

# Architettura degli Elaboratori

Appello del 5 settembre 2011

Riportare su tutti i fogli consegnati nome, cognome, matricola, corso, programma d'esame. I risultati saranno comunicati appena disponibili sulle pagine web dei docenti.

## Domanda 1 (tutti)

Una unità di elaborazione  $U$  contiene, nella sua Parte Operativa, un componente logico memoria  $M$ , di capacità complessiva  $C = 1G$  parole, costituito da 8 componenti logici memoria  $M_0, \dots, M_7$  aventi la stessa capacità. Gli indirizzi di  $M$  sono distribuiti secondo l'organizzazione interallacciata.

Le operazioni esterne sono definite come segue:

- *operazione 0*: ricevendo un indirizzo  $IND$ , con il vincolo che  $IND \bmod 8 = 0$ , legge un blocco di 8 parole consecutive a partire da  $IND$ . Tali parole sono inviate su interfacce di uscita distinte;
- *operazione 1*: ricevendo un indirizzo  $IND$  e un dato  $X$ , scrive  $X$  all'indirizzo  $IND$ ;
- *operazione 2*: ricevendo un dato  $X$ , con il vincolo che tale valore sia presente in  $M$  una e una sola volta, invia in uscita l'indirizzo della locazione avente contenuto uguale a  $X$ .

a) Scrivere il microprogramma di  $U$  e determinare il tempo medio di elaborazione di ognuna delle tre operazioni esterne in funzione di  $C$  e  $t_p$ . È richiesto di minimizzare il numero medio di cicli di clock per ogni operazione esterna. Ogni componente  $M_j$  ha tempo di accesso  $10t_p$ , una ALU ha ritardo di stabilizzazione  $5t_p$ . Fornire adeguate spiegazioni.

b) Sia  $\tau_U$  è il ciclo di clock di  $U$ . Si supponga che  $U$  venga connessa ad una CPU con collegamenti aventi latenza di trasmissione trascurabile. Determinare, in funzione di  $N$  e  $\tau_U$ , il tempo di completamento dei due seguenti programmi, entrambi costituiti da un ciclo di  $N$  iterazioni.

b1) Nel primo programma, le interazioni con  $U$  consistono solo in operazioni di scrittura di una singola parola, una per ogni iterazione. Senza considerare il ritardo introdotto da  $U$ , le richieste di scrittura sarebbero distanziate di un tempo medio uguale a  $\tau_U/4$ .

b2) Nel secondo programma, le interazioni con  $U$  consistono solo in operazioni di lettura di blocco, una ogni 8 iterazioni. Senza considerare il ritardo introdotto da  $U$ , le richieste di lettura blocco sarebbero distanziate di un tempo medio uguale a  $\tau_U$  (inclusa la copia del blocco internamente alla CPU).

## Domanda 2 (tutti)

Si consideri la seguente computazione operante sugli array di  $N$  interi  $A$  e  $B$ , con  $F$  funzione nota:

$$\square i = 0 \dots N - 1 : B[i] = F(A[i])$$

Spiegare cosa c'è da aspettarsi circa le differenze sul tempo di completamento implementando la computazione sulle seguenti architetture:  $S1$  basata su una unità di elaborazione dedicata,  $S2$  basata su una CPU convenzionale e (**solo per NEW e OLD-0**)  $S3$  basata su una CPU pipeline. Tutte le architetture hanno lo stesso ciclo di clock e la stessa gerarchia memoria principale – cache.  $A$  e  $B$  sono allocati in memoria.

## Domanda 3 (NEW, OLD-0)

Per una architettura CPU pipeline, spiegare la struttura, il funzionamento, ed eventuali ottimizzazioni dinamiche, del sottosistema DM.

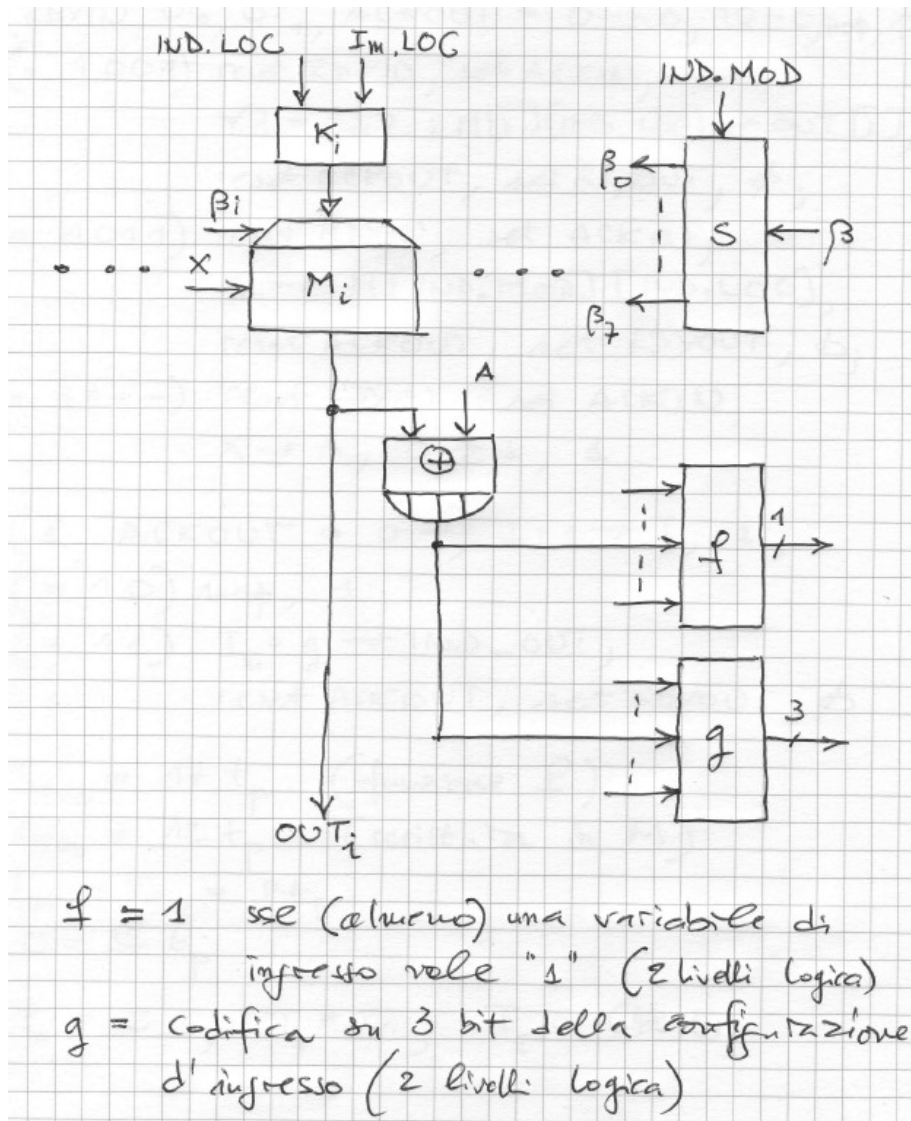
## Domanda 3 (OLD-1)

Spiegare le funzionalità di creazione e di caricamento di un processo, specificando le entità coinvolte, le azioni svolte e le strutture dati utilizzate.

### Traccia di soluzione

#### Domanda 1 (tutti)

a) La struttura del componente logico memoria M e delle principali reti combinatorie utilizzate per le operazioni esterne è la seguente:



Per un indirizzo  $addr$  di  $M$ , con  $addr.MOD$  si indica l'identificatore del modulo, dato dai 3 bit meno significativi, e con  $addr.LOC$  si indica l'indirizzo all'interno del modulo, dato dai 27 bit più significativi.

Il vincolo della minimizzazione del ciclo di clock implica che la prima e la seconda operazione esterna devono avere tempo medio di elaborazione uguale a un ciclo di clock, e la terza uguale a circa 64K cicli di clock con un algoritmo di ricerca sequenziale. Infatti, la ricerca viene effettuata in parallelo su un blocco 8 parole ad ogni iterazione dell'algoritmo.

Poiché non è richiesto di minimizzare il tempo di elaborazione complessivo, ma solo il numero di cicli di clock, vanno bene sia una soluzione che, nella terza operazione esterna, utilizzi una variabile di condizionamento

complessa (confronto in parallelo su un blocco), oppure una soluzione con controllo residuo. Scegliamo la prima soluzione.

Utilizzando le funzioni  $f$  e  $g$  di figura (più volte spiegate nel testo e in altri esercizi), il microprogramma è il seguente:

$\phi$ . ( $RDYIN, OP_0, OP_1, ACKOUT = 0--0, 10-0$ )  $nop, \phi$ ;  
 (= 1001) reset  $RDYIN$ , set  $ACKIN$ ,  
 $\forall i = 0..7 : M[i][IND.LOC] \rightarrow OUT[i]$ ,  
 reset  $ACKOUT$ , set  $RDYOUT, \phi$ ;  
 (= 1011) reset  $RDYIN$ , set  $ACKIN$ ,  
 $X \rightarrow M[IND.MOD][IND.LOC]$ ,  
 reset  $ACKOUT$ , set  $RDYOUT, \phi$ ;  
 (= 11--) reset  $RDYIN$ , set  $ACKIN$ ,  
 $X \rightarrow A, 0 \rightarrow I, 1$

1. ( $\phi, ACKOUT = 0-$ )  $I+1 \rightarrow I, 1$ ;  
 (= 10)  $nop, 1$ ;  
 (= 11)  $I_n \circ g \rightarrow IND\_OUT$ ,  
 reset  $ACKOUT$ , set  $RDYOUT, \phi$

$T_{wP_0} = 17 t_p$  (funzione  $f$ )  
 $T_{SP_0} = 12 t_p$  (scrittura in  $M$ )  
 $T_{wPC} = T_{SPC} = 2 t_p$   
 $\tau = 32 t_p$

$T_0 = \tau, T_1 = \tau, T_2 \sim 64k \tau$

Il calcolo del ciclo di clock è mostrato nella stessa figura.

**b)**  $\tau_U$  è il ciclo di clock calcolato in precedenza. Essendo  $U$  una singola unità di elaborazione, la memoria interallacciata viene sfruttata con la massima banda solo per le letture di blocco (8 accessi per  $\tau_U$ ), mentre la banda offerta alla scritture è di *una sola scrittura* per  $\tau_U$ . Quindi:

b1) La CPU deve adattarsi alla banda in scrittura di  $U$ , da cui:

$$T_c = N \tau_U$$

b2) La CPU alterna 8 iterazioni, di durata complessiva  $\tau_U$ , ed una lettura di blocco, da cui:

$$T_c = N \tau_U / 4$$

### **Domanda 2 (tutti)**

La computazione data:

$$\forall i = 0 .. N - 1 : B[i] = F(A[i])$$

è caratterizzata da sola località negli accessi ad A e B, quindi, in qualsiasi architettura, il tempo di completamento è dato da:

$$T_c = T_{c-id} + N T_{trasf} / \sigma$$

Il secondo addendo è (largamente) indipendente dall'architettura. Assumendo che nelle architetture S2 e S3 la lettura delle istruzioni non influisca sensibilmente sulle prestazioni, il primo addendo è diverso da architettura a architettura per quanto riguarda il parallelismo esplicitabile dall'algoritmo (si veda la teoria e altri esercizi più volte discussi). Rispetto alla S1, le architetture S2 e S3 impiegano un numero di cicli di clock  $a$  volte maggiore, dove la costante  $a$  è dell'ordine di 5 – 10 per S2, e dell'ordine di 2 – 3 per la CPU pipeline scalare, riducibile a circa 1 per CPU superscalari.

*Questa spiegazione deve essere opportunamente espansa da parte dello studente.*

### **Domanda 3 (NEW, OLD-0)**

Vedere il materiale didattico:

- testo, cap. XI, sez. 1, 3;
- materiale integrativo, in particolare sez. 2.2 e 3.3.3.

### **Domanda 3 (OLD-1)**

Vedere il testo, Cap. V, sez. 2 con integrazioni nell'Errata-Corrige, e Cap. VI, sez. 1.