Architettura degli Elaboratori
8 - Memorie, CPU e Bus

Ugo Dal Lago

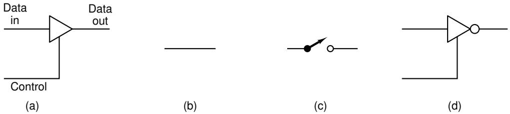
Dipartimento di Scienze dell'Informazione
Università degli Studi di Bologna

Anno Accademico 2007/2008

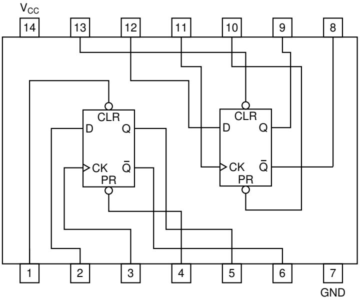
Sommario

La Memoria

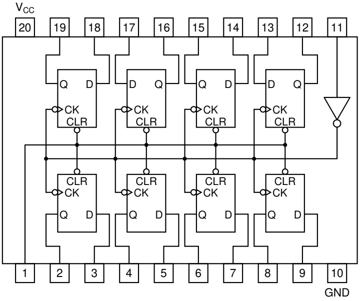
- È venuto il momento di parlare delle memorie.
- L'abbiamo già fatto nella prima parte del corso, ma ora abbiamo acquisito competenze specifiche relative al **livello logico-digitale**.
  - Ciò ci permetterà di capire meglio il funzionamento interno delle memorie.
- I blocchi di memoria che abbiamo incontrato sinora sono essenzialmente di tre tipi: **registri**, **cache** e **memoria centrale**.
  - Tutti sono realizzati tramite opportune reti sequenziali.
  - Gli ulitmi due hanno bisogno di ingressi per gli indirizzi, oltre che per i dati.
- Le porte logiche utilizzate sono quelle che conosciamo già. A volte, però, risulta utile avere a disposizione un **dispositivo a tre stati**:



Registri: Esempi

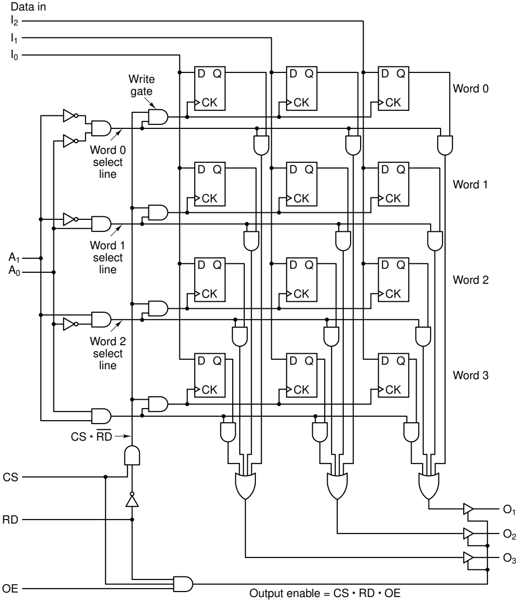


(a)



(b)

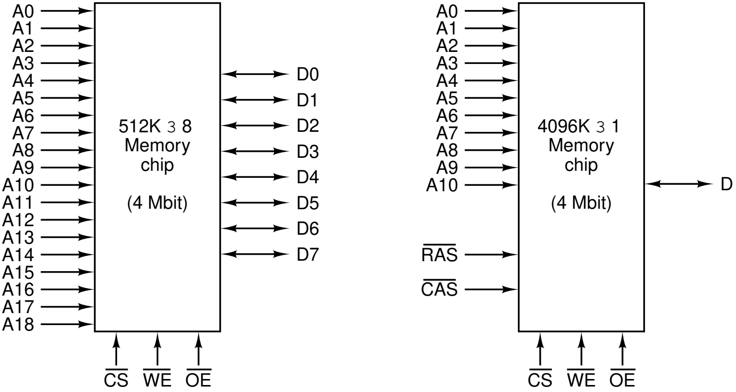
Memoria 4 × 3


Data in
I2
I1
I0
Write gate
Word 0 select line
Word 1 select line
Word 2 select line
Word 0
Word 1
Word 2
Word 3
A1
A0
CS • RD
CS
RD
OE
D Q
CK
O1
O2
O3
Output enable = CS • RD • OE

Chip di Memoria

- La struttura della memoria che abbiamo appena descritto facilita l'**ampliamento** del numero di parole e della dimensione delle parole.
  - Si potrebbe passare ad una memoria 8×3 oppure ad una memoria 4×8.
- La legge di Moore vale anche in questo contesto.
  - La dimensione dei chip di memoria disponibili sul mercato aumenta con un andamento esponenziale rispetto al tempo.
- Esistono moltissimi modi diversi per organizzare una memoria, ognuno con i suoi pregi e i suoi difetti.
- Se il numero di ingressi per gli indirizzi diventa troppo grande, si può utilizzare uno schema **a matrice**, in cui ogni parola è univocamente associata ad una coppia di indirizzi, anziché ad un singolo indirizzo.
  - In questo modo, però, è necessario più tempo per leggere (o scrivere) una parola: occorre comunicare al chip due sequenze binarie anziché una.

Chip di Memoria: Esempi



(a) (b)

Chip di Memoria Volatile: Tecnologia

- Le memorie che abbiamo considerato finora possono essere lette e scritte e si chiamano, per motivi storici, **RAM** (Random Access Memory).
- Esistono due tipi di RAM: le **RAM statiche** (o **SRAM**) e le **RAM dinamiche** (o **DRAM**).
  - Le RAM statiche sono costruite con circuiti simili ai flip-flop D.
  - Le RAM dinamiche sono costruite con celle costituite da un transistor e da un condensatore.
- Il vantaggio delle RAM dinamiche sta nella loro capacità, mentre lo svantaggio sta nella loro lentezza.
- Esistono vari tipi di RAM dinamica:
  - **FPM** (Fast Page Mode), organizzato a matrice di bit.
  - **EDO** (Extended Data Output), in cui più operazioni di lettura e scrittura possono essere sovrapposte.
  - **SDRAM** (Synchronous DRAM), una RAM ibrida, parte statica e parte dinamica.
  - **DDR** (Double Data Rate).

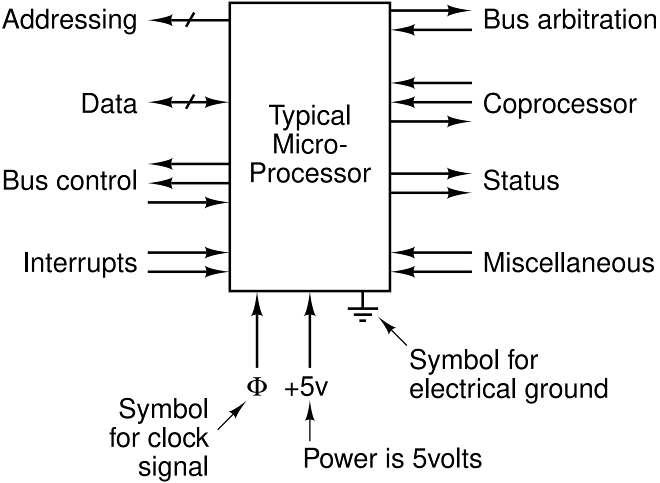
Chip di Memoria Non Volatile: Tecnologia

- In molte applicazioni, risulta utile avere a disposizione una memoria:
  - **Non volatile**, in modo tale che i dati non vengano persi quando è tolta l'alimentazione.
  - **A sola lettura**.
- Il contenuto delle memorie **ROM** (Read-Only Memory) non può essere modificato né cancellato.
  - I dati sono inseriti nella memoria durante la fabbricazione.
- Le memorie **PROM** (Programmable ROM), possono essere programmate, una sola volta.
  - Una delle tecniche che si utilizzano nelle PROM sono i fusibili.
- Le memorie **EPROM** (Erasable PROM), permettono anche la cancellazione, che però richiede un tempo relativamente lungo
- Nelle memorie **EEPROM** (Electronically EPROM) la cancellazione può essere effettuata in poco tempo.
  - La dimensione delle EEPROM è relativamente piccola.
- Le **memorie flash** ha tutti i vantaggi delle EEPROM, con l'unico svantaggio che risultano inutilizzabili dopo circa 100.000 cancellazioni.

Chip per la CPU

- Tutte le CPU moderne sono contenunte in un unico chip.
- I pin di una CPU possono essere divisi in tre categorie:
  - Pin per gli **indirizzi**.
  - Pin per i **dati**.
  - Pin di **controllo**.
- La comunicazione tra la CPU e la memoria avviene attraverso il bus e coinvolge i pin per gli indirizzi, i dati e i pin di controllo.
- Il numero m dei pin per gli indirizzi e il numero n dei pin per i dati sono due parametri fondamentali di una CPU.
- I pin di controllo possono essere (approssimativamente) raggruppati come segue:
  - **Controllo del Bus**.
  - **Interrupt**.
  - **Arbitraggio del Bus**.
  - **Comunicazione con il Coprocessore**.
  - **Stato**.
  - **Altro**.

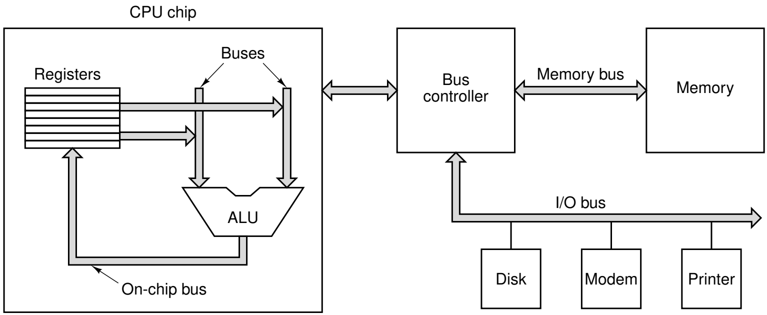
Tipico Chip per la CPU



Bus Del Calcolatore

- Un **bus** è un collegamento elettrico costituito da una sequenza di fili paralleli, che serve a far comunicare componenti diverse di un sistema di calcolo.
- In questa parte del corso ci occuperemo dei bus che connettono la CPU con la memoria e con i dispositivi di I/O.
  - Esistono anche bus interni alla CPU. Ne parleremo in una parte successiva del corso.
- Nei primi PC vi era un unico **bus di sistema**, mentre oggi c'è:
  - Almeno un bus per la comunicazione tra CPU e memoria.
  - Almeno un bus per la comunicazione tra CPU e dispositivi di I/O.
- Spesso ai bus esterni alla CPU possono essere connessi dispositivi diversi. Per questo occorre stabilire:
  - Un'insieme di regole di funzionamento, detto **protocollo del bus**.
  - Un'insieme di **specifiche meccaniche ed elettriche**.

Sistema di Calcolo con Più Bus



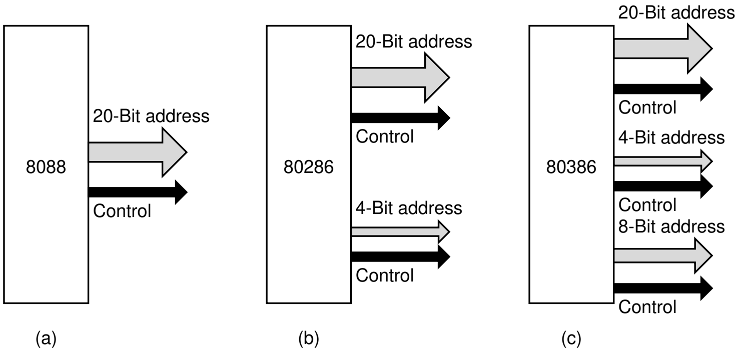
Bus Del Calcolatore

- Tra i componenti che sono connessi ad un bus, distinguiamo:
  - I componenti che sono attivi e possono iniziare un trasferimento dati (chiamati **master**).
  - I componenti che sono passivi e restando in attesa di una richiesta (chiamati **slave**).
- I dispositivi sono connessi al bus attraverso un **driver del bus**, un **ricevitore del bus** oppure un **trasmettitore-ricevitore del bus**, che dal punto di vista elettronico:
  - Possono essere dispositivi a tre stati.
  - Possono essere **collettori aperti**, il cui comportamento è simile ad una porta OR.
- I fili del bus hanno una semantica simile a quella dei pin della CPU.
  - Non è però detto che ci sia una corrispondenza uno-a-uno tra i fili del bus e i pin della CPU.

Ampiezza del Bus

- Un certo numero di fili tra quelli componenti un bus sarà dedicato agli **indirizzi**.
- Se un bus ha n linee d'indirizzi, la CPU potrà indirizzare 2^n locazioni di memoria.
- Occorre trovare un compromesso tra l'esigenza di indirizzare un grande numero di locazioni di memoria e l'esigenza di non occupare troppo spazio.
- Come le linee d'indirizzi, anche le linee di dati tendono a crescere di numero, aumentando quindi la **larghezza di banda** dei dati sul bus.
- Anche la velocità del bus può essere aumentata, ma non oltre un certo limite.
  - I dati su linee distinte viaggiano a velocità leggermente distinte (**disallineamento del bus**).
- Per aggirare il problema di bus troppo ampi, i progettisti optano a volte per un **bus multiplexato**: le stesse linee vengono utilizzate sia per gli indirizzi che per i dati.

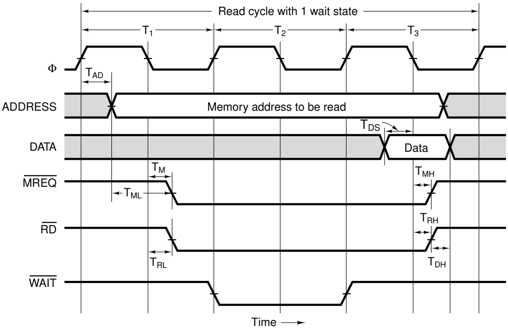
Crescita dell'Ampiezza del Bus



Temporizzazione del Bus

- I bus possono essere separati in due categorie distinte in base alla loro temporizzazione.
- Nei **bus sincroni**, esiste una linea pilotata da un clock, con frequenza generalmente compresa tra 5 e 100 MHz. Tutte le operazioni sul bus richiedono un numero intero di questi cicli.
- Nei **bus asincroni**, invece, non c'è un orologio principale: le operazioni sul bus possono avere una qualsiasi lunghezza.
- Per ragioni storiche e per rendere facile il progetto, i bus sincroni sono utilizzati molto più spesso dei bus asincroni.
- I bus asincroni, però, hanno un fondamentale vantaggio: permettono di sfruttare pienamente la velocità delle periferiche in gioco.
  - I bus sincroni, d'altro canto, devono sempre essere regolati alla velocità della periferica più lenta.

Bus Sincrono

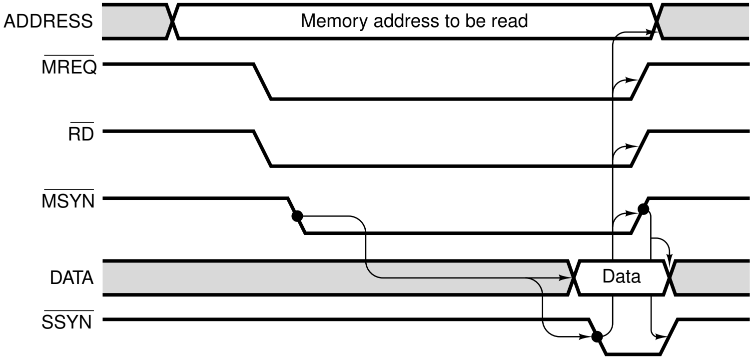


(a)

Symbol	Parameter	Min	Max	Unit
T_{AD}	Address output delay		11	nsec
T_{ML}	Address stable prior to $\overline{MREQ}$	6		nsec
T_M	$\overline{MREQ}$ delay from falling edge of Φ in T_1		8	nsec
T_{RL}	RD delay from falling edge of Φ in T_1		8	nsec
T_{DS}	Data setup time prior to falling edge of Φ	5		nsec
T_{MH}	$\overline{MREQ}$ delay from falling edge of Φ in T_3		8	nsec
T_{RH}	$\overline{RD}$ delay from falling edge of Φ in T_3		8	nsec
T_{DH}	Data hold time from negation of $\overline{RD}$	0		nsec

(b)

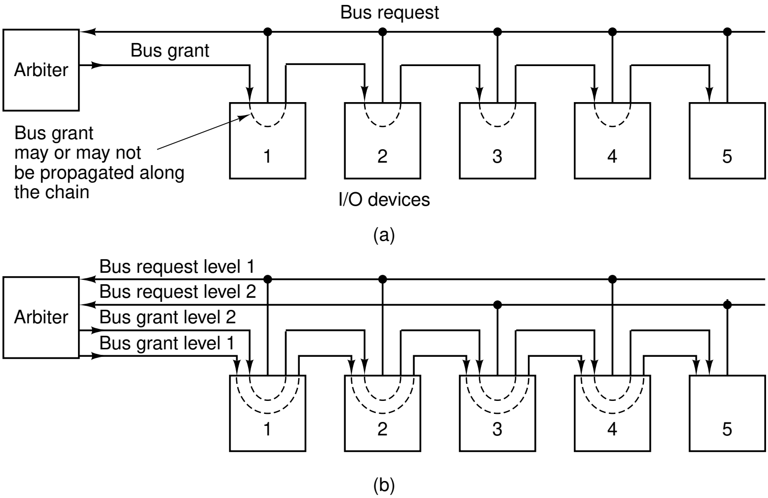
Bus Asincrono



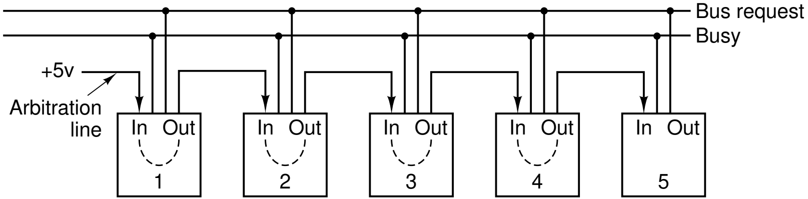
Arbitraggio del Bus

- Quando più dispositivi connessi allo stesso bus possono fungere da master, si rende necessaria qualche forma di **arbitraggio del bus**.
- L'arbitraggio del bus può essere **centralizzato**: un singolo arbitro determina chi sarà il prossimo master.
  - Esistono due linee: una per la richiesta del bus e una per la concessione del bus.
  - Questo schema è anche chiamato **daisy chaining** e assegna implicitamente una priorità più alta ai dispositivi fisicamente più vicini all'arbitro.
  - Per ovviare a questa rigidità, si possono utilizzare schemi con priorità.
- L'arbitraggio del bus può però anche essere **decentralizzato**:
  - Per esempio, potrebbe esserci una linea di richiesta per ogni dispositivo, ciascuna con la propria priorità.
  - Alternativamente, si potrebbe avere uno schema con una singola linea di richiesta, una linea asserita dal master corrente e una terza linea per la gestione dell'arbitraggio.

Arbitraggio Centralizzato



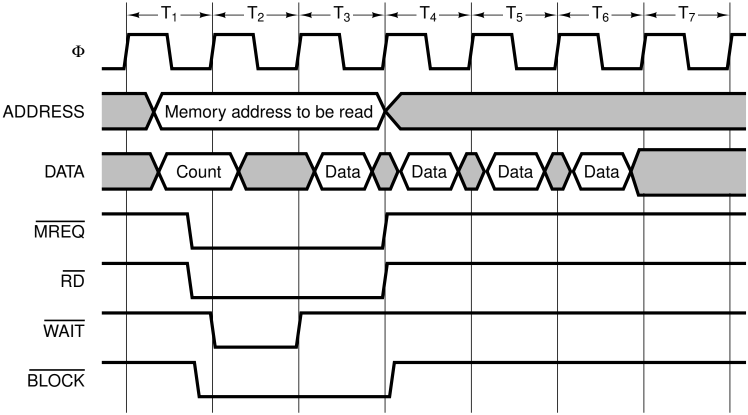
Arbitraggio Decentralizzato



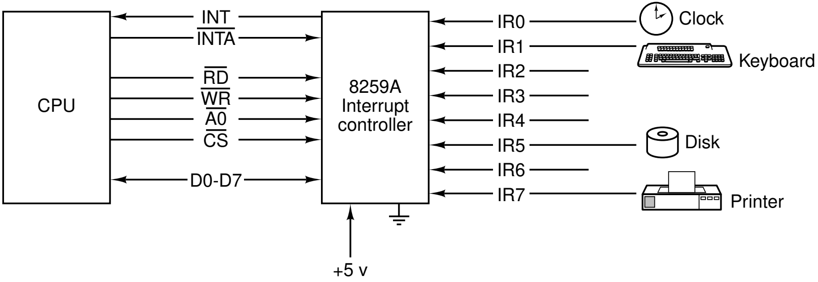
Operazioni del Bus

- Oltre ai trasferimenti di dati tra un master e uno slave, esistono altri tipi di eventi che riguardano i bus esterni.
- Prima di tutto, si possono effettuare trasferimenti di **blocchi di dati**, anziché di singole parole di memoria.
  - Utili soprattutto in presenza di una cache.
- Può esistere uno speciale ciclo di bus che consente ad una CPU di leggere una parola dalla memoria, analizzarla e riscriverla in memoria.
  - Utile per gestire l'accesso concorrente alle strutture dati.
- Infine, attraverso il bus viaggiano anche i segnali di **interrupt**.
  - Tra l'altro, servono a comunicare alla CPU la fine di un'operazione di I/O.
  - Diverse periferiche potrebbero aver bisogno di interrompere la CPU contemporaneamente. Serve quindi un meccanismo di arbitraggio, gestito da un **controllore degli interrupt** (per esempio il chip Intel 8259A).

Trasferimento di un Blocco di Dati



Controllore degli Interrupt



Pentium 4

- Retrocompatibile con tutti i modelli prececedenti della famiglia: 8088, 80286, 80386, etc.
- Fino a 55 milioni di transistor, fino a $3,2$ GHz, larghezza di linea di $0,09$ micron.
- Può scambiare dati con la memoria in unità da 64 bit, anche se gli indirizzi e i registri sono a 32 bit.
- Due ALU che lavorano in parallelo e che permettono di eseguire due programmi molto velocemente.
- Almeno due livelli di cache.
  - Possibili problemi di incoerenza, risolti tramite il cosiddetto **snooping**.
- Due bus: uno per la memoria e uno (PCI) per i dispositivi di I/O.
- Problema: dissipazione del calore.
- Soluzione: cinque stati possibili, che variano dalla piena esecuzione al “sonno profondo”.

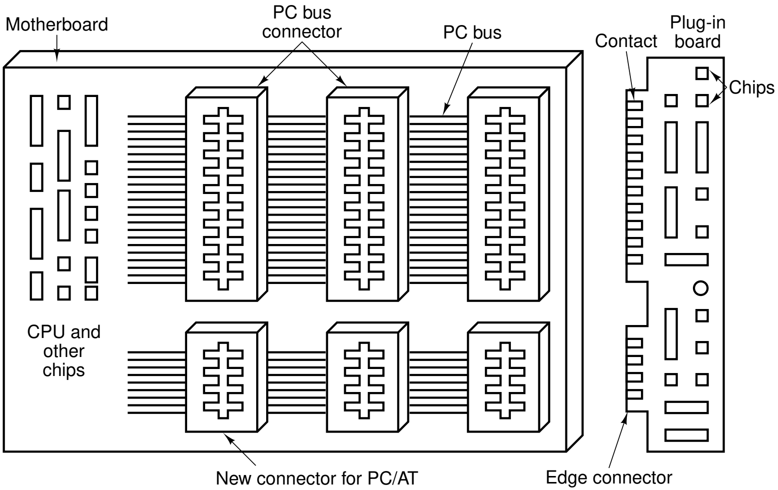
UltraSPARC III

- È un processore a 64 bit.
- 1369 pin, disposti in una matriche quadrata di $37 \times 37 - 1$ pin.
- Inizialmente: 29 milioni di transistor, 600 MHz e 29 milioni di transistor.
  - Poi: aumentato.
- Due cache interne alla CPU.
  - 32 KB per le istruzioni e 64 KB per i dati.
- Una cache di secondo livello, esterna al chip.
- 43 linee di bus per gli indirizzi e 128 linee di bus per i dati.
- Sun ha sviluppato **UPA**, che consente alla CPU di comunicare con memorie di tipi diversi.

Bus ISA

- Il bus dei **primi PC** aveva 62 linee, di cui 20 per gli indirizzi e 8 per i dati.
  - Ogni dispositivo era collegato alla scheda madre del PC con dei connettori distanziati 2 cm uno dall'altro.
- Quando IBM introdusse il **PC/AT**, aggiunse 36 linee, rimanendo comunque compatibile con i primi PC.
- Con il successore del PC/AT, chiamato **PS/2**, IBM decise di progettare un bus completamente nuovo.
  - Il mercato continuò però ad utilizzare un bus simile al bus del PC/AT, ribattezzandolo **ISA**.
- Successivamente il bus ISA fu esteso a 32 bit, diventando il bus **EISA**.

Bus PC/AT

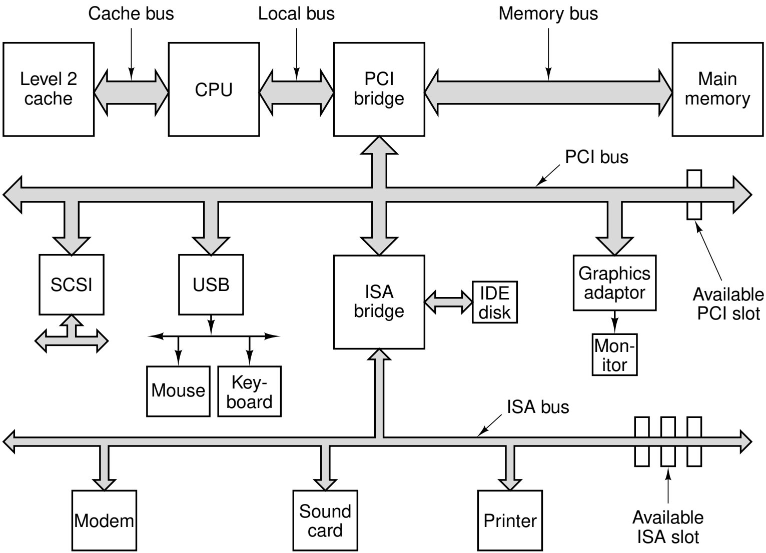


Motherboard
PC bus connector
PC bus
Contact
Plug-in board
Chips
CPU and other chips
New connector for PC/AT
Edge connector

Bus PCI

- I bus ISA, EISA sono troppo lenti per tutte le applicazioni che richiedono una certa larghezza di banda.
- Attorno al 1990, Intel progettò un nuovo bus con una larghezza di banda molto più ampia e lo chiamò **PCI** (Peripheral Component Interconnect).
  - La larghezza di banda del bus PCI originario era di 133 MB al secondo, contro una larghezza di banda di $16,7$ MB al secondo per il bus ISA.
  - Nelle versioni successive, la larghezza di banda per il bus PCI aumenta ulteriormente.
- Nei PC moderni, i bus in gioco sono in genere molti, in modo tale da bilanciare esigenze diverse.
- Il bus PCI è sincrono.
- Le linee degli indirizzi e dei dati sono multiplexate.

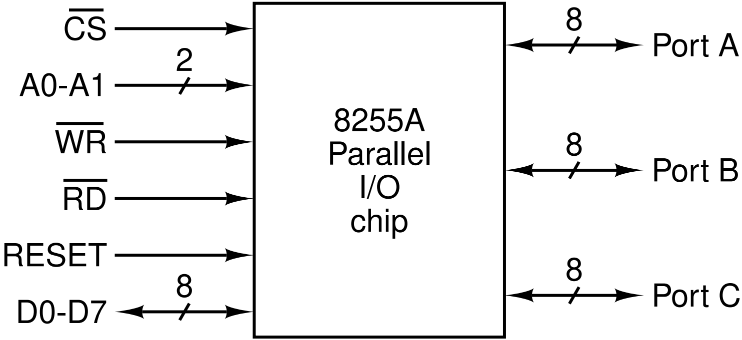
Architettura di uno dei primi Pentium



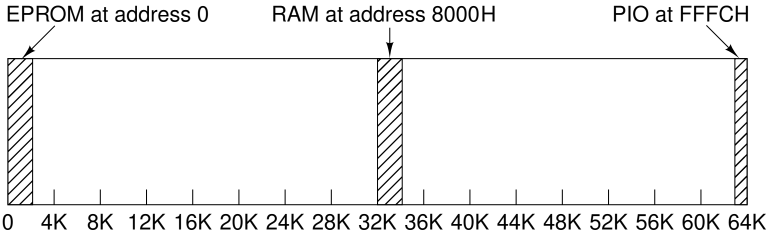
Il Chip 8255A

- Il Chip **8255A** è prodotto da Intel.
- Il suo compito è quello di interfacciare il calcolatore con dispositivi esterni.
- La CPU può comunicare con il dispositivo modificando o leggendo il contenuto di tre registri (a 8 bit) A, B e C interni al chip 8255A.
  - La CPU può accedere ai registri attraverso 8 linee di I/O, modificando opportunamente alcuni ingressi di controllo.
- I registri A, B e C possono anche essere scritti dai dispositivi esterni. In questo modo la CPU può ricevere dati in input dal dispositivo.
- Esistono due modi in cui la CPU si può interfacciare con il chip:
  - Considerando il chip come un vero e proprio I/O. In questo caso occorrerà una linea, nel bus, che indica che ci si sta riferendo ad un dispositivo di I/O.
  - Usando l'**I/O mappato in memoria**, assegnando al chip quattro indirizzi tra quelli appartenenti allo spazio di indirizzamento della memoria.

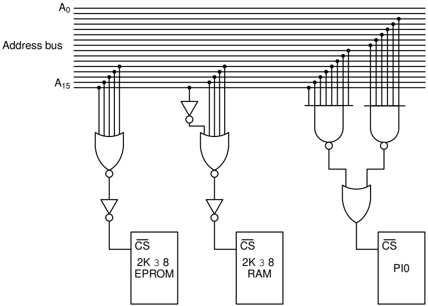
II Chip 8255A



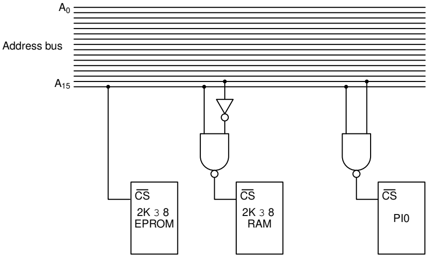
II Chip 8255A



II Chip 8255A



(a)



(b)