Esercizi di ricapitolazione – parte III

Cos'è il branch prediction buffer? Cosa si intende per pipeline superscalare, superpipeline e scheduling dinamico? Che cosa sono le reorder station? E le reservation station? A cosa serve l'operazione di renaming dei registri? E' corretto dire che internamente la CPU di un Pentium IV ha un'ISA RISC e perchè? Quante cache primarie ha un Pentium IV e perchè? Cosa si intende per esecuzione fuori-ordine? Che differenza c'e' tra un multi-core e un cluster? Cos'è lo snooping? Chi lo attua? Perchè? Cos'è la coerenza? Come può essere garantita dall'hardware?

Cosa si intende per esecuzione speculativa? Cos'è un branch prediction buffer? Come si può migliorare il funzionamento di una pipeline per la gestione dei salti condizionati?

Definire lo schema di massima dell'Architettura della pipe-line di un Pentium IV e descrivere quali sono i componenti principali e quali possono essere i problemi principali.

Qual'è il ruolo della memoria? Quali funzioni si possono eseguire sulla memoria? Cosa rappresenta l'altezza e l'ampiezza della memoria e come si calcola la capacità? Come è definita la parola di memoria? Qual'è la relazione tra capacità di memoria e numero di bit di indirizzamento?

Cosa esprime il principio di località di una memoria? Cosa contiene una memoria cache? Può il contenuto di una memoria cache essere diverso dal contenuto della memoria principale?

Cosa si intende per Hit e Miss? Hit rate e miss rate? Può essere la somma di Miss rate e Hit rate maggiore di 15? Cosa si intende per Write through e Write back in una memoria cache? Cosa si intende per LRU? Cosa si intende per modalità a buffer circolare?

Data una memoria cache di 64Kbyte ed una RAM di 1Gbyte, a quanti bit devo dimensionare il campo TAG in caso di una memoria cache a mappatura diretta? E in caso di una cache a 2 vie?

Dato un indirizzo di memoria di 32 bit, specificare il significato dei singoli bit nel caso di utilizzo di una memoria cache con le seguenti caratteristiche:

Cache a mappatura diretta di 128Kbyte, con linee contenenti 8 parole di 4 byte ciascuna.
Cache a 2 vie di 128Kbyte, con 2 banchi e linee contenenti 8 parole di 4 byte ciascuna.
Cache associativa di 128Kbyte con linee ciascuna contenente 8 parole di 4 byte ciascuna.
Per ciascuna delle 3 cache disegnare il circuito di lettura e scrittura.

Disegnare le seguenti tre memorie cache:
Cache a mappatura diretta di 128 byte con linee contenenti 2 parole di 4 byte.
Cache a 2 vie di 128 byte con linee contenenti 2 parole di 4 byte.
Cache associativa con linee contenenti 2 parole di 4 byte.
Data l'istruzione lw $t0, 1024($zero), specificare all'interno delle cache a), b), c) dove si trova la parola che deve essere letta?

Specificare per una memoria cache a chi viene inviato: il segnale di Miss, il segnale di Hit, il dato letto; e da dove proviene il dato scritto.

Cos'è l'interleaving di una memoria? Come è costruita una memoria SRAM? Cosa significa l'acronimo SRAM? Qual'è il ruolo dell'uscita "three-state" in una memoria SRAM? Discuterlo con un esempio. Come vengono gestiti i banchi di memoria?

Qual'è il principio di funzionamento di una DRAM? Cosa è una SDRAM? Quando una SDRAM lavora in "burst mode"? Cosa rappresentano i segnali CAS e RAS di una DRAM? Cos'è il refresh della memoria? Si può leggere la memoria durante il refresh? Perché? Può una SDRAM essere letta in modalità asincrona?

Qual è il vantaggio / svantaggio dell'organizzazione a matrice di una memoria?

Dato il seguente segmento di codice, descrivere istruzione per istruzione cosa succede in una cache a 2 vie, dove ciascun banco è di 1Kbyte con linee di 4 word, e con tutte le linee con dati non validi:
lw $s0, 64($zero)
lw $s0, 8($zero)
lw $s0, 1032($zero)
lw $s0, 2056($zero)
lw $s0, 4($zero)
sw $s1, 4($zero)

Disegnare la porta di lettura e scrittura di una cache a mappatura diretta di 2Kbyte e 8 linee, di una cache a 2 vie con banchi di 2Kbyte e linee di 8 word e di una memoria cache associativa di 2Kbyte e 8 linee.

Dato un indirizzo di 32 bit, come vengono utilizzati i bit per indirizzare una memoria cache a k-vie?

Quali requisiti imporreste alla cache primaria? E alla cache secondaria? Data una dimensione di cache, cosa succede all'aumentare della lunghezza di una linea e all'aumentare del numero di vie?

Cosa si intende per split-cache?

Come si utilizza la tecnologia three-state all'interno delle memorie? Si applica indifferentemente alle memorie SRAM e DRAM? Perché?

Cos'è un bus? Cos'è l'arbitraggio? Cosa rappresenta il segnale di "bus grant"? Descrivere gli schemi di arbitraggio centralizzato e gli schemi di arbitraggio distribuiti (con e senza autoselezione).

Descrivere i requisiti di funzionamento dei tre tipi di bus principali: processore-memoria, backplane e I/O. Come viene sincronizzata la trasmissione di dati sul bus?

Cosa è il device controller? Quali sono i suoi componenti principali? A cosa serve? Come vengono indirizzate le periferiche?

Descrivere una procedura di handshaking tipica per l'accesso ad un bus sincrono ed asincrono. Cosa si intende per bus-master? A cosa si riferisce: al device controller, al bridge o al dispositivo?

Cosa si intende per transazione sul bus?

Come viene gestito l'I/O dall'ISA di un'architettura MIPS e da un'architettura INTEL?

Come viene gestito l'I/O a controllo di programma? Cos'è il polling? Come viene gestito l'I/O tramite interrupt? Cosa è il DMA? Che cos'è lo "spin lock"?

Esercizio. Supponiamo di valutare il costo per una CPU con frequenza pari a 1Ghz per trasferire 64Mbyte di dati da un *Hard-disk. Supponiamo che ad ogni accesso vengano trasferiti 64 byte. Il tempo richiesto ad una*

CPU per il trasferimento di 64 byte è di 200 cicli_clock che devono essere sommati al tempo per accedere alla periferica che è di 300 cicli di clock in modalità a controllo di programma e di 400 cicli di clock in modalità interrupt. Confrontate le prestazioni con un trasferimento in DMA nel quale vengono trasferiti 6,400 parole tenendo conto che il tempo richiesto per l'avviamento della DMA è di 500 cicli di clock e per la chiusura è di 800 cicli di clock. Quale tra le modalità a controllo di programma e ad interrupt è più efficiente in questo caso (motivare la risposta)?
Valutiamo le diverse modalità di trasferimento:
a controllo di programma (300 cicli di clock)
ad interrupt (400 cicli di clock)
mediante DMA (500 cicli clock per l'avviamento e 800 cicli di clock per la chiusura).
A controllo di programma: 64Mbyte / 64byte = 1M accessi. Tempo di CPU: 1Maccessi * (200+300)cicli = 500 * 10^6. Percentuale di sfruttamento della CPU: 50%.
Mediante interrupt: 64Mbyte / 64 byte = 1M interrupt. Tempo di CPU: 1Minterrupt * (200+400) cicli di clock. Percentuale di utilizzo della CPU: 60%.
Mediante DMA: 64Mbyte / 6400 word = 2,500 DMA => 2,500 * (500 + 800) = 3,25 Mega cicli di clock. Percentuale di utilizzo della CPU: 0,325%.
Ripetete le valutazioni per trasferire dati da tastiera (i quali richiedono una frequenza di 10byte/s, un tempo di CPU di 10 cicli_clock / byte per l'operazione di I/O e trasferisce 1 byte per ogni accesso).
A controllo di programma: 10 accessi / s * (10 + 300) cicli_clock = 3100 cicli_clock. Percentuale di sfruttamento della CPU: 3,1 / 10^6.
Mediante interrupt: 10 accessi / s * (10 + 400) cicli_clock = 4100 cicli_clock. Percentuale di utilizzo della CPU: 4,1 / 10^6.
Mediante DMA. Non ha senso. Trasferisco 1 byte alla volta.

Descrivere l'utilizzo principali dei registri nelle architetture IA-32. Cosa sono i segmenti? Modalità di indirizzamento dei dati nelle architetture IA-32. Come vengono gestite le operazioni di I/O nelle IA-32? Come funzionano i modificatori delle istruzioni dell'ISA delle architetture IA-32?

Cosa è lo SPEC? Cosa è il CPI? Perchè sono nati i benchmark? Enunciare la legge di Amdhal. Sotto quali ipotesi riesco ad aumentare la velocità di esecuzione in architetture multi-processori? E la velocità di un server?

Si deve valutare un miglioramento di una macchina per l'aggiunta di una modalità vettoriale. La computazione vettoriale è 20 volte più veloce di quella normale. La *percentuale di vettorizzazione* è la porzione del tempo che può essere spesa usando la modalità vettoriale.
Disegnare un grafico che riporti lo speedup come percentuale della computazione effettuata in modo vettoriale.
Quale percentuale di vettorizzazione è necessaria per uno speedup di 2?
Quale percentuale di vettorializzazione è necessaria per raggiungere la metà dello speedup massimo?
La percentuale di vettorizzazione misurata è del 70%. I progettisti hardware affermano di potere raddoppiare la velocità della parte vettoriale se vengono effettuati significativi investimenti. Il gruppo che si occupa dei compilatori può incrementare la percentuale d'uso della modalità vettoriale.
Quale incremento della percentuale di vettorizzazione sarebbe necessario per ottenere lo stesso guadagno di prestazioni?
Quale investimento raccomandereste?

Descrivere l'architettura dei bus di un'architettura Intel (Pentium 4). Quale funzione svolge ciascun bus?

Dove si trova fisicamente il circuito di arbitraggio di un bus? E' un circuito sequenziale o combinatorio (motivare la risposta)?

Quali circuiti vengono utilizzati per un SRAM: flip-flop o latch? Motivare la risposta.

Descrivere la struttura fisica interna di un disco. Fare uno schizzo ed indicare quali sono i settori e le tracce.

Cosa si intende per multiple zoned recording? Perché si utilizza?

Cosa rappresenta il tempo di seek? E' costante all'interno di uno stesso disco? E' costante per tutto i dischi?

Cosa rappresenta il tempo di attesa? Può essere uguale su dischi diversi e perché?

Cosa si intende per gerarchia di memoria? Perché viene richiesta una gerarchia?

Cosa si intende per bus seriale? Che differenza c'è tra bus PCI-Express e PCI64?

Definire il ruolo dei bridge nella architetture? Dove si trovano? Cosa prevedete ci sia all'interno di un bridge?

Descrivere la struttura di una GPU nVidia della serie 8000. Cosa si intende per CUDA? Quali sono i punti più significativi dell'architettura? Cosa si intende per fragment processor? Per Kernel? Per Blocco? Per griglia? Che differenza c'e' tra Shared Memory, Local Memory, Global Memory, Constant Memory e Texture Memory in una scheda nVidia della serie 8000? Quali accortezze occorre avere nella