Architettura degli Elaboratori

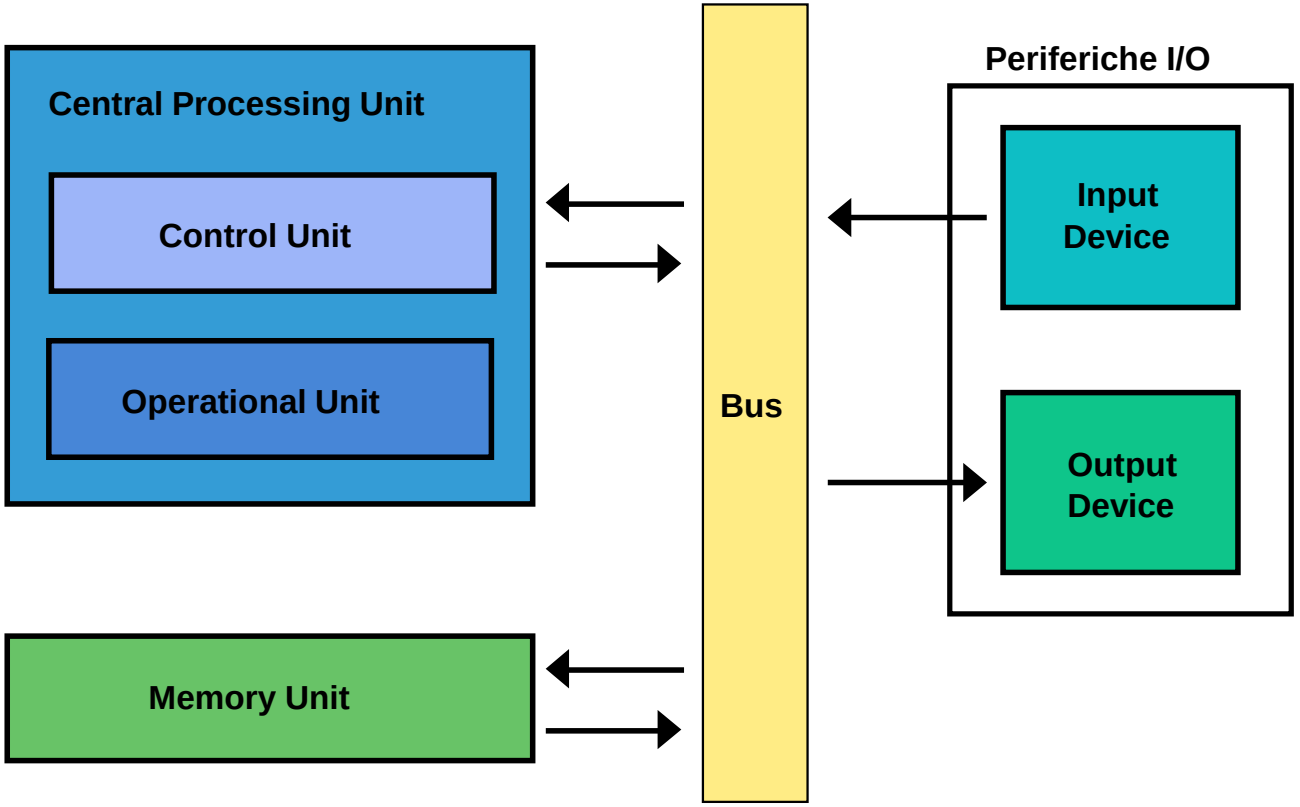
Lezione 14-15 – Memorie

Giuseppe Cota
Dipartimento di Scienze Matematiche Fisiche e Informatiche
Università degli Studi di Parma

Indice

- Gerarchia della memoria
- Spazio degli indirizzi

Architettura di von Neumann



Memoria

- I dati e i programmi all'interno del calcolatore devono essere conservati in memoria.
- Esistono più tipi di memoria:
  - Dischi rigidi
  - Memoria centrale (comunemente chiamata RAM)
  - Cache
  - Registri
  - …

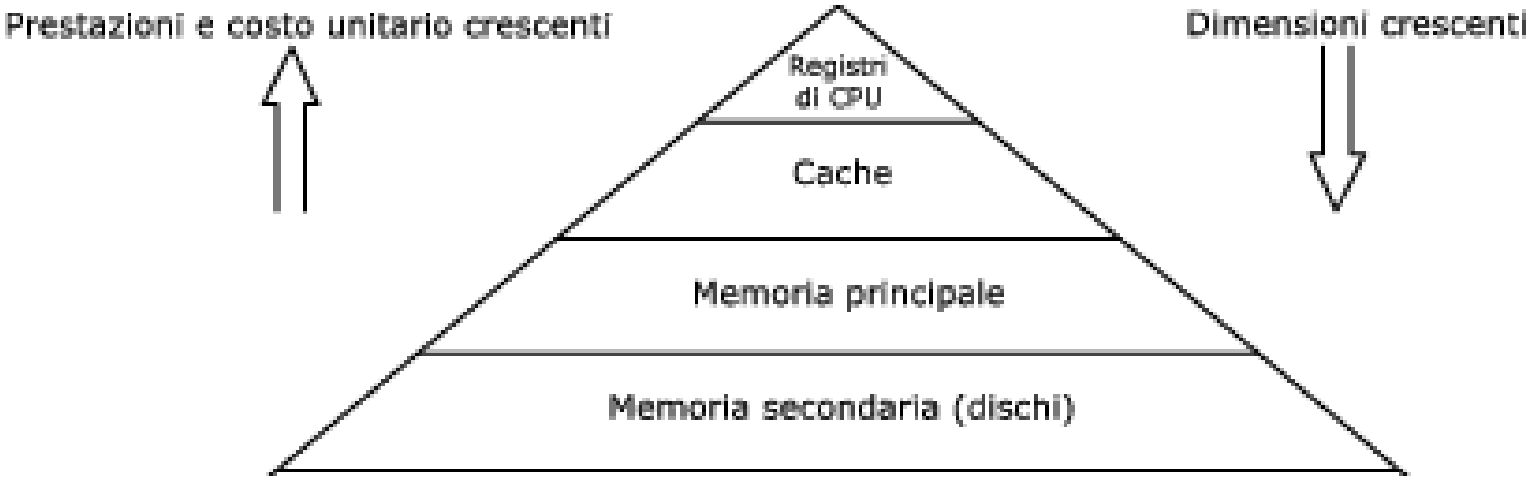
Caratteristiche di una memoria

- Una memoria è caratterizzata da diversi parametri:
  - **Dimensione o capacità**: quanti dati riesce a memorizzare, solitamente misurata in multipli di byte.
  - **Velocità o tempo di accesso**: intervallo di tempo tra la richiesta del dato e il momento in cui è disponibile
  - **Potenza o consumo**: potenza media assorbita (dalle memorie elettroniche)
  - **Costo per bit**: costo materiale per un bit, non è un costo fisso dipende anche dalle dimensioni della memoria.

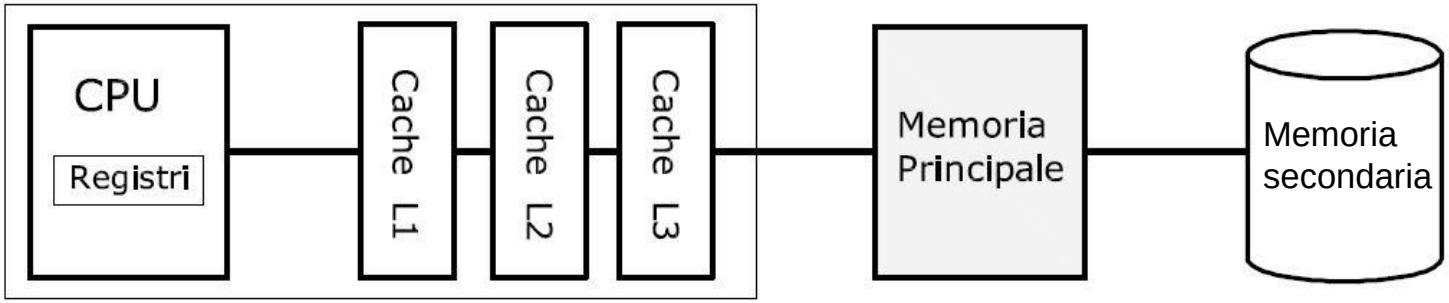
Gerarchia della memoria

- Idealmente un calcolatore dovrebbe avere quanta più memoria possibile, ad alta velocità, basso consumo e minimo costo.
- Tuttavia non è possibile avere un'unica memoria grande e con un alto rapporto prestazioni/costo → **bisogna raggiungere un compromesso**.
- È necessario strutturare il sistema di memoria con memorie di diverso tipo strutturate in maniera gerarchica.
- **Gerarchia della memoria**:
  - ai livelli più alti: memorie con prestazioni e costo elevati e dimensioni piccole.
  - ai livelli più bassi: memorie con prestazioni e costo bassi e dimensioni grandi.

Gerarchia di memoria



Gerarchia di memoria

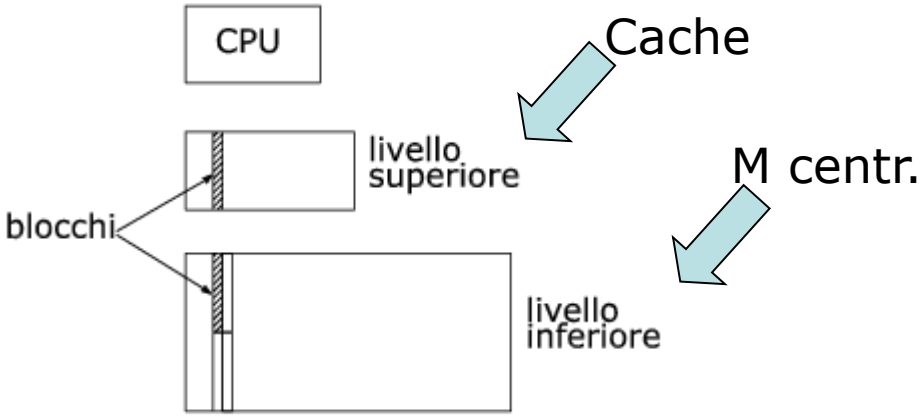


- **Registri della CPU**: centinaia di byte
- **Cache di livello 1 (L1 cache)**: ~decine di KB
- **Cache di livello 2 (L2 cache)**: ~512 KB
- **Cache di livello 3 (L3 cache)**: ~4 MB - 128MB
- **Memoria Centrale (RAM):** ~4 GB – 32GB
- **Memoria secondaria interna (HDD o SSD)**: centinaia di GB
- **Memoria secondaria esterna (dischi ottici, penne USB)**

Principi di località spaziale e temporale

- **Principio di località spaziale**: se un programma, nel corso della sua esecuzione, fa riferimento ad una particolare cella di memoria, è molto probabile che, *nell'immediato futuro*, faccia riferimento a celle vicine a essa.
- **Principio di località temporale**: se un programma, nel corso della sua esecuzione, fa riferimento ad una particolare cella di memoria, è molto probabile che, *nell'immediato futuro*, faccia riferimento alla stessa cella.

Terminologia



- **Blocco**: la minima unità di informazione che può essere trasferita tra due livelli adiacenti della gerarchia.
  - Il trasferimento dei dati avviene tra due blocchi adiacenti.
- **Hit (successo):** l'informazione richiesta è presente nel livello acceduto
- **Miss (fallimento)**: l'informazione richiesta non è presente nel livello acceduto
  - Bisogna accedere al livello inferiore della gerarchia per recuperare il blocco contenente l'informazione richiesta.
  - Alla fine quando l'informazione viene trovata, il blocco che la conteneva viene trasferito al livello di memoria dove è avvenuto il miss.

Prestazioni della gerarchia di memoria

- **Tasso di hit** h: rapporto tra il numero di hit di un livello di memoria e il numero totale di accessi.
- **Tempo di hit** t_h: tempo di accesso di un livello di memoria, compreso il tempo per determinare se si verifica un hit o un miss.
- **Tasso di miss** m: rapporto tra il numero di miss di un livello di memoria e il numero totale di accessi.

$$m = 1 - h$$

- **Penalità o tempo di miss** t_m: tempo per trasferire il blocco dal livello più basso (ignoriamo per semplicità il caso di miss a più livelli).
- **Tempo medio di accesso alla memoria** $\bar{t}$:

$$\bar{t} = t_h + m \cdot t_m = t_h + (1 - h)t_m$$

- Se siamo al livello più alto della gerarchia, per avere buone prestazioni h deve essere il più vicino possibile a 1.

Tipologie di memoria
Funzionalità

- **Memoria volatile**: la memorizzazione richiede l'alimentazione elettrica. Quando una memoria volatile viene spenta, tutto il contenuto viene cancellato.
- **Memoria persistente**: permette la persistenza dei dati per più tempo. Se una memoria persistente viene spenta, il contenuto non viene cancellato, ma *persiste*.
- **Memoria di sola lettura**: Read Only Memory (ROM)
- **Memoria di lettura/scrittura**

Tipologie di memoria
Tecnologia

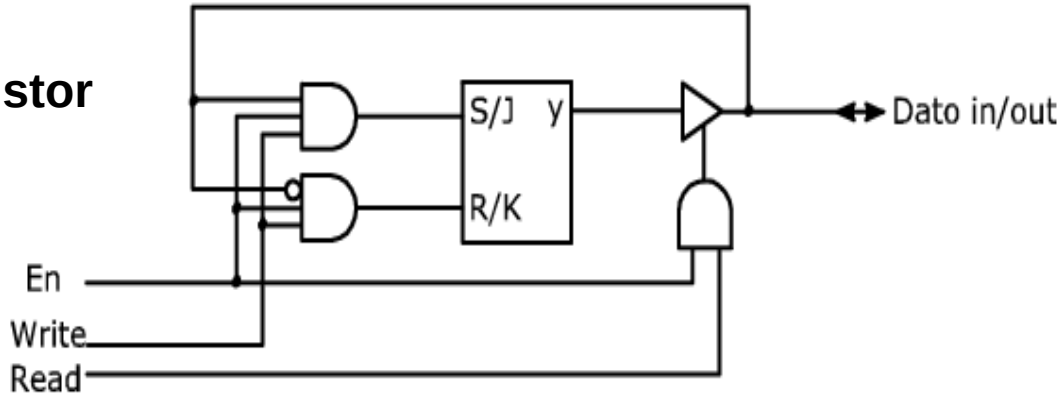
- **Memoria elettronica:** si usano componenti elettronici (ROM, RAM, flash).
- **Memoria magnetica**: memorie di massa come dischi rigidi e nastri magnetici. Principale caratteristica è la persistenza.
- **Memoria ottica**: dischi ottici per memorizzazione di lungo termine (CD, DVD, Blu-ray)

Tipologie di memoria
Modalità di accesso

- **Memoria ad accesso casuale**, in inglese **Random Access Memory (RAM)**: il tempo di accesso ad una cella è indipendente dalla sua posizione. La sigla RAM viene impropriamente utilizzata per indicare la memoria centrale del calcolatore (ma anche le memorie flash e la ROM sono ad accesso casuale).
  - A volte in italiano si preferisce la locuzione *memorie ad accesso diretto*.
- **Memoria ad accesso sequenziale**: l'accesso ad una generica cella avviene scorrendo sequenzialmente la memoria stessa.
  - Esempio: nastri magnetici.
- **Memoria ad accesso semicasuale**: dato l'indirizzo di una cella, con un accesso diretto si accede ad un blocco di celle all'interno del quale la singola cella viene individuata con una ricerca sequenziale.
  - Esempio: dischi rigidi
- **Memoria ad accesso per contenuto**: In lettura, la memoria risponde restituendo l'indirizzo della posizione che contiene il dato.

Memorie RAM

- **SRAM: Static RAM (RAM Statiche)**
  - **un FF per bit**
    - **Costituito da 6 transistor**
  - Alto Consumo
  - Alto Costo
  - Alta velocità
  - **Usate per le cache**



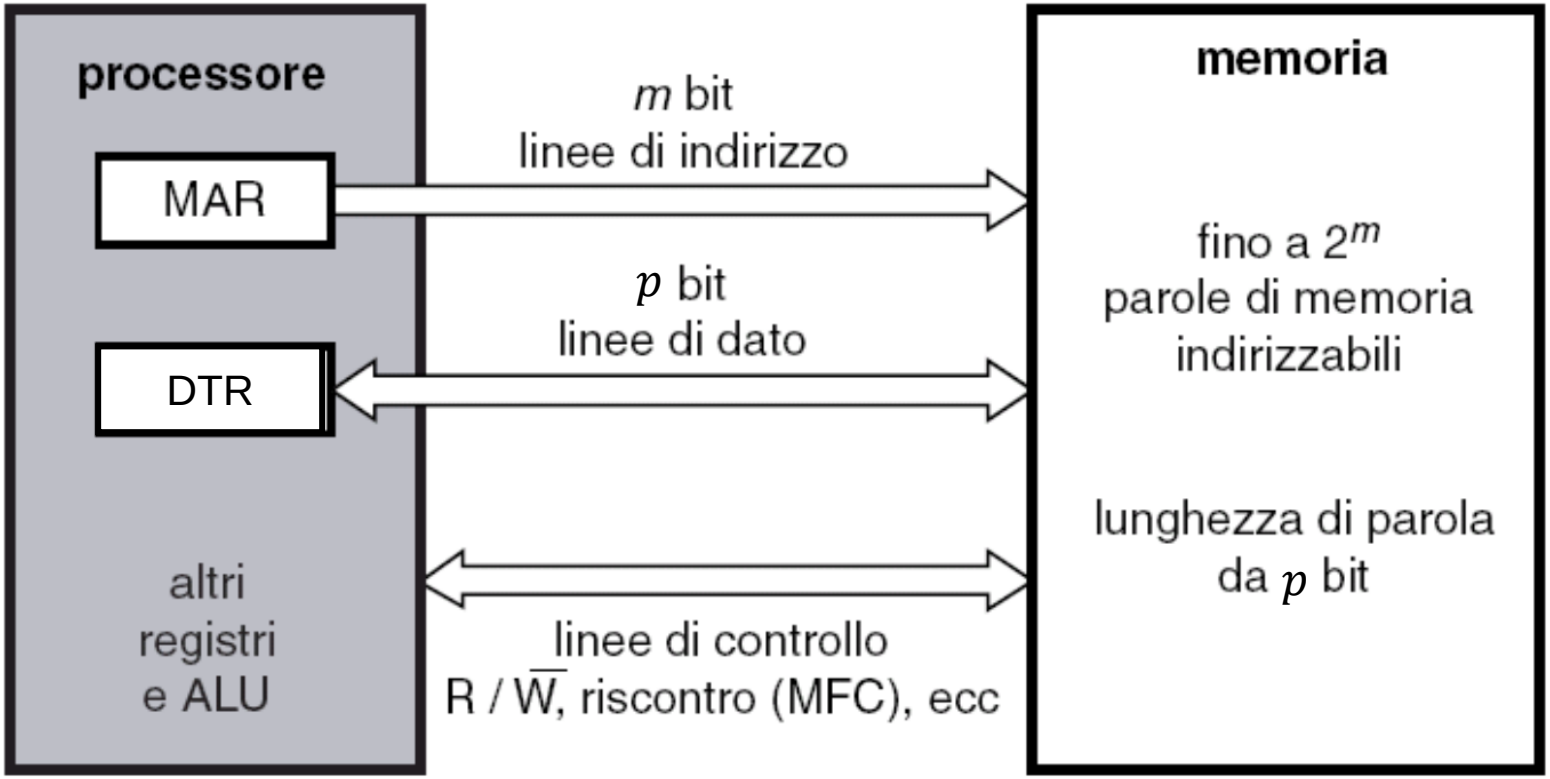
- **DRAM: Dynamic RAM (RAM Dinamiche)**
  - **un transistor e un condensatore per bit**
  - Basso Consumo
  - Basso costo
  - Bassa velocità
  - **Usate per le "RAM"**

SRAM e DDR

- SDRAM: Synchronous DRAM
  - Sono DRAM sincrone, introdotte nel 1996
- I comandi agganciati a un clock
- DDR (Double Data Rate SDRAM) operano su ambedue i fronti del clock
  - Correntemente DDR4, dal 2021 DDR5

Spazio degli indirizzi

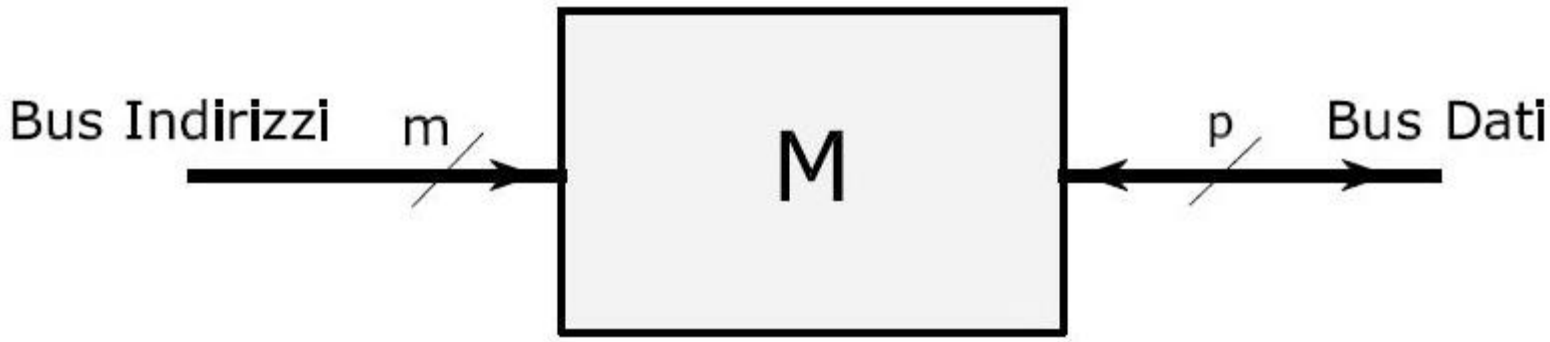
Connessione tra memoria e processore



- **Per leggere** è necessario fornire l'indirizzo (A_0 .. A_{m-1}) e il comando di lettura
- **Per scrivere** sono necessari indirizzo, dato (D_0 .. D_{d-1}) e il comando di scrittura

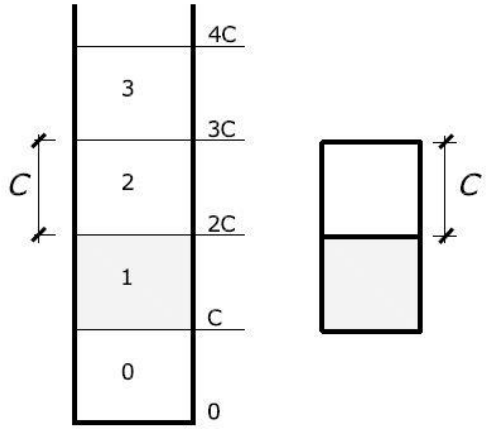
Spazio di indirizzamento

- Una memoria è costituita da celle di p bit, pari al grado di parallelismo del bus dei dati.
- Ad ogni cella è assegnato un indirizzo univoco.
  - Nella pratica corrente è convenzione assegnare degli indirizzi ai byte.
- Per poter leggere/scrivere una cella è necessario presentare il suo indirizzo sul bus degli indirizzi.
- Con m linee di indirizzo si indirizzano 2^m posizioni
  - Si dice che la memoria ha un'estensione pari a $M = 2^m$

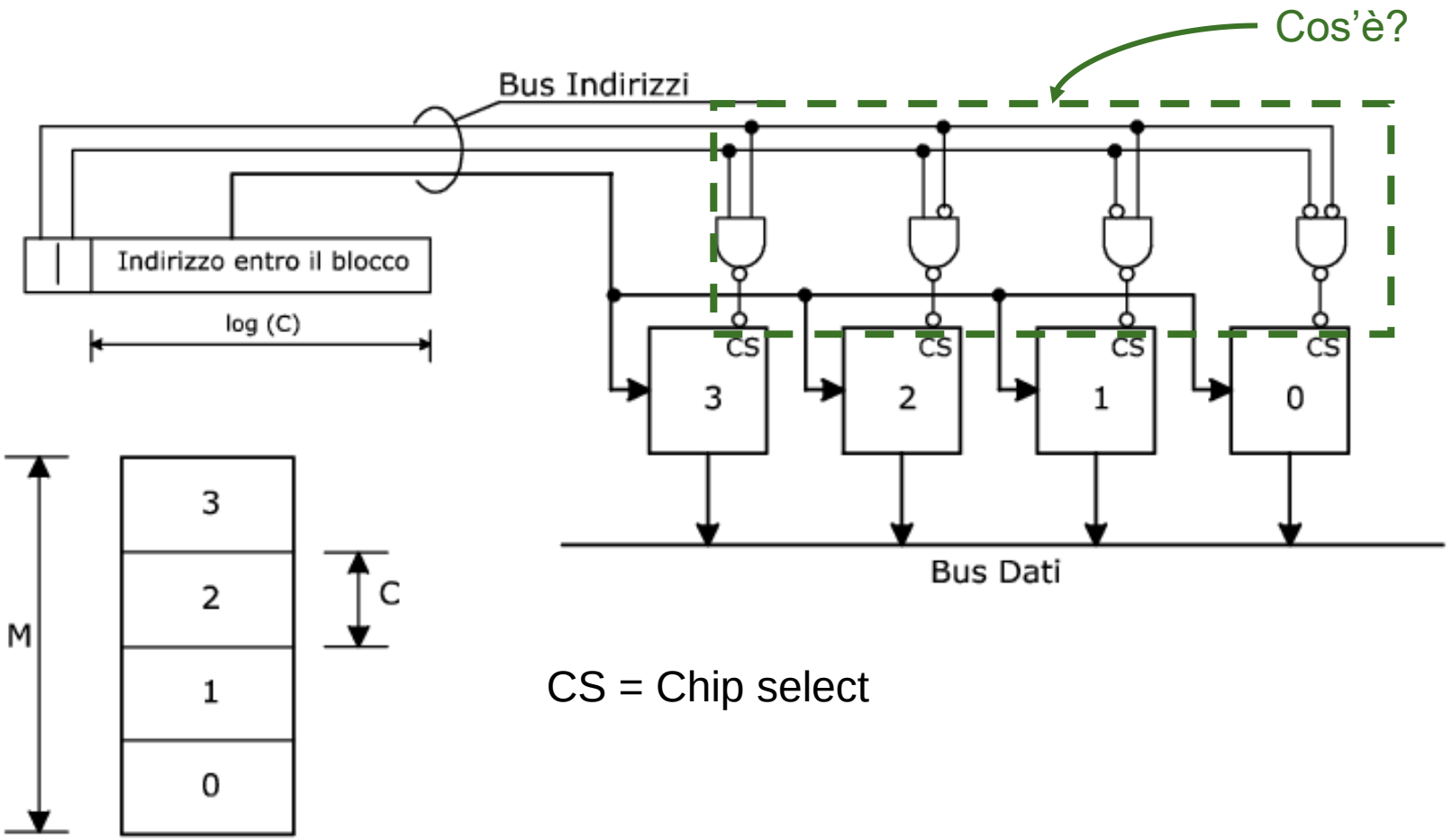


Spazio di indirizzamento
Blocchi di memoria

- Il complesso delle possibili posizioni di memoria costituisce lo spazio degli indirizzi
  - Di norma non tutto lo spazio di indirizzamento viene utilizzato.
- Inoltre la memoria potrebbe essere divisa in blocchi di memoria di una certa dimensione $C = 2^c$ (quindi con $c = \log_2 C$ linee di indirizzo).
- Potrei costruire una memoria con n circuiti integrati con estensione C
- Un indirizzo di memoria può essere visto come composto da due parti: il numero (indirizzo) del blocco e l'indirizzo della cella all'interno del blocco.



Spazio di indirizzamento
Blocchi di memoria

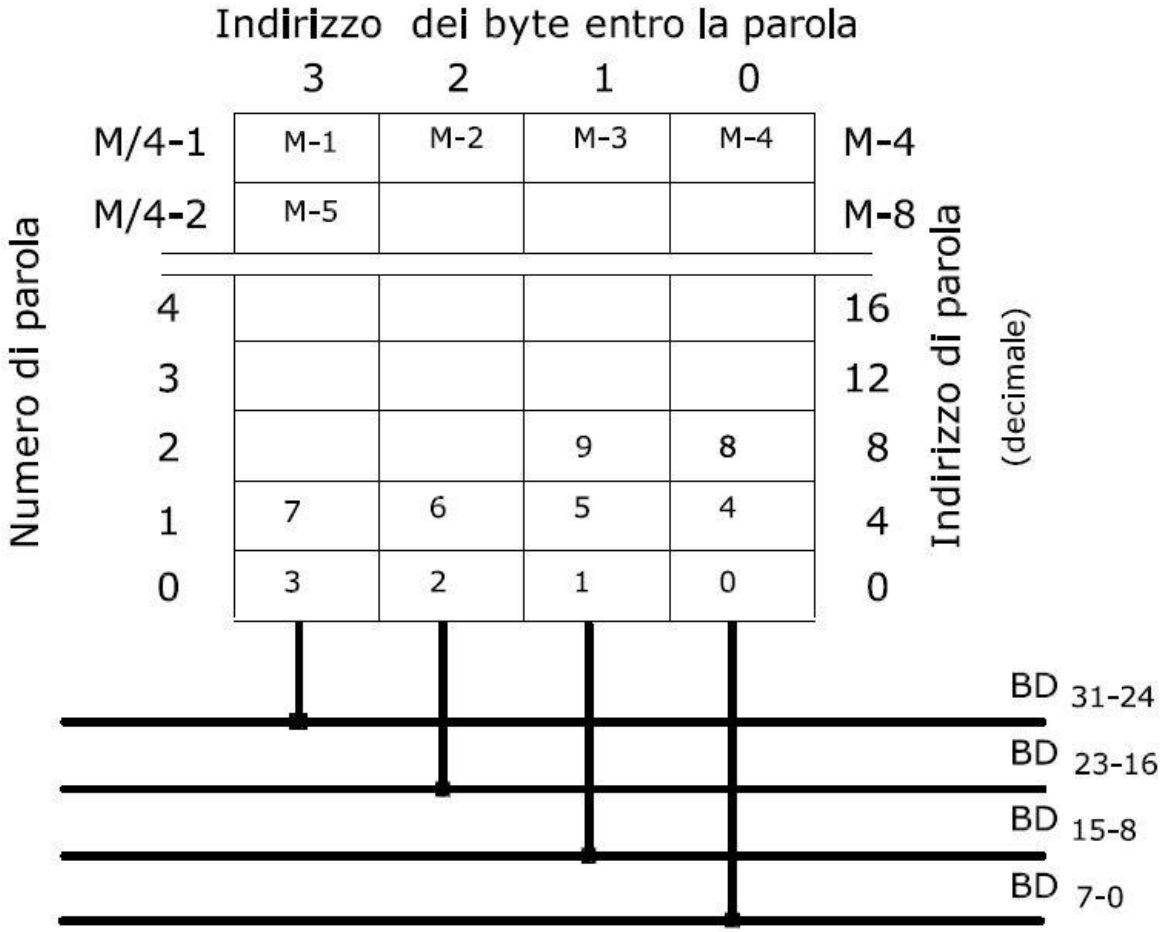


CS = Chip select

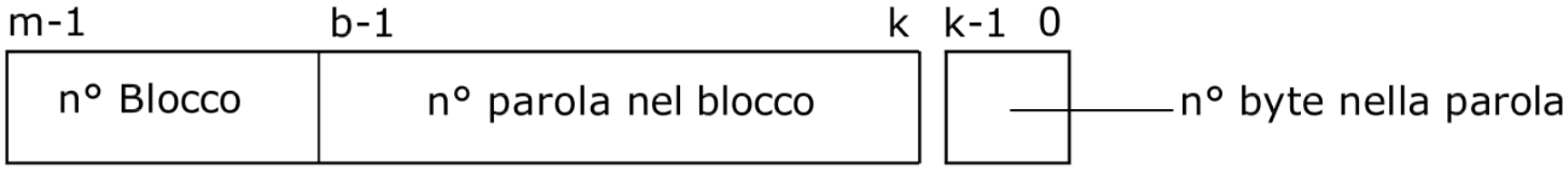
Parallelismo superiore al byte

- Il parallelismo più semplice è quello a 8 bit, tuttavia il parallelismo evolve ed ora abbiamo memorie con parallelismo a 64 bit.
  - Usiamo il termine *parola* o *word* per indicare i bit di parallelismo della memora (pari all'ampiezza del bus dati)
- Se ho una memoria con parallelismo a k byte, allora una parola di k byte è allineata se il suo byte meno significativo è a un indirizzo multiplo di k.
  - In generale una entità a l byte è allineata se il suo byte meno significativo è a un indirizzo multiplo di k
- È convenzione (anche per retrocompatibilità) usare il byte come cella elementare e si assume che gli indirizzi siano sempre associati ai byte.
  - Di conseguenza solitamente l'estensione della memoria è misurata in byte.

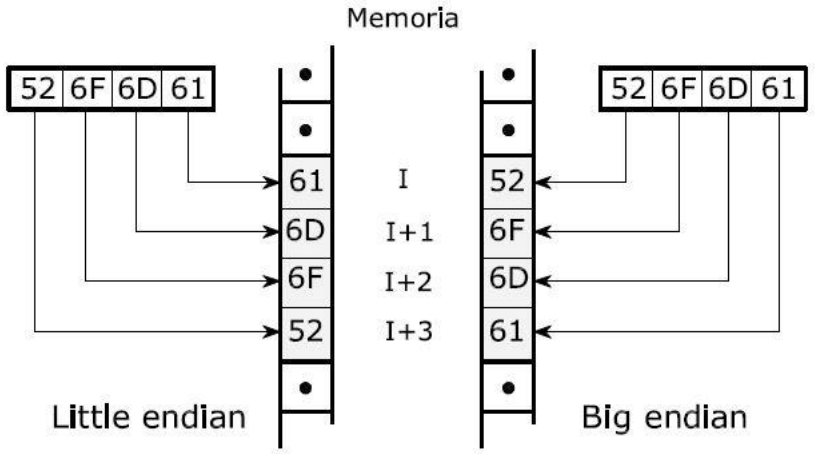
Parallelismo superiore al byte

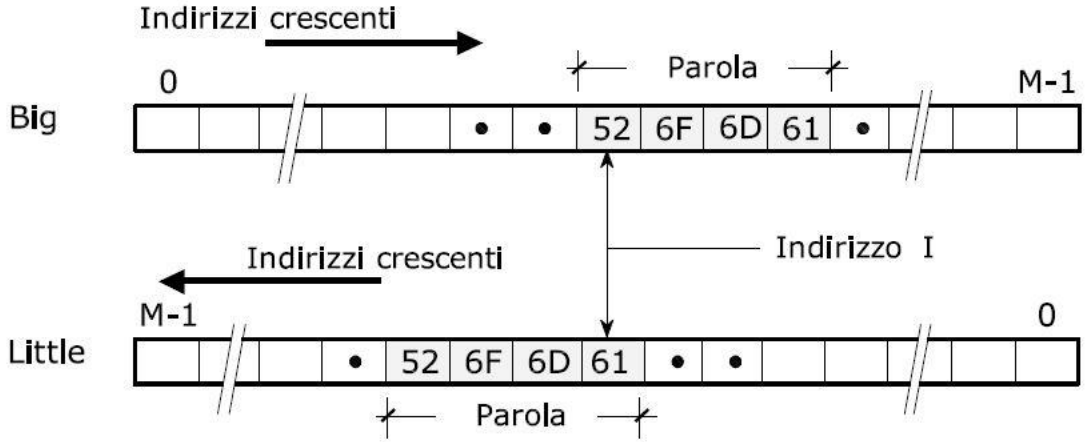


Indirizzamento byte con parallelismo superiore al byte



Big endian e little endian





Domande?

Riferimenti principali

- Capitolo 4 di ***Calcolatori elettronici. Architettura e Organizzazione,*** Giacomo Bucci. McGraw-Hill Education, 2017.