Gerarchia delle memorie

Leonardo Bizzoni

May 20, 2023

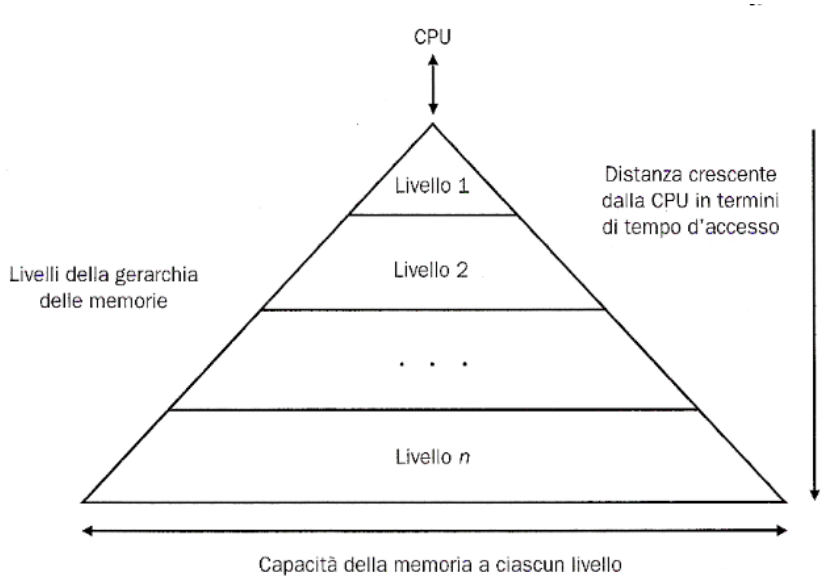
Un programma in certo istante di tempo non accede a tutto lo spazio di memoria indirizzabile con la stessa probabilità.

I 2 **principi di località** stabiliscono che un programma accede ad una porzione di memoria relativamente piccola ad ogni istante di tempo:

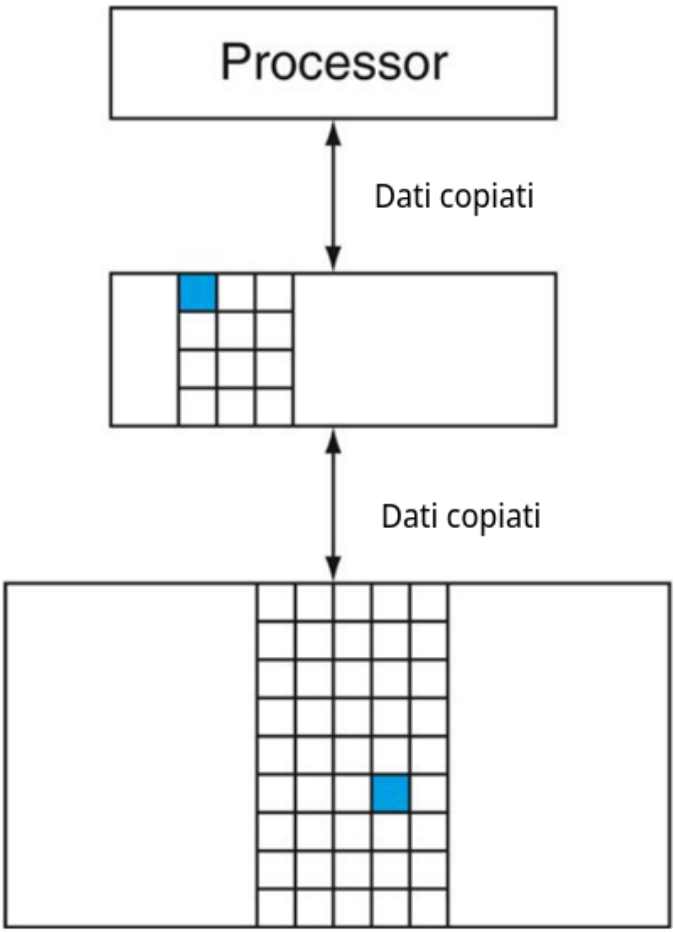
- **Località temporale**: quando si fa riferimento ad una cella di memoria c'è la tendenza a fare riferimento allo stesso elemento dopo poco tempo.
- **Località spaziale**: quando si fa riferimento ad una cella di memoria c'è la tendenza a fare riferimento ad indirizzi vicini dopo poco tempo.

Applicando i principi di località è possibile dare l'illusione di avere a disposizione una grande quantità di memoria e riuscire ad accedervi rapidamente, ovvero creare una **gerarchia di memoria**:

- Memoria **interna alla CPU** costituita da registri
- Memoria cache
- Memoria centrale indirizzabile, di capienza superiore ma più lenta
- **Memorie secondarie/permanenti**



Un livello di memoria più vicino al processore contiene un sottoinsieme di dati memorizzati nel livello direttamente sottostante. Tutti i dati si trovano nel livello più basso.



L'unità di informazione più piccola che può essere presente/assente in una gerarchia di memoria viene chiamata **blocco/linea**.

Se un dato richiesto dalla CPU è presente nel livello di memoria direttamente precedente, la richiesta è un **hit** (*successo nell'accesso*). Se invece non dovesse essere presente è un **miss** (*fallimento nell'accesso*) e la richiesta viene passata ai livelli inferiori.

L' **hit rate** rappresenta la frazione degli accessi alla memoria nei quali il blocco è stato trovato. Il **miss rate** rappresenta la frazione deli accessi alla memoria nei quali il blocco non è stato trovato ($1 - \text{HitRate}$).

Il **tempo di hit** è il tempo di accesso al livelli di memoria direttamente

precedente compreso il tempo necessario per determinare se è un hit o un miss.

Il **tempo di miss** è il tempo necessario a sostituire un block del livello precedente con un nuovo blocco dal livello inferiore incluso il tempo per trasferire il dato alla CPU.

Frequenza di hit = $\frac{\text{numero di hit}}{\text{numeri di richieste}}$

Frequenza di miss = $\frac{\text{numero di miss}}{\text{numeri di richieste}}$