CACHING

Benvegnù Nicholas, 2019

Motivazioni

Memorie diverse e periferiche diverse hanno velocità e tempi di risposta differenti.

Per poter riuscire a sfruttare appieno le memorie e periferiche più veloci è possibile inserire delle piccole memorie rapide quanto quella rapida, cercando di limitare i colli di bottiglia e massimizzarne l'uso.

La cache è comoda

La cache è una memoria molto rapida per permettere di eseguire le operazioni senza dover aspettare il recupero dei dati trattati e facilmente raggiungibile dal suo utilizzatore principale senza dover interagire con i vari sistemi di controllo della memoria eliminando i passaggi e rallentamenti non voluti.

La cache è preziosa

Queste memorie sono molto specializzate e possono occupare aree dei die importanti.

Tali caratteristiche le rendono molto costose da produrre ed inserire.

Anche la gestione di grandi quantità di memoria cache risulta complessa da realizzare.

La cache è preziosa

La loro quantità limitata, può condizionare la logica e con la quale le parti più critiche in termini di velocità e memoria dei programmi vengono scritte ed eseguite.

Questo vale particolarmente per programmi che necessitano un elevato sfruttamento dell'HW.

Cache vs Buffer

Quando si usano questo tipo di memorie, è possibile dividerle in due categorie:

- Se tale memoria inserita all'interno dell'utilizzatore più rapido viene utilizzata per rendere più rapide le elaborazioni di specifici dati in modo da sfruttarne al massimo la velocità d'elaborazione, viene indicata col termine «cache»;

Cache vs Buffer

Quando si usano questo tipo di memorie, è possibile dividerle in due categorie:

- Se la memoria si trova all'interno della periferica più lenta in operazioni di input/output al fine di massimizzare l'utilizzo del bus dove passano i dati e gli stessi semplicemente vi transitano, si è soliti indicarla come «buffer».

Tipi di Cache

Non esiste un singolo tipo di memoria cache (e di buffer) perché vengono esistono molti ambiti d'utilizzo per la cache, con un'enorme variabilità per quanto riguarda l'HW e lo scopo per il quale il sistema ed i componenti vengono previsti.

I due tipi di cache in HW più noti sono delle CPU e degl'HDD.

Tipi di Cache – CPU

I sistema di cache all'interno delle CPU, è composto da più livelli di memorie veloci alle quali i singoli core possono accedere direttamente, senza dover passare per il controller della memoria RAM, il bus di memoria e la RAM stessa.

Il passaggio per il controller viene fatto nei due sensi.

Tipi di Cache – CPU

L'organizzazione delle cache all'interno delle CPU segue un *ordine gerarchico*, che prevede

- Le memorie di livello gerarchico più basso sono accessibili immediatamente solamente al core cui sono associate;
- Le memorie di livello superiore sono visibili e condivise fra i vari core.

Tipi di Cache – CPU

Quando la CPU vuole accedere ad un dato in memoria, questo viene dapprima cercato all'interno della cache di livello inferiore, in caso non fosse presente, passa alla cache di livello superiore e così via.

Se non viene trovato neanche lì, passa ad un livello di cache ulteriormente più alto e così via.

Tipi di Cache – CPU

Se il dato non è presente neanche nella cache di livello massimo, viene richiesto al controller della memoria di recuperare il dato dalla RAM.

Quando un dato cercato nella cache è presente si ha in «cache hit» mentre se esso non è presente si verifica un «cache miss».

Algoritmo di cache

In caso di cache hit il dato è subito utilizzabile dalla CPU, mentre in caso di cache miss, oltre a dover andare a recuperare il dato da un'altra memoria, parte anche l'operazione per l'aggiornamento della cache.

Il rapporto fra cache hit e cache miss è un indice della qualità dell'algoritmo di cache.

Algoritmo di cache

Visto che in caso di cache miss, il tempo per completare le operazioni inerenti i dati presenti in cache e la cache stessa è notevole, soprattutto se interviene anche la RAM, sono stati sviluppati algoritmi specifici per il riempimento e l'aggiornamento dei dati in cache, che possono partire da un semplice FIFO a complesse euristiche.

Tipi di Cache – HDD

Gli Hard Disk meccanici sono naturalmente più lenti di diversi ordini di grandezza rispetto a tutta l'elettronica che li circonda in quanto hanno parti meccaniche in movimento ed il modo con il quale sono organizzati i dati, anche sfruttando meccaniche sofisticate e tecniche d'arciviazione specializzate, questa differenza resta grande.

Tipi di Cache – HDD

Per ridurre questa differenza, all'interno degli Hard Disk vengono inserite delle memorie cache che aiutano sia in caso di lettura (con algoritmi che selezionano i dati più letti) che in caso di scrittura (solitamente si comporta come un semplice buffer, accompagnato da ottimizzazioni per la successiva lettura sequenziale).

Tipi di Cache – SW

Il concetto di cache viene usato anche in ambito SW, con aree di memoria (RAM ed HD) che vengono riservate per contenere dati per poterli riutilizzare senza doverli elaborare nuovamente.

Come per la cache in HW la scelta di quali dati salvare e per quanto tempo può influenzare la reattività di un'applicazione.

Tipi di Cache – SW

A differenza di quanto visto precendentemente, generalmente è possibile inserire in memoria una quantità di dati maggiore, ma è comunque necessario limitarne l'utilizzo per non occupare troppa memoria inutilmente.

Questo concetto viene sfruttato da quasi ogni sistema informatico odierno.

Tipi di Cache – SW

La cache viene sfruttata da Sistemi Operativi, applicazioni PC/cellulare, browser web, server, DB, motori di ricerca…

Un'importante parte di questi utilizzi sono correlabili alle naturali caratteristiche di ridotta velocità, instabilità ed imprevedibilità delle trasmissioni via Internet.

Cache dei browser

La cache sfruttata dai browser consente di non dover riscaricare e rielaborare le risorse già trattate.

Il limite di questa tecnica riguarda le pagine contenenti dati che s'aggiornano frequentemente o con dati che sono per loro natura variabili in base alle diverse richieste effettuate.

Cache dei browser

Anche i vari server usano la cache per ottimizzare i loro tempi di risposta, salvando in memoria le pagine più richieste.

Nel caso una risorsa non sia presente nativamente nel server, ma fosse una risorsa salvata (od un mirror), sarà necessario aspettarne la propagazione degli aggiornamenti.

Cache nei database

I database sfruttano automaticamente il caching per alleggerirne il carico, con le stored procedure che vengono salvate una volta eseguite in previsione di un loro uso futuro; mentre le *query semplici* vengono valutate ogni singola volta come fossero query nuove e differenti (anche se al loro interno la query è identica).