UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

Memory organization

A.Carini – Microcontrollers

Memory systems

- The term *memory system* (not just memory) wants to highlight the fact that information storing is performed by a set of solid state memory elements having different electrical characteristics, access times, capacities, and maybe even different technologies.
- The distinction between *immediate access storage* and *mass storage* used in PC has not much meaning in DSPs and µCs.
- Memories of DSPs and µCs are always *immediate access.*
- They are realized exclusively as integrated circuits, in CMOS technology, *on chip,* i.e., on the same chip of the processor, or *off chip*, as external components. The latter can have a large size and act as mass storage.

Memory systems

- Memories can be classified as :
- *Volatile*
  - *SRAM (static random access memory)*
  - *DRAM (dynamic random access memory)*
- *Non volatile*
  - *ROM (read only memory)*
  - *PROM (Programmable ROM)*
  - *EPROM (Electrically programmable ROM)*
  - *EEPROM (Electrically programmable and erasable ROM)*
  - *Flash EPROM*
- The main memory parameters are the memory *capacity*, *speed, power dissipation, integration density.*

Memory capacity

- The memory capacity is defined in bits, bytes, or words.
- Sometime, there is an indication of the internal organization, e.g., 64k x 4 bit, which means there are 64*1024 cells of 4 bits.

Memory speed

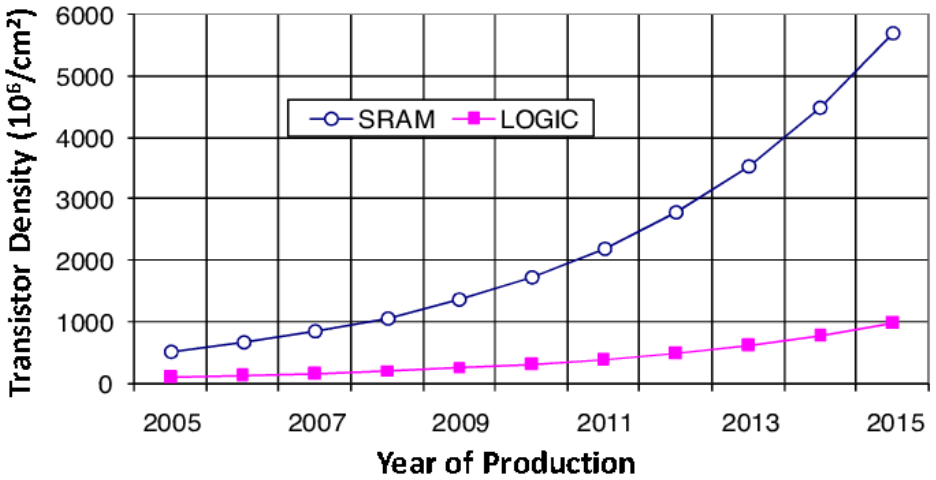
- Is described in terms of
- *Access time:* time (typically in ns) from the request of a data and the moment it is retrieved (read cycle). It is composed of two parts: the time necessary to decode the address and localize the data and the time to retrieve it.
- *Latency time:* time from the beginning of a write cycle and the stabilization of the data in memory. (Access time and latency time are often equal).
- *Bandwidth:* provides an indication of the speed at which data can be transferred form a processor to memory and vice versa. It is measured in byte per second.

Power dissipation

- The parameter refers to *external* memory chips. For *on chip* memories, it is never separately specified, but added to the power dissipation of the processor.
- Generally it defines the *average electrical power* required by the memory circuit.
- In some cases, producers specify the power dissipation in *stand-by* or the power dissipation during the read or write cycles.

Integration density

- This parameter depends on the technology used for fabricating the memory, in particular to the photolithographic definition used in producing the *masks*.
- The memory density is directly related to the integration density.



From: https://www.researchgate.net/figure/Trend-of-transistor-density-in-logic-elements-and-SRAM_fig3_224091827

Memory organization

- All memory systems are organized *hieratically*.
- The information the processor exchanges with the memory is characterized by a *spatial* and *temporal* locality.
- In fact, statistically, data in close spatial proximity inside the memory tend to be used after a short time. On the other hand, when a data is used by a processor, very often is used again after a short time.
- Performance can be drastically improved by combining a small amount of very fast (and expensive) memory with larger amounts of slower (but less expensive) memories.
- The most valuable processor memory is represented by the *registers* (tens).
- The next lower level of internal memory is the *SRAM.* Can store tens or hundreds of kB but is much slower than registers.

Memory organization

- In Harvard architectures, we have separate memories for instructions and data.
- In *modified Harvard architectures,* this subdivision can be extended in order to allow simultaneous access to an instruction and multiple data.
- In many cases, we have three memory banks (for 1 instruction and 2 operands), and three bus systems, which are prolonged off chip.
- In some cases, instead of multiple memories, there are *dual port* memories. They allow two simultaneous access to the memory because they duplicate all the circuits for address decoding and memory cell access.
- *Multi port* memories (with up to 4 simultaneous possible accesses) also exists.

Memory organization

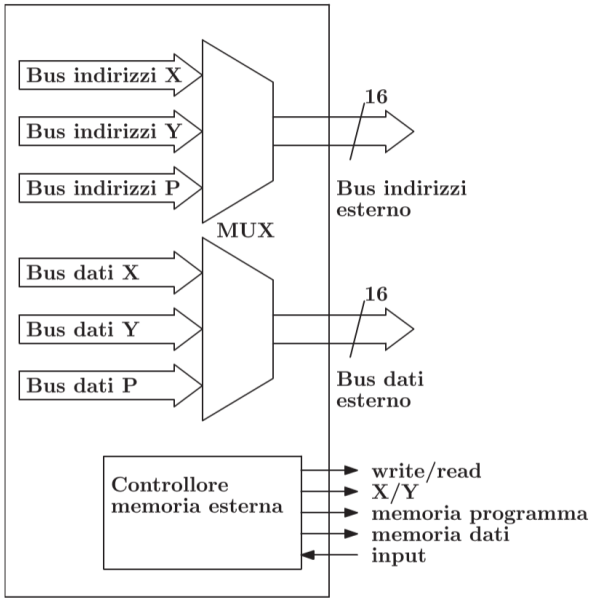


Figura 7.1: Esempio di schema di espansione della memoria interna su *due* bus esterni a 16 bit. Internamente, sono disponibili 3 sistemi di bus, uno per le istruzioni (P) e due per i dati (X e Y). Solo uno di essi, selezionabile tramite un multiplexer, MUX, è però portato all'esterno.

Memory organization

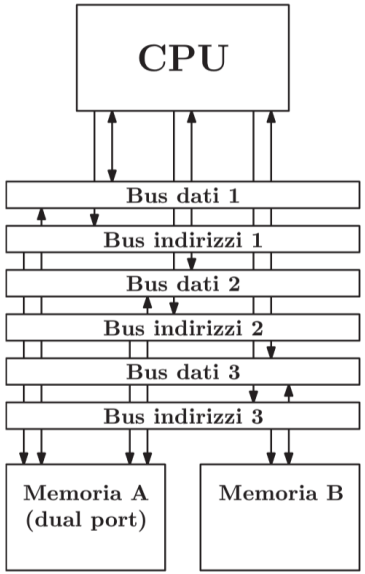


Figura 7.2: Organizzazione di tipo misto: una memoria dual port (banco A) e una memoria standard (banco B), permettono l'accesso simultaneo a una istruzione (nel banco B) e due dati (entrambi nel banco A). Una simile soluzione si trova nei DSP di NXP della serie 56F8xxx.

External memory controller

- In DSPs and µCs, control of external memory is performed by an appropriate circuit, realized on the same chip (without additional external controllers).
- This circuit is a first example of *peripheral unit*, i.e., a section of the integrated circuit that allow interactions between the processor and the external world.
- The characteristics of the integrated controllers are very variable.
- Some generate only the essential signals for the external bus (select, strobe, …)
- Others are more flexible and manage *wait states* and PC-like external memories (e.g., page mode DRAM).
- Wait states are necessary any time the external memory is slower than the processor.

Static RAM (SRAM)

- In DSPs and µCs, *volatile* read and write memories are always *SRAM* (Static Random Access Memory).
- The use of SRAM memories is justified by its high speed (with access time of 1 ns) and easy management (they do not need a refresh phase).
- Main limitations are: power dissipation (as high as 1 mW per kbyte) and silicon area occupation (6 transistors per cell).
- In the past, SRAMs were used also for external memories, but DRAM are now preferred for their lower power dissipation and higher density for similar costs.

Static RAM (SRAM)

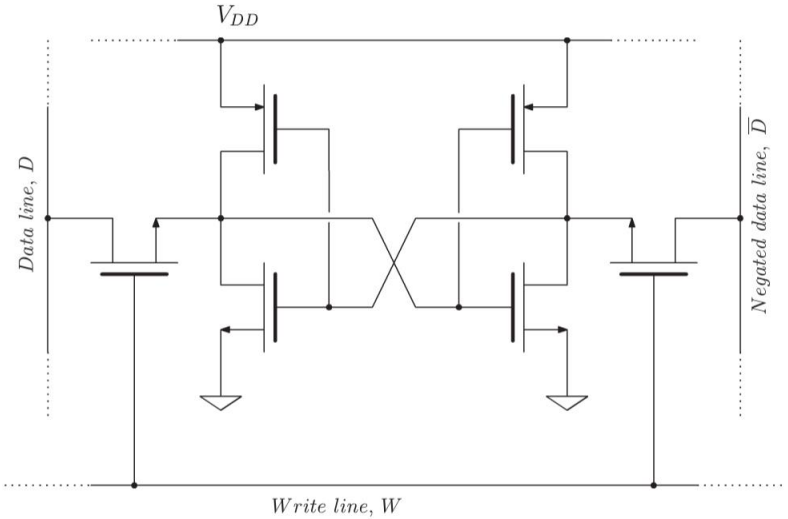


Figura 7.3: Cella di memoria SRAM. La scrittura avviene portando la write line W a livello alto e asserendo la linea dati desiderata.

Static RAM (SRAM)

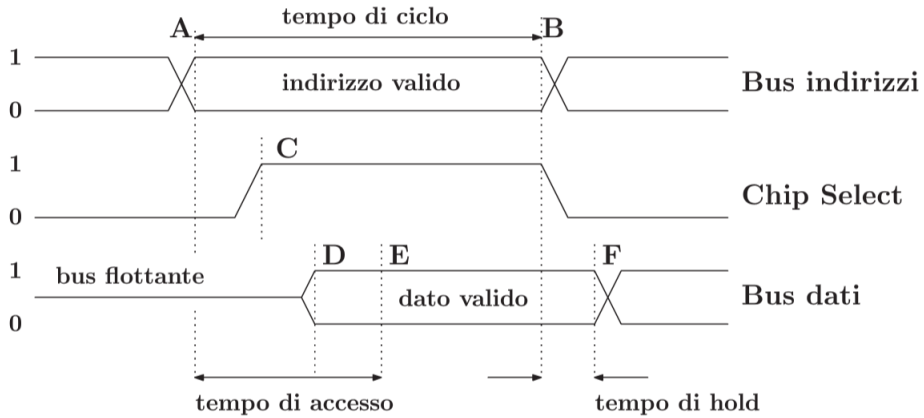


Figura 7.4: Diagramma temporale dei segnali che controllano la lettura di un dato in una SRAM.

Dynamic RAM (DRAM)

- DRAMs are often used as external memory expansions, as an alternative to Flash EPROMs.
- DRAMs are characterized by a high density and low cost, but also require a refresh phase, which imposes the use of a dedicated external control (on chip).
- This control must manage dynamic wait-states, i.e., must progressively prolong the wait-state if the memory response time get longer.
- DRAMs have a lower cost and a higher density (4 times higher) that SRAMs.
- They also have lower power dissipation (at most 0.1 mW/kbyte).
- Information is memorized in the form of a charge stored in a MOS condenser.
- DRAM cycle times are higher that SRAM but still on the order on 1 ns.

Dynamic RAM (DRAM)

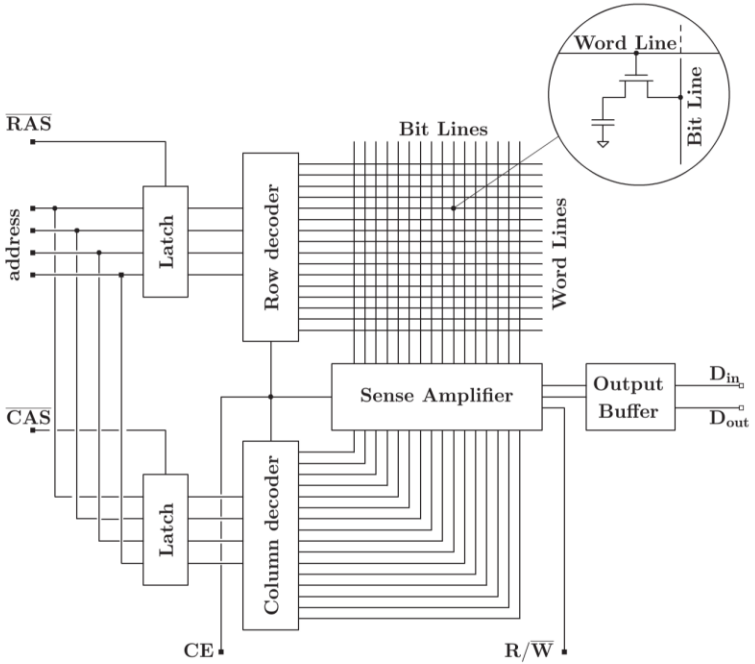


Figura 7.5: Struttura di principio di una memoria DRAM. Si notano i decodificatori di riga e colonna e l'amplificatore di lettura. Nell'inserto è visibile lo schema circuitale di un bit, in cui appaiono due transistori MOS; il primo viene usato come un vero e proprio interruttore comandato, il secondo solo per immagazzinare carica elettrica.

Dynamic RAM (DRAM)

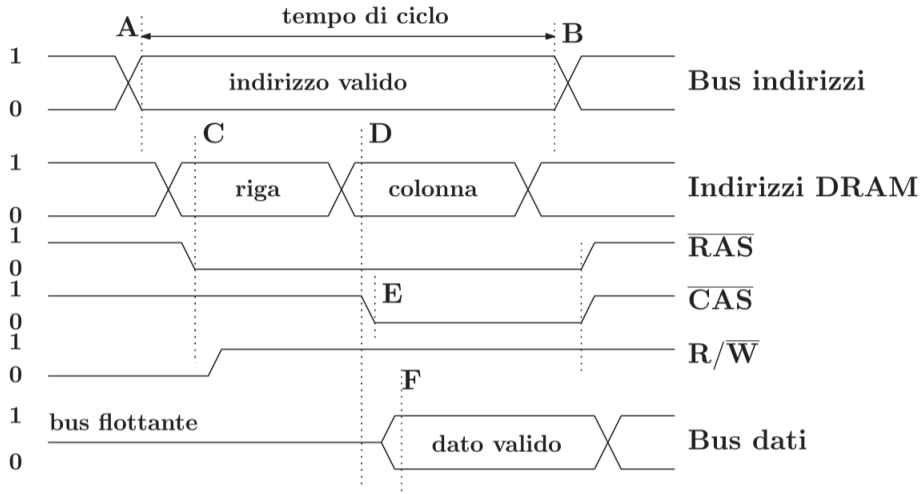


Figura 7.6: Diagramma temporale per i segnali che controllano la lettura di un dato in una DRAM. Il circuito di controllo individua i bit di interesse attraverso un indirizzo di riga e di colonna, che transitano in tempi diversi sul bus indirizzi del chip. I valori letti sono resi disponibili sul bus dati e mantenuti fino al temine del tempo di ciclo.

Read only memories (ROMs)

- DSPs and µCs often include on chip a small amount of read only memory.
- This can be the only memory on chip in case of very high production volumes with a well defined application code.
- More commonly, we can find on chip or off chip some re-programmable ROM memories.
- The most common are:
  - PROM or OTP ROM - one time programmable ROM
  - Flash EPROM, non volatile ROM which can be electrically written and cancelled many times.

PROM or OTP ROM

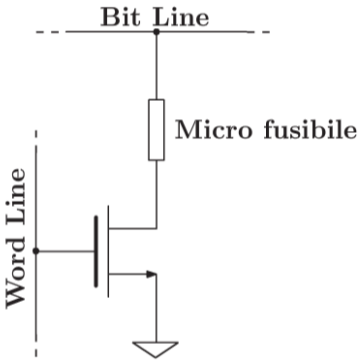


Figura 7.7: Struttura di principio di una cella di memoria PROM. La programmazione avviene, in modo irreversibile, interrompendo il micro-fusibile.

- Once programmed, they are very fast with access times of ~ 1ns

Flash EPROM

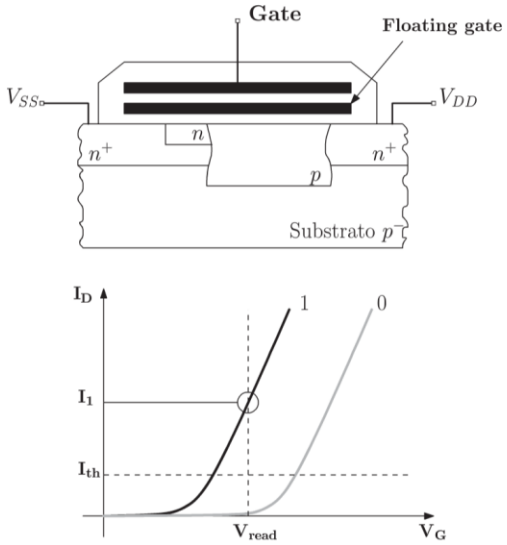


Figura 7.8: Struttura di principio di una cella di memoria Flash EPROM. Essa sfrutta una delicata struttura, denominata gate flottante (floating gate). In basso sono visualizzate le caratteristiche tensione corrente della cella, nel caso di nessuna carica intrappolata, 1, e di carica intrappolata, 0.

Flash EPROM

- Information is stored in the form of a charge, accumulated in the floating gate.
- The charge is taken to the floating gate through the thin silicon oxide between floating gate and drain using *hot electron injection* (with G and D at 10 V).
- The negative charge in the floating gate raises the threshold voltage of the MOS transistor, and thus the MOS is turned off.
- By applying a positive voltage to the source with respect to the gate, the electrons can leave the floating gate thanks to the Fowler-Nordheim tunneling effect (requires small oxide thickness < 10nm).
- Flash memories can be written only a finite number of times (> 10^5). It is better, to write on different sectors.
- Erase and write times can be long (on the order of ms per 256 kB)
- Read and write time (of the order of ns) are variable with parameters (age, temperature, fabrication process) and require a dynamic control.

Flash EPROM

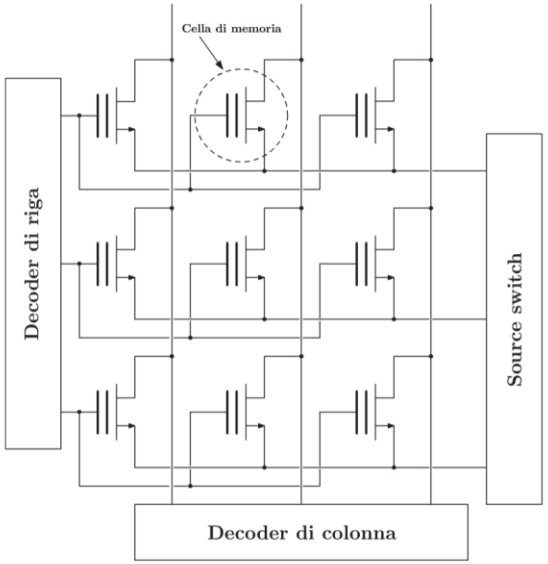


Figura 7.9: Struttura di principio di un blocco di memoria Flash EPROM. I source dei dispositivi dello stesso blocco sono connessi ad un sistema di deviatori che consente di connetterli tutti insieme a V_{SS}, oppure, in cancellazione, a V_{DD}. Non è quindi possibile cancellare o programmare una singola cella.

Cache memories

- Performance can be significantly improved with limited costs using the concepts of *virtual* or *cache* memory.
- Cache memory can improve performance only in those cases where memory cannot be read in one clock cycle.
- This is not the case in many µCs that have low clock frequencies.
- It is the case of high end DSPs and DSCs, which can work with clock frequencies of hundreds of MHz.
- Cache memories used in DSPs/DSCs are simpler than GPPs. Cache is used only for the instruction memory, not for data (no write back).
- In the most simple cases, the cache is a simple *buffer* that stores small sequences of instructions: *repeat buffer.*

Cache memories

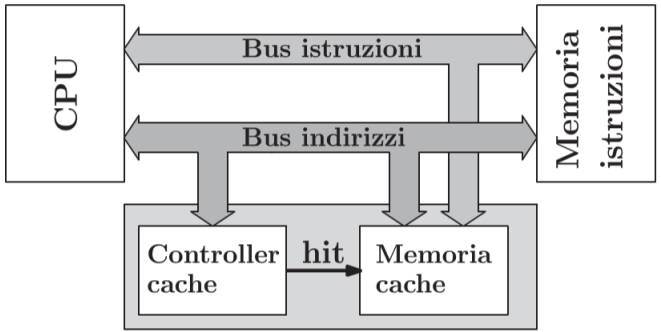


Figura 7.10: Struttura di principio di una cache per la memoria istruzioni di un DSP. L'organizzazione prevede lo sdoppiamento dei bus istruzioni e indirizzi e un opportuno circuito di controllo.

Cache memories

- The main parameter of cache memories is the *hit-ratio h,* defined as the ratio between the memory accesses performed on cache and the total number of memory accesses.
- The *speed-up* ratio is then: $$S = \frac{1}{1 + h \cdot (\frac{t_c}{t_{acc}} - 1)}$$
- Where t_{acc} is the access time to the main memory, and t_c to the cache.
- Taking into account the fraction *q* of machine cycles spent for activities that do not involve memory:

$$T_{cycle_{AVG}} = q \cdot T_{clock} + (1 - q) \cdot (h \cdot t_c + (1 - h) \cdot t_{acc}),$$

Cache memories

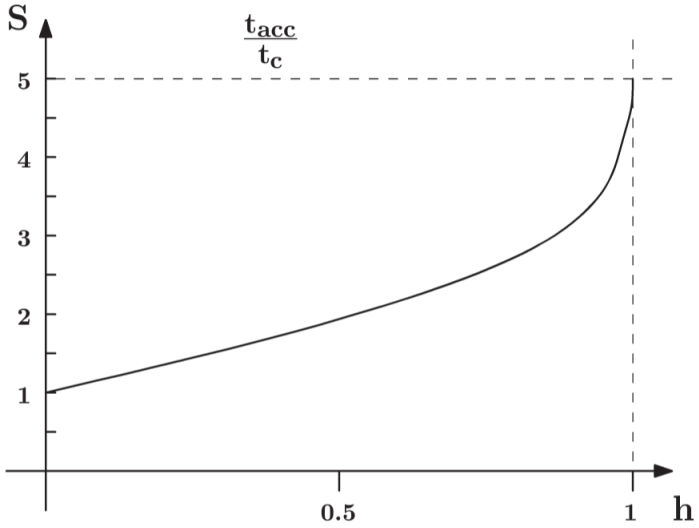


Figura 7.11: Speed-up ratio di una cache in funzione dell'hit ratio. La figura assume che sia $\frac{t_{acc}}{t_c} = 5$.

Direct mapping cache

- The *direct mapping cache* is the most common cache used in DSPs.
- The DSP memory is logically segmented into 2^B blocks, each with 2^R rows of 2^W words, i.e., the address is divided into *B,R,* and *W* bits.
- The cache memory has the same size of a block, i.e., 2^R rows. Each row can come from any of the 2^B blocks.
- Internally, the cache is divided into two sections *DATA* and *TAG*. The first section stores the data, the second the block address of each row.
- During read, the row is localized in the cache using the *R* bits.
- Simultaneously, the *B* is compared with the row *TAG.*
- If *B==TAG,* we have a *hit* and the desired word is placed on the bus.
- If *B/=TAG,* we have a *miss,* and the searched row is copied from the external memory to the cache at index *R,* replacing also the *TAG.*

Direct mapping cache

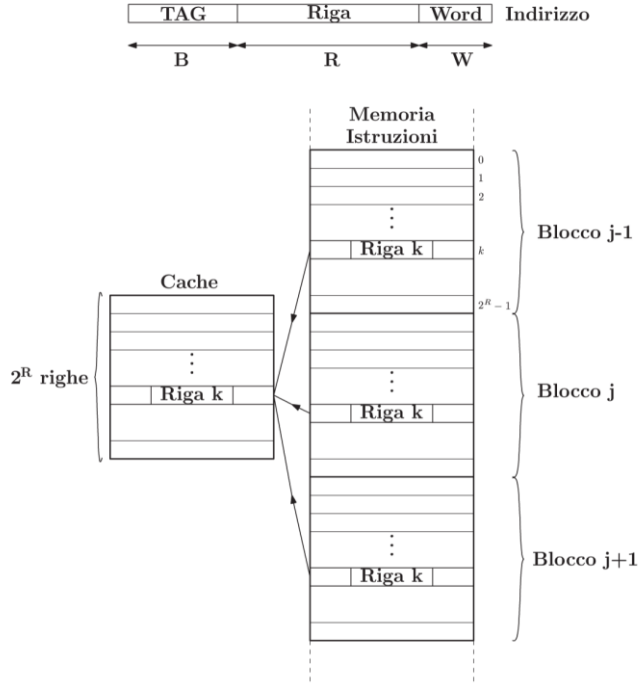


Figura 7.12: Struttura di una cache a mappatura diretta. La memoria è suddivisa in righe, ciascuna delle quali contiene 2^W istruzioni. A loro volta, 2^R righe costituiscono un blocco. La cache ha tante righe quante sono quelle di un blocco.

Direct mapping cache

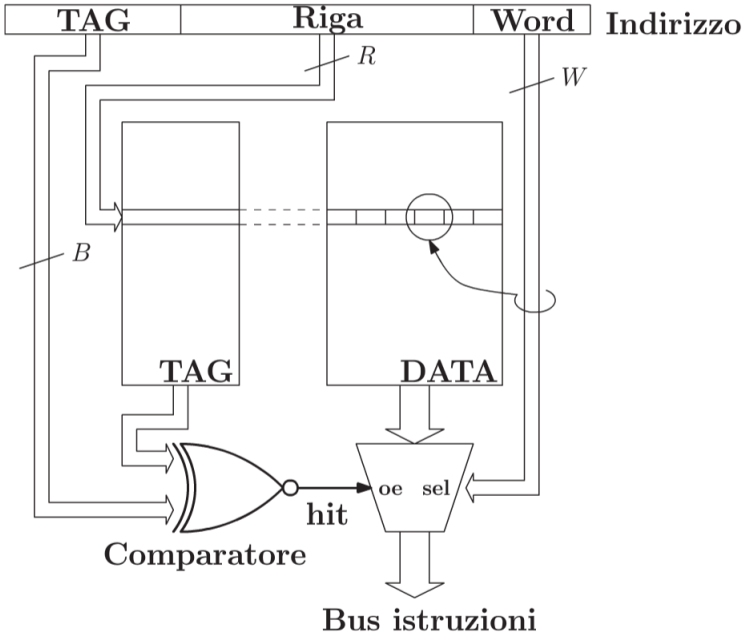


Figura 7.13: Funzionamento di una cache a mappatura diretta. Per poter rapidamente discriminare una condizione di *hit* da una di *miss* vengono comparati due TAG, uno proveniente dal bus indirizzi del processore e uno presente nella cache alla stessa riga.

See:

- Simone Buso, «Introduzione alle applicazioni industriali di Microcontrollori e DSP» Società editrice Esculapio, 2018
  - Chapter 7