Corso di Laurea in Informatica
Architettura degli elaboratori

Architettura degli Elaboratori 2020-2021

Input-Output
Periferiche
Interrupt
DMA

Claudia Raibulet
Claudio Ferretti

Tecniche di gestione ingresso-uscita

- Input-output – Periferiche
- Bus di sistema
- Periferiche mappate in memoria
- Controllo di programma
- Banda passante e latenza
- Input-Output gestito direttamente dal programma
- Input-Output con interruzioni
- Input-Output con DMA (Direct Memory Access)
- Periferiche in SPIM

Input/Output

- I/O – insieme di architetture e dispositivi per il trasferimento di informazioni da e verso l'elaboratore
- Dispositivi eterogenei per:
  - Velocita' di trasferimento
  - Latenze
  - Sincronizzazione
  - Modalita' di interazione (sia con l'uomo sia con la macchina)

Bus di Sistema

- Esistono vari tipi di bus nei computer di oggi
- In questo corso: bus comune, di sistema che collega la CPU sia con la memoria sia con le periferiche
- Il bus di sistema e' composto da:
  - Bus di dati: le linee per trasferire dati e istruzioni da/verso dispositivi
  - Bus di controllo: trasporta informazioni per la definizione delle operazioni da compiere e per la sincronizzazione tra i dispositivi
  - Bus degli indirizzi: la CPU trasmette gli indirizzi di memoria o di periferica che identificano i dati da leggere/scrivere dalla memoria/periferiche
- Tutte le unita' dell'elabratore sono conesse al bus

Bus di Sistema

- Vantaggi: elevata flessibilita', semplicita', basso costo
- Svantaggi: gestione complessa del canale condiviso

Periferiche

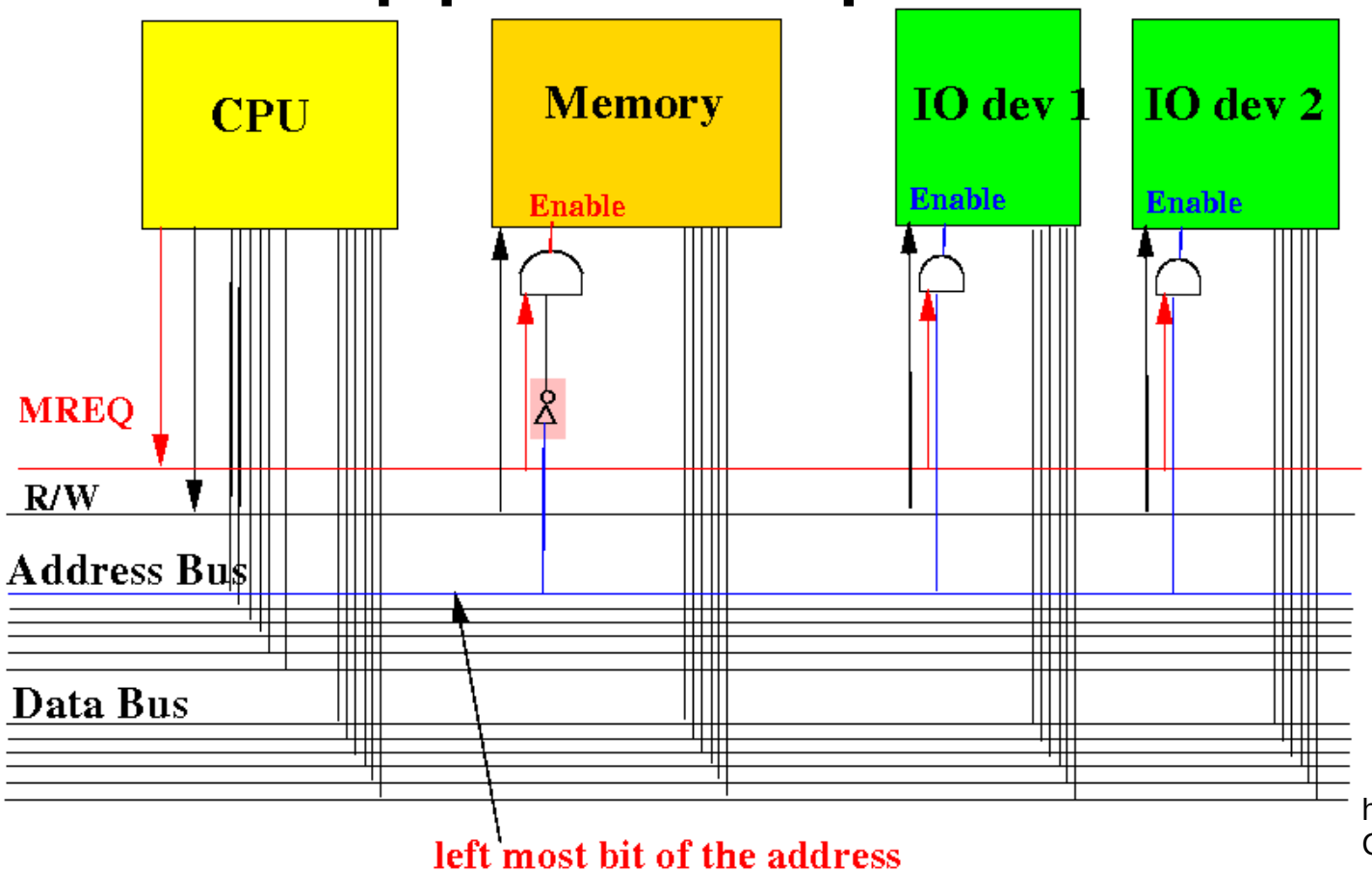
- Dispositivi per I/O di informazioni
- Collegato alle CPU tramite il bus di sistema e/o interfacce
- Le interfacce sono standardizzate per la comunicazione
- Le interfacce hanno una componente hardware (e.g., il controllore della periferica) e una componente software (e.g., driver)

- Struttura di un'interfaccia (e.g., tramite i registri della periferica):
  - Contiene registri di dati (oppure buffer)
  - Ha associato registri di stato

Periferiche mappate in memoria

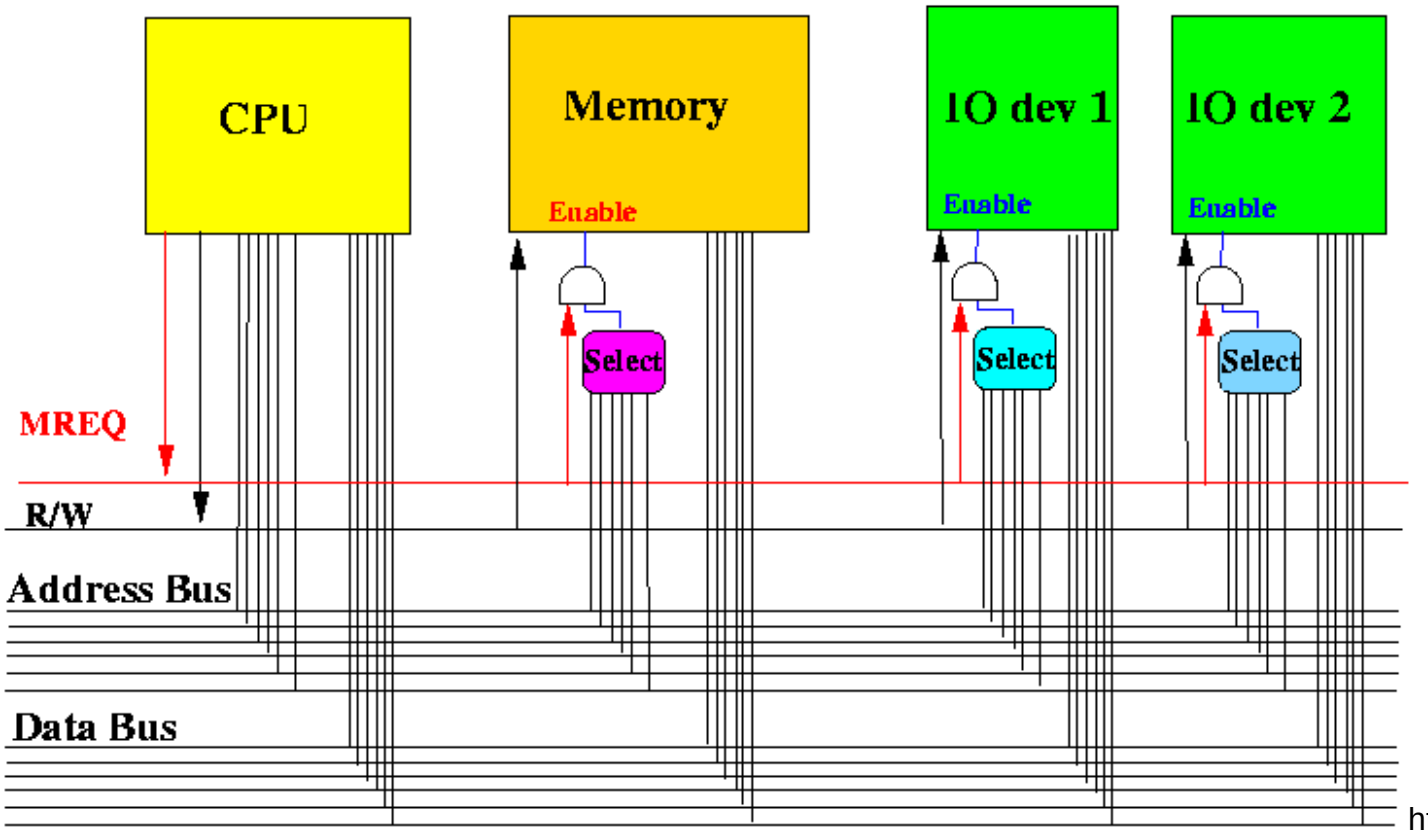
- Una parte della memoria riservata al sistema viene usata per la comunicazione con le periferiche
- Ogni periferica ha uno spazio dedicato nella memoria; a ogni periferica viene assegnato un identificatore unico (i.e., indirizzo unico)
- I registri (dell'interfaccia della periferica) sono mappati in memoria
- Tali registri non sono accessibili da un programma utente
- I programmi utente devono passare dal SO per accedere a questi registri riservati
- L'accesso alla periferica e' simile all'accesso alla memoria (e.g., lw, sw) dal punto di vista della CPU
- Osservazione: non tutte le architetture degli elaboratori usano questa soluzione

Un semplice selettore per mappare dispositivi in memoria



http://www.mathcs.emory.edu/~cheung/Courses/355/Syllabus/6-io/mem-map-io2.gif

Un selettore generale per mappare dispositivi in memoria



http://www.mathcs.emory.edu/~cheung/Courses/355/Syllabus/6-io/mem-map-io.gif

Periferiche mappate in memoria

- Una parte della memoria riservata al sistema viene usata per la comunicazione con le periferiche
- Ogni periferica ha uno spazio dedicato nella memoria; a ogni periferica viene assegnato un identificatore unico (i.e., indirizzo unico)
- I registri (dell'interfaccia della periferica) sono mappati in memoria
- Tali registri non sono accessibili da un programma utente
- I programmi utente devono passare dal SO per accedere a questi registri riservati
- L'accesso alla periferica e' simile all'accesso alla memoria (e.g., lw, sw) dal punto di vista della CPU
- Osservazione: non tutte le architetture degli elaboratori usano questa soluzione

Registri di interfaccia della periferica

- Registro di stato della periferica
  - Rappresenta lo stato della periferica
  - Viene letto dalla CPU
  - Esempi di stato: periferica pronta per ricevere dati, dati richiesti disponibili per il trasferimento
- Registro dei dati della periferica
  - Rappresenta i dati di input o di output in base al tipo della periferica

Passi di I/O (versione molto semplificata)

- CPU interroga lo stato della periferica
- La periferica restituisce il suo stato
- Se la periferica e' pronta a trasmettere/ricevere dati, la CPU richiede il trasferimento dei dati
- La CPU invia o riceve i dati

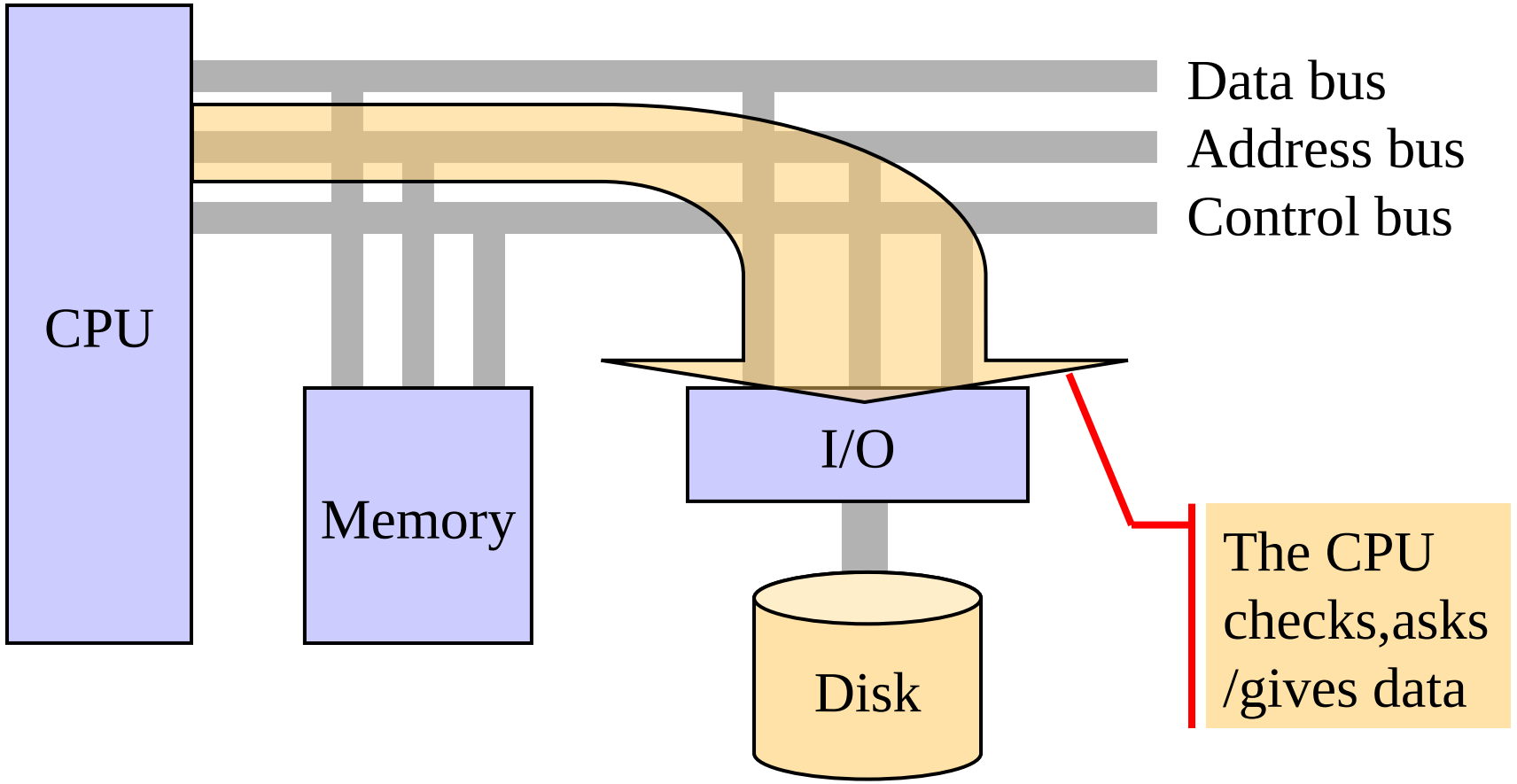
Tecniche di gestione I/O

- I/O gestito da programma (Programmed I/O)
- I/O guidato da interrupt (Interrupt Driven I/O)
- Accesso Diretto alla Memoria (Direct Memory Access –DMA)

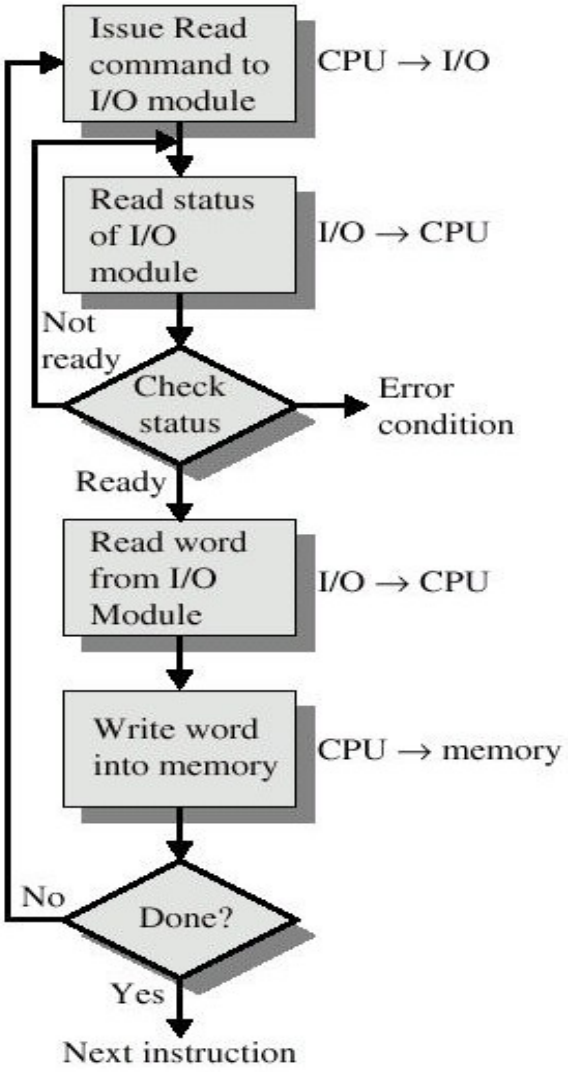
	No Interrupts	**Use of Interrupts**
I/O-to-memory transfer through processor	Programmed I/O	Interrupt-driven I/O
Direct I/O-to-memory transfer		Direct memory access (DMA)

I/O gestito da programma

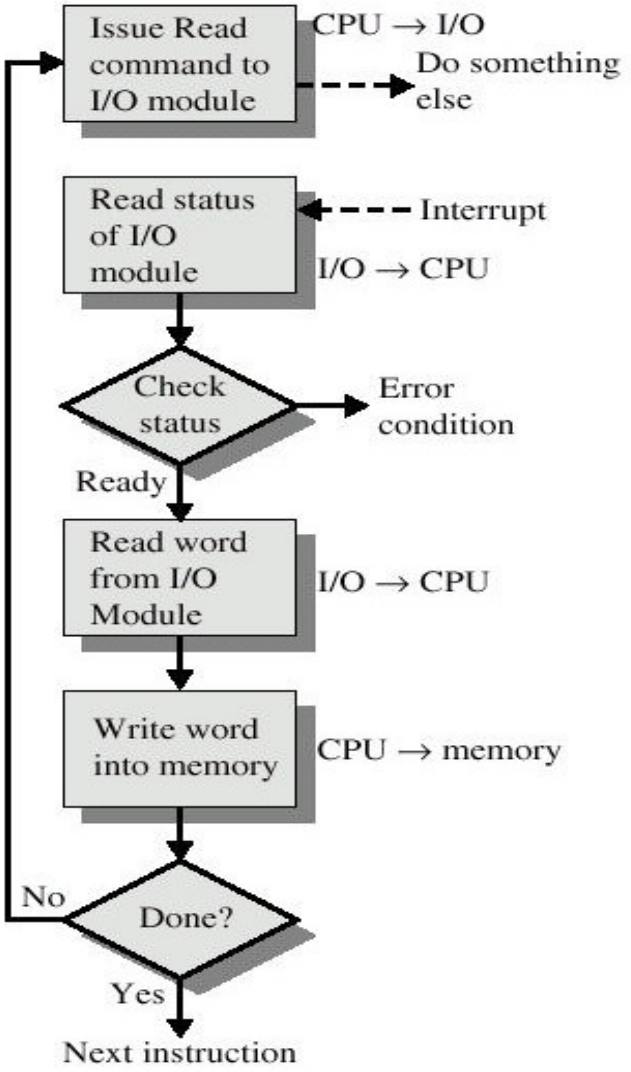
- I/O gestito dal controllo di programma
- La periferica ha un ruolo passivo
- La CPU si occupa sia del controllo sia del trasferimento dati
- La CPU predispone il controllore della periferica all'esecuzione dell'I/O
- La CPU si ferma e interroga il registro di stato della periferica in attesa che sia pronta (e.g., ready bit assume un determinato valore)
- Vantaggio: risposta veloce al ready bit
- Svantaggio: la CPU bloccata in stato di busy waiting



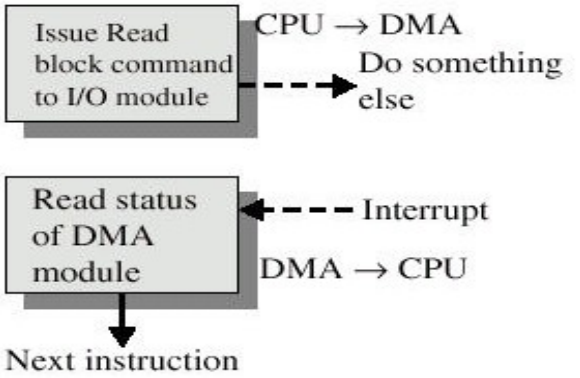
UNIVERSITÉ YORK UNIVERSITY



(a) Programmed I/O



(b) Interrupt-driven I/O



(c) Direct memory access

Prestazioni – Banda passante e latenza

- Misure che permettono di valutare l'efficienza della gestione delle operazioni di I/O
- Banda passante – rappresenta la quantita' dei dati che si puo' trasferire per unita' di tempo; rappresenta una misura di flusso
- Latenza – rappresenta il tempo che intercorre tra l'istante in cui una periferica e' pronta per il trasferimento e l'istante in cui il dato viente trasferito; e' una misura di tempo

Prestazioni – I/O gestito da programma

- Banda passante: alta perche' la CPU trasferira subito il dato (e.g., molti dati per l'unita' di tempo) e la gestione della periferica richiede poche istruzioni
- Latenza: minima in quanto la CPU noterá' subito lo stato della periferica (tempo massimo: un ciclo intero di busy waiting)

- Esempio: se si trasferiscono piu' dati (1KB) si ha un ciclo interno per il trasferimento di ogni byte e un ciclo esterno per il trasferimento di tutti i byte

Su latenza VS banda passante

1.2.1 Latency vs. Bandwidth

Loosely, latency is efficiency to the user, bandwidth is overall efficiency of the system.

More accurately, latency is defined as time to complete a specific operation. Bandwidth (also called throughput) is the number of units of work that can be completed over a specific time unit. These two measures are very different. Completing a specific piece of work quickly to ensure minimum latency can be at the cost of overall efficiency, as measured by bandwidth.

For example, a disk typically takes around 10ms to perform an access, most of which is time to move the head to the right place (*seek* time), and to wait for rotation of the disk (*rotational delay*). If a small amount of data is needed, the minimum latency is achieved if only that piece of data is read off the disk. However, if many small pieces of data are needed which happen to be close together on the disk, it is more efficient to fetch them all at once than to do a separate disk transaction for each, since less seek and rotational delay time is required in total. An individual transaction is slower (since the time to transfer a larger amount of data is added onto it), but the overall effect is more efficient use of the available bandwidth, or higher overall throughput.

Balancing latency and bandwidth requirements in general is a hard problem. Issues

Testo di Philip Machanick,
https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.6415&rep=rep1&type=pdf

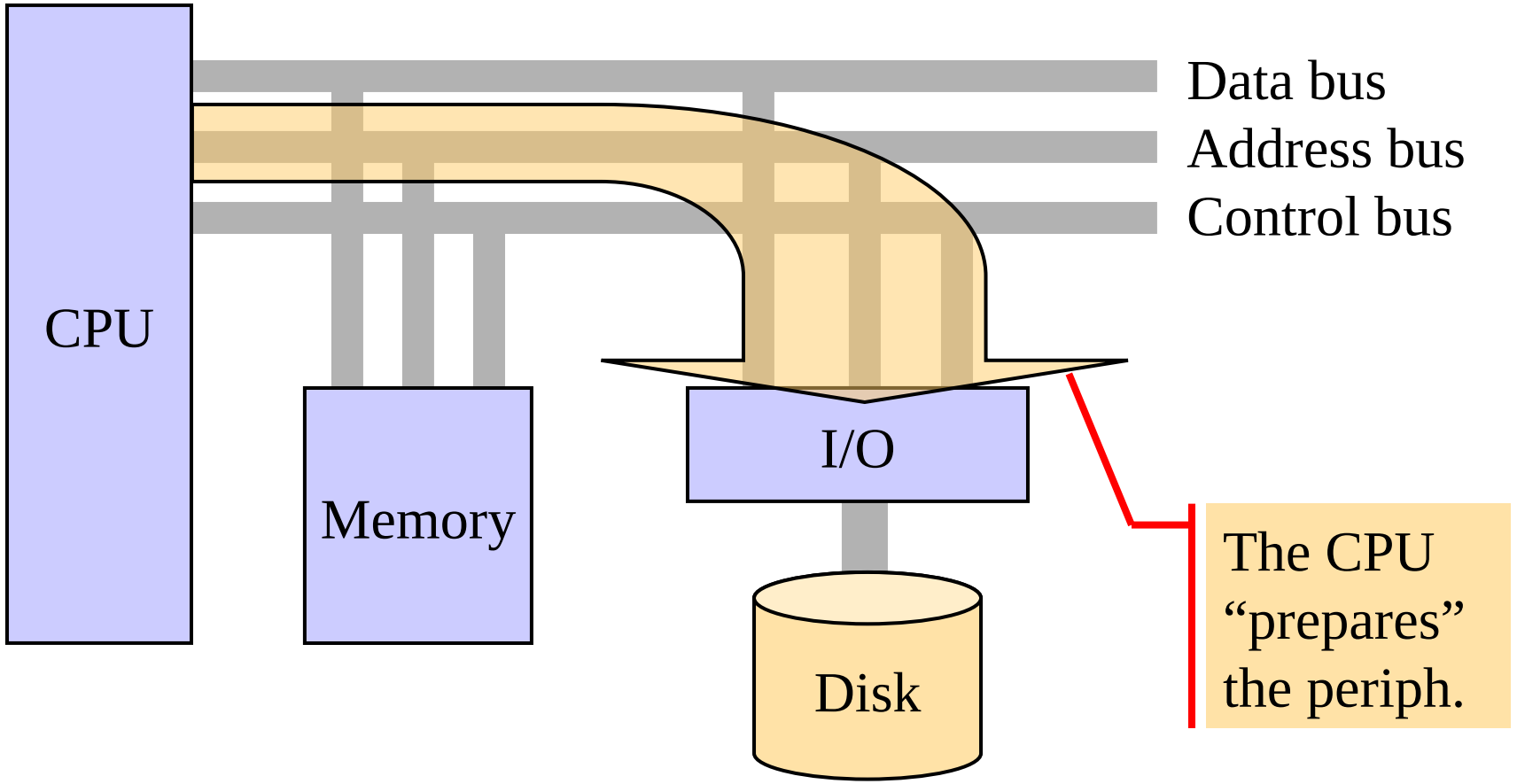
I/O guidato da interrupt

- Interrupt – evento asincrono che genera l'interruzione del normale funzionamento del processore
- La periferica segnala alla CPU di aver bisogno di attenzione mediante un segnala sul bus di controllo
- La periferica avvisa la CPU attivando un segnale interrupt request
- Quando il processore se ne accorge (in una fase di fetch) informa la periferica con un segnale interrupt acknowledge
- La CPU interrompe l'esecuzione del programma corrente (salvando il contesto dell'esecuzione del programma per poter riprendere la sua esecuzione) ed esegue la procedura di risposta all'interrupt
- Terminala l'esecuzione della procedura di interrupt, la CPU riprende l'esecuzione del programma interrotto
- Il programma utente continua la sua esecuzione

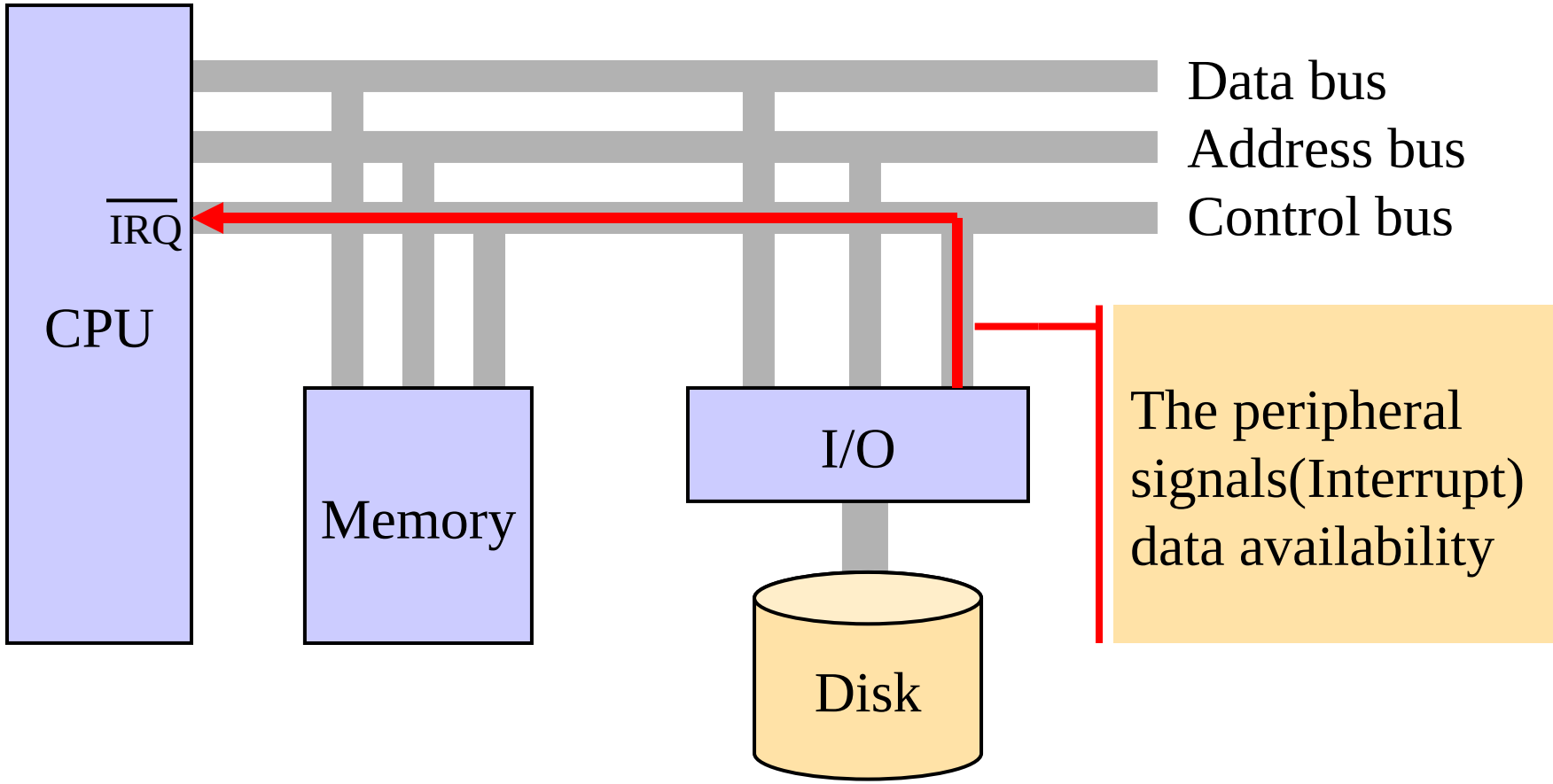
I/O guidato da interrupt

- Vantaggio: la CPU non fa piu' busy waiting come per I/O gestito da programma
- Svantaggio: la CPU deve comunque gestire le operazioni di trasfermineto

- Per evitare l'intervento della CPU nella fase di trasferimento dati e' stato introdotto il protocollo di trasferimento DMA – Direct Memory Access

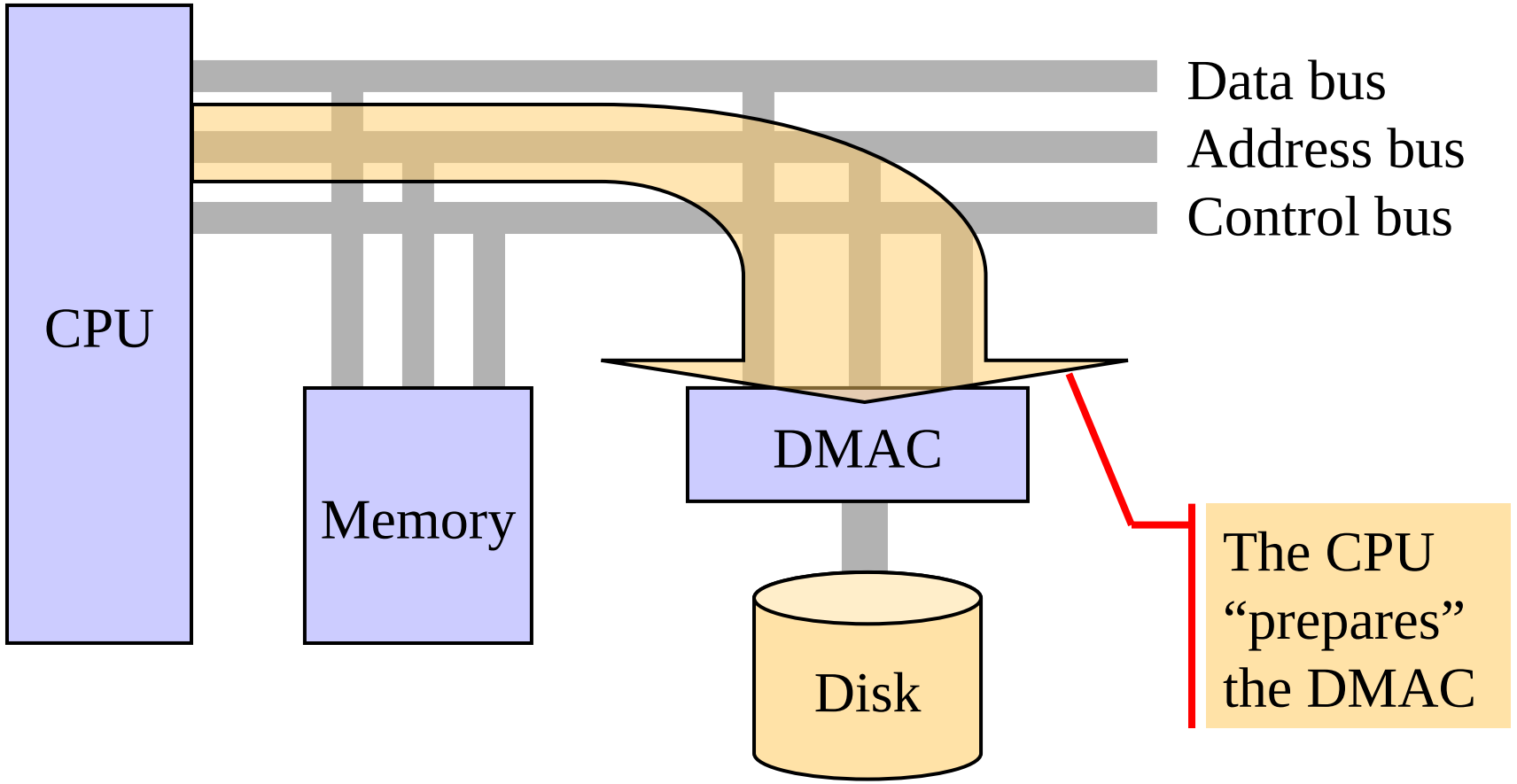


UNIVERSITÉ YORK UNIVERSITY

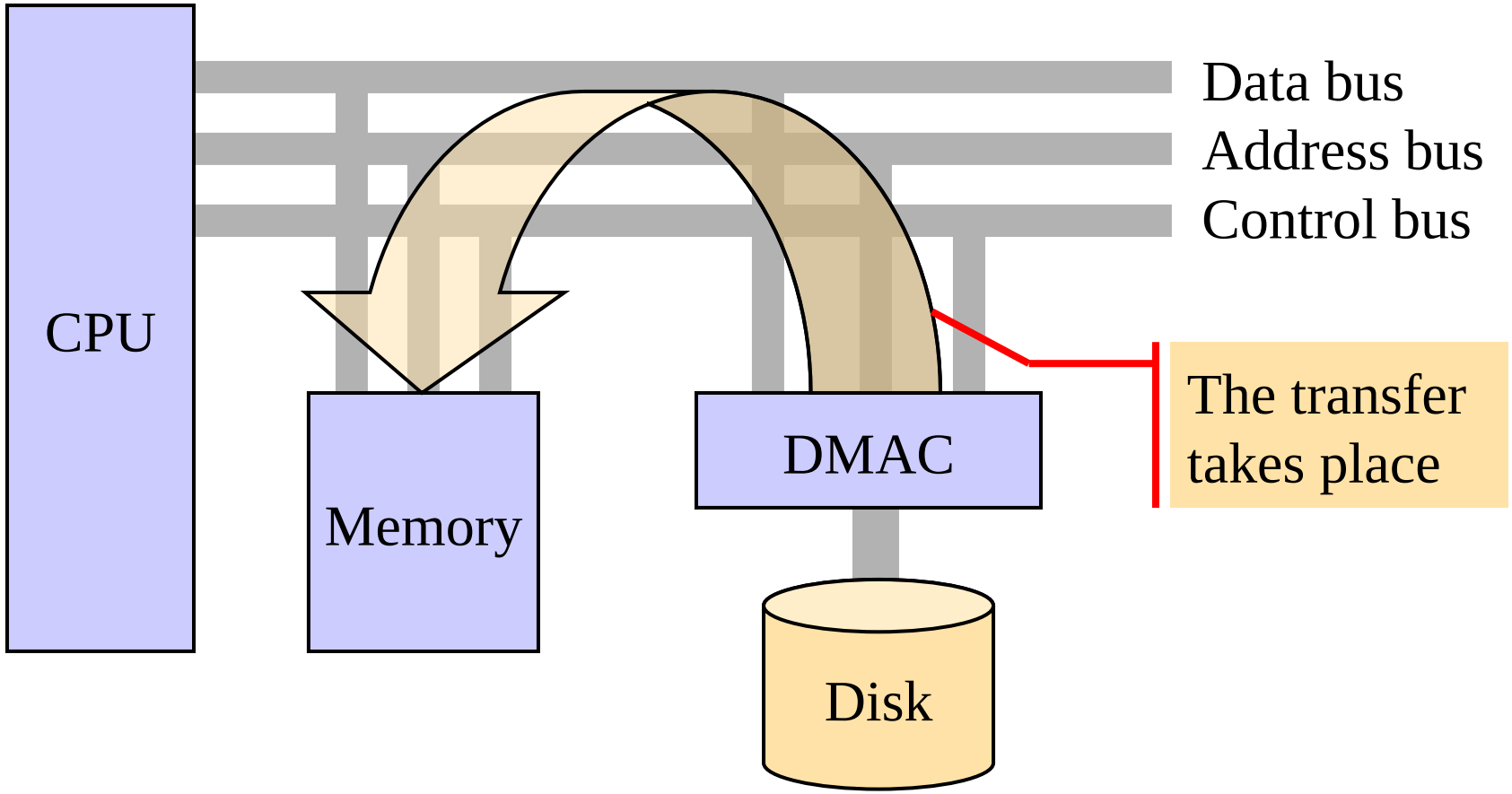

Data bus
Address bus
Control bus
IRQ
CPU
Memory
I/O
Disk
The peripheral signals(Interrupt) data availability

DMA – accesso diretto alla memoria

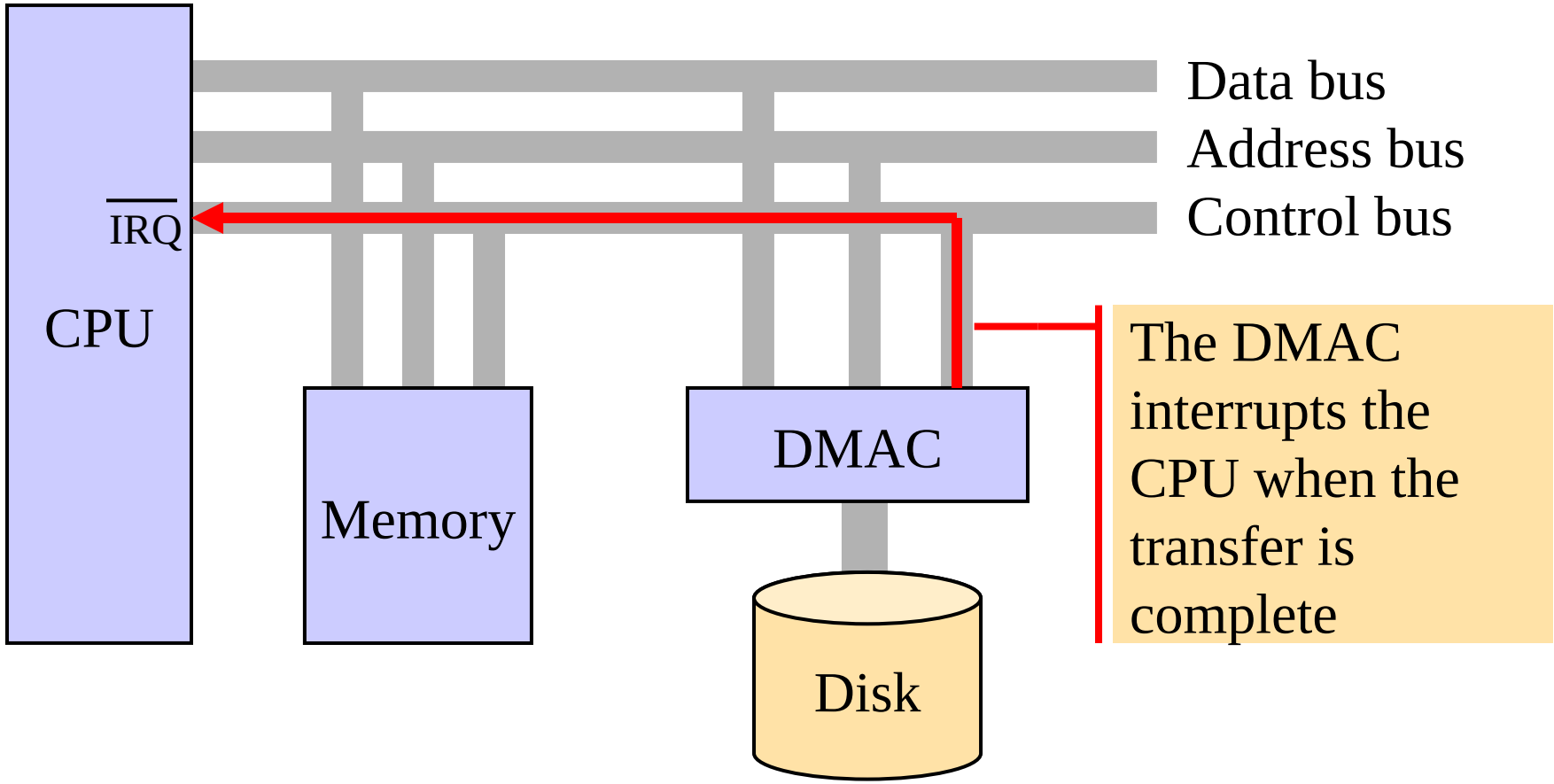
- Usato quando si trasferiscono velocemente grandi quantita' di dati
- Con DMA la periferica diventa autonoma nell'accessi alla memoria
- E' la periferica che gestisce i trasferimenti; la CPU non interviene nel trasferimento dei dati
- Necessita' 2 registri in piu' per ogni periferica oltre altre registro di stato e registro dei dati:
  - Un registro che indichi l'indirizzo di memoria da/dove trasferire i dati
  - Un registro che indichi la quantita' dei dati da trasferire
- Anche questi 2 registri aggiuntivi sono mappati in memoria
- Alla fine del trasferimento la periferica invia un interrupt alla CPU per segnalare il completamento del trasferimento

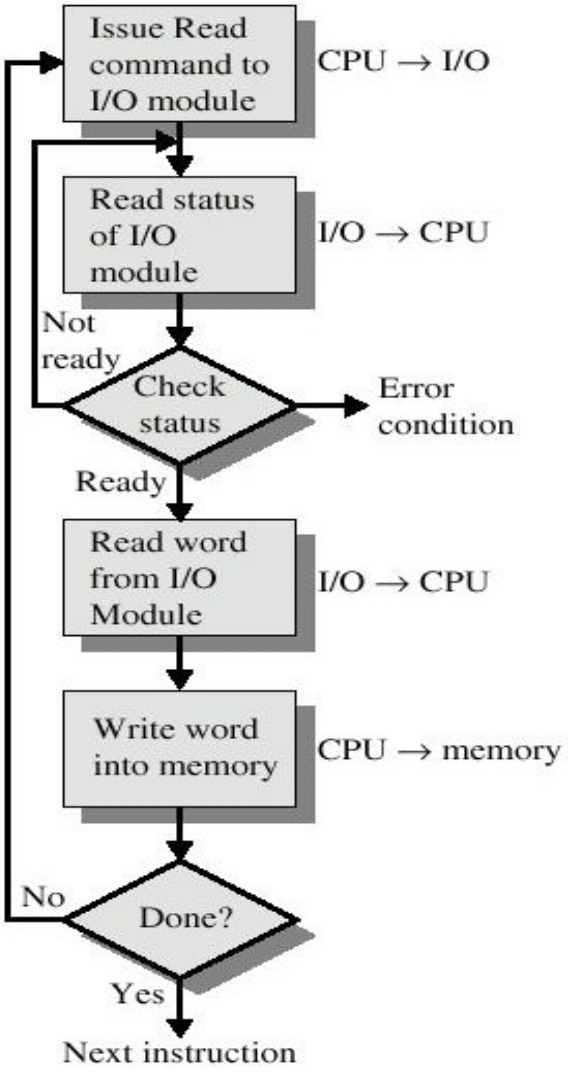


UNIVERSITÉ YORK UNIVERSITY

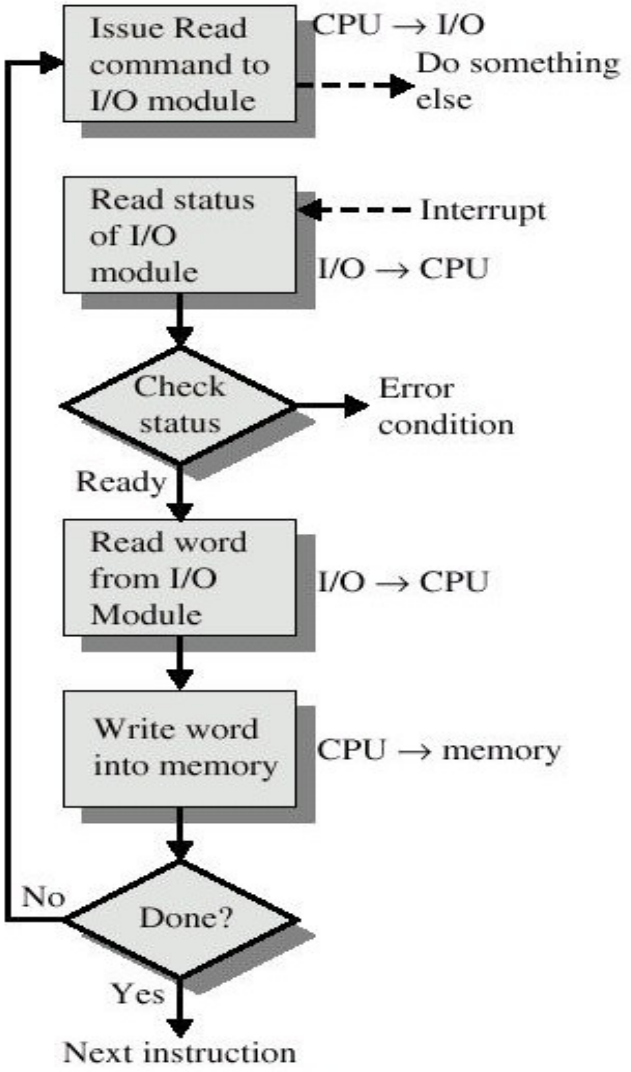


UNIVERSITÉ YORK UNIVERSITY

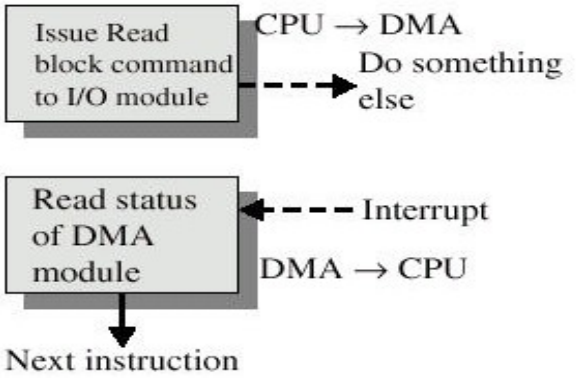

Data bus
Address bus
Control bus
IRQ
CPU
Memory
DMAC
Disk
The DMAC interrupts the CPU when the transfer is complete



(a) Programmed I/O



(b) Interrupt-driven I/O



(c) Direct memory access

Prestazioni – Banda passante e latenza

- Misure che permettono di valutare l'efficienza della gestione delle operazioni di I/O
- Banda passante – rappresenta la quantita' dei dati che si puo' trasferire per unita' di tempo; rappresenta una misura di flusso
- Latenza – rappresenta il tempo che intercorre tra l'istante in cui una periferica e' pronta per il trasferimento e l'istante in cui il dato viente trasferito; e' una misura di tempo

Prestazioni – I/O gestito da interrupt

- Banda passante: minor banda passante in quanto il trasferimento di ogni dato necessita' di piu' tempo
- Latenza: maggior latenza per la maggior quantita' di operazioni da eseguire

Prestazioni – DMA

- Banda passante: massima perche' la CPU non deve eseguire nessuna istruzione
- Latenza: minima dato che nessuna istruzione e' eseguita dalla CPU