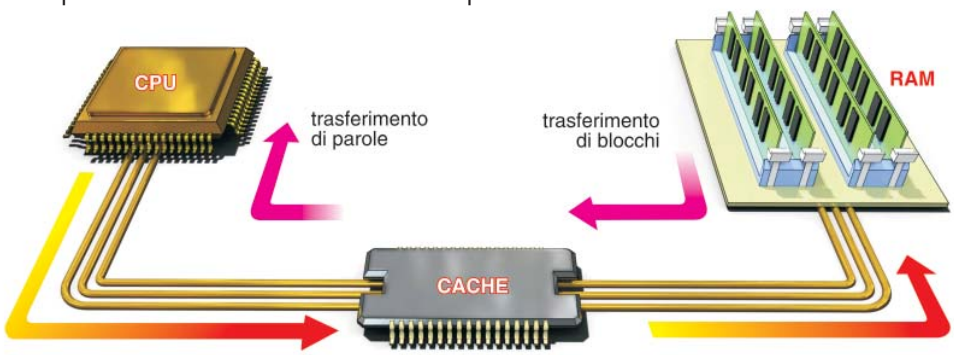
La memoria *cache* e la gerarchia delle memorie

La **memoria cache** è una memoria temporanea utilizzata per trasferire dati da un dispositivo a un altro operanti a velocità di lavoro diverse (uno veloce e l'altro lento).
L'introduzione di queste memorie consente di intervenire efficacemente in tutte le situazioni nelle quali l'avanzamento dei processi è condizionato dai rallentamenti dovuti ai tempi diversi di elaborazione da parte delle unità del sistema.
Possiamo trovare memorie cache:

- nella comunicazione tra memoria RAM e unità a disco,
- nel trasferimento dati tra memoria RAM e CPU quando la RAM non è in grado di servire tempestivamente le richieste della CPU.

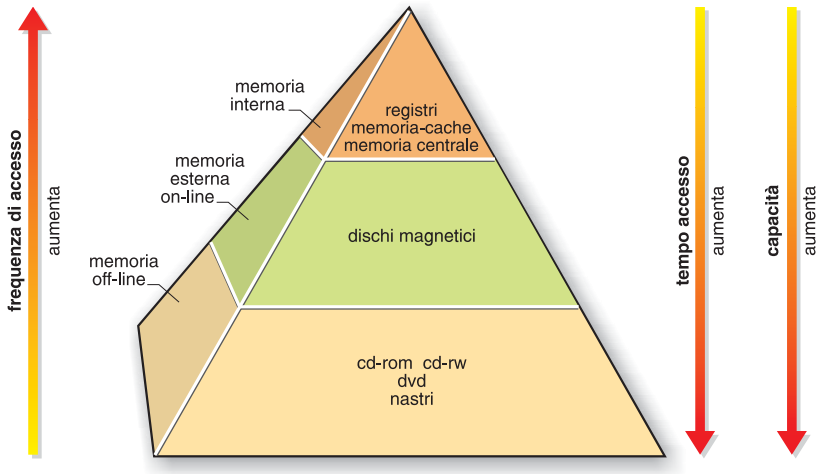
In commercio esistono infatti memorie RAM caratterizzate da velocità e costi diversi, adatte quindi a operare con CPU funzionanti a frequenze diverse.



Una CPU veloce, in assenza di cache, dovrebbe compiere a vuoto molti cicli di elaborazione (colpi di clock) prima che la RAM sia pronta a rispondere, data la notevole differenza di velocità tra CPU e RAM, oltre ai cicli necessari per attivare la connessione sul bus e l'invio delle richieste. Per questo si affianca alla normale memoria, supponiamo con tempo medio di accesso di 100 ns, una piccola *cache memory*, di dimensioni tipiche di 256 o 512 KB nel caso di personal computer, di velocità superiore (per esempio con tempo di accesso di 10 ns), che serve da rifornimento veloce per la CPU.
Quando il processore richiede un dato, per esempio una parola di memoria di 32 bit, lo cerca nella cache: se c'è, lo trasferisce nei registri del processore e lo utilizza; se non c'è, trasferisce nella cache un blocco di dati (per esempio di 16 parole) che comprende la parola cercata e porta la parola cercata dalla cache nei registri per utilizzarla. Riassumendo, se il dato è nella cache, il processore lo preleva in 10 ns; se la parola cercata non è nella cache, il processore impiega 10+100=110 ns per recuperarla. Il tempo medio di accesso dipende dalla probabilità di trovare il dato in cache (*cache hit*). Con i valori dell'esempio considerato, ipotizzando una percentuale di successo del 97%, si avrebbe un tempo medio di accesso di 13 ns.
Poiché nel corso dell'esecuzione dei programmi si evidenziano valori di *cache hit* superiori al 95%, l'idea di frapporre una memoria piccola, veloce e costosa tra il processore e la memoria centrale grande e lenta permette di costruire un sistema di memorie a due livelli, con la proprietà di avere capacità pari a quella della memoria più grande e la velocità di quella più veloce.

La ragione per cui i dati nella memoria cache sono utilizzati con maggior frequenza di quelli in memoria centrale risiede in due **principi di località**: se nel corso dell'esecuzione di un programma si fa riferimento a una cella di memoria è molto probabile che nell'immediato futuro si farà riferimento a celle di memoria vicine alla precedente (*località spaziale*); è molto probabile che una cella di memoria appena richiesta sarà richiesta nell'immediato futuro (*località temporale*).

Le diverse forme di memorizzazione presenti in un computer (registri, memoria cache, memoria centrale, dischi, nastri e dischi ottici) costituiscono una **gerarchia di memoria** a più livelli come quella illustrata nella figura della pagina seguente. In essa ogni coppia di memorie in livelli adiacenti può essere pensata come un sistema di memorie a due livelli simile a quello sopra descritto.



Scendendo nella gerarchia si osservano memorie caratterizzate dall'essere sempre più grandi, sempre più lente e sempre meno costose (per bit memorizzato). La caratteristica alla base del successo di questa organizzazione delle memorie è però la seguente: scendendo nella gerarchia le memorie sono di accesso sempre meno frequente, per il principio di località.
È per questa ragione che quando serve un dato che sta sul disco si trasferisce l'intero blocco che contiene quel dato: è molto probabile che nell'immediato futuro anche altri dati di quel blocco saranno utilizzati. È altresì molto probabile che nell'immediato futuro i dati del blocco saranno ancora necessari.
La tabella seguente sintetizza una serie di valori caratteristici nei personal computer per le prestazioni delle memorie della gerarchia in figura.

Tipo di memoria	Tempo di accesso	Capacità	Caratteristiche
Registri di memoria	1 - 3 ns	< 1KB	Interna all'unità centrale
Memoria cache	3 - 10 ns	512 KB - 4 MB	
Memoria centrale	50 - 200 ns	1 - 4 GB	
Disco magnetico	20 - 30 ms	50 GB - 1 TB	Esterna all'unità centrale in linea
Nastro	> 1 s	4 GB - 300 GB	Esterna all'unità centrale
Dischi ottici	> 1 s	650 MB - 4,7 GB	

Si osservino i salti nelle prestazioni quando si passa dalle forme di *memoria interna*, con tempi di accesso misurati in nanosecondi, alle forme di *memoria esterna* (rispetto all'unità centrale) che evidenziano tempi di accesso dell'ordine dei millesimi di secondo, fino alle forme di memoria *esterna fuori linea* con tempi di accesso misurati in secondi.
Se la memoria principale del computer è la memoria interna o *primaria* e la memoria su disco costituisce una forma di memorizzazione *secondaria*, le memorie esterne fuori linea, chiamate così per la necessità di dover montare il supporto fisico di memorizzazione nel drive prima di utilizzarle, prendono il nome di memoria *terziaria*.