \times 10^{-3}$ sec (millisecondi)	0,30 €	8 - 32 Gb	NO
	Floppy Disk	100×10^{-3} sec	0,25 €	1,4 Mb	NO
Ottica	Compact Disk	$< 100 \times 10^{-3}$ sec	0,01 €	650 Mb	NO
	Digital Versatile Disk	$< 100 \times 10^{-3}$ sec	0,05 €	4,7 - 18 Gb	NO