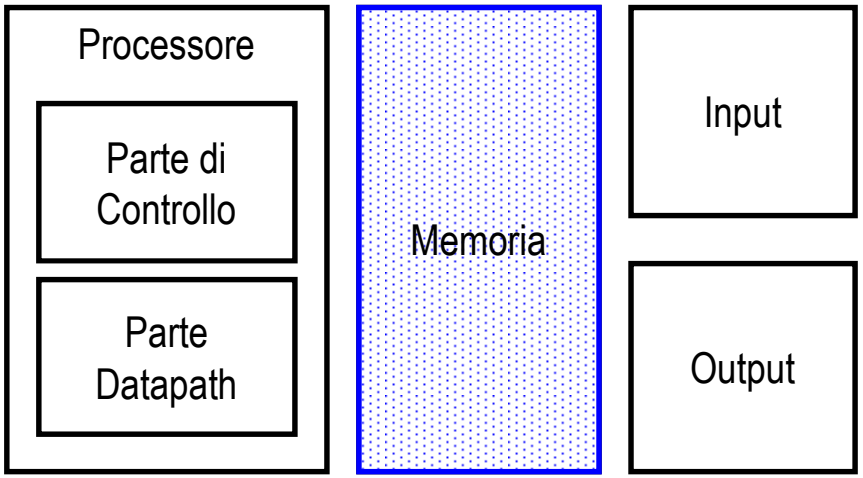
Lezione 16
Introduzione al sottosistema di memoria

http://www.dii.unisi.it/~giorgi/didattica/arcal1

Dove siamo

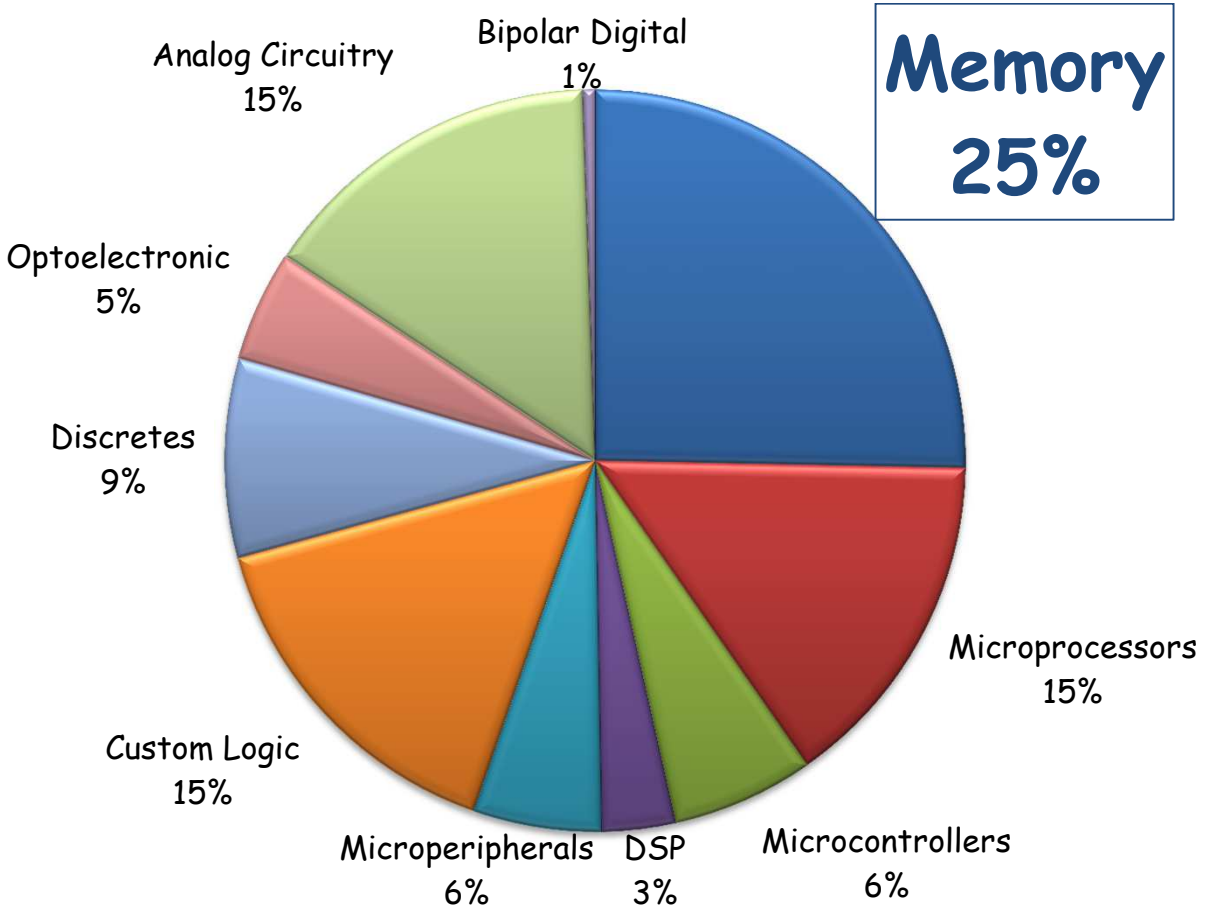
- I Cinque Componenti Classici di un Calcolatore



- Scaletta:
  - Tecnologie usate per realizzare le memorie "volatili": DRAM e SRAM
  - Tecnologie usate per l'accesso alla memoria DRAM: EDO, SDRAM, DDR, ...

Market Share dei Principali Componenti Elettronici

SITUAZIONE RIFERITA ALL'ANNO 2000
SOURCE: Primer on Semiconductors, SG Cowen, 3rd Ed. Jan. 2001



ANNO 2007: MEMORIA CIRCA 37% DEL MERCATO TOTALE PER HIGH PERFORMANCE COMPUTING (SOURCE: BCCRESEARCH.COM)
PREV. 2012: MARKET SHARE COSTANTE, CRESCITA CIRCA 9% ANNUO

Tecnologie usate per realizzare le memorie

- Tecnologie ad accesso sequenziale
  - Il tempo di accesso dipende in maniera lineare dalla locazione fisica
    - Nastri

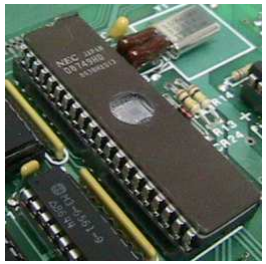
- Tecnologie ad accesso semicasuale
  - Il tempo di accesso dipende sia dalla locazione fisica dell'informazione che dal particolare istante in cui faccio accesso
    - Dischi: dischi-fissi, CDROM, DVD, ...

- Tecnologie ad accesso casuale (Random Access)
  - Il tempo di accesso e' indipendente dalla locazione fisica dell'informazione
    - Memoria principale
    - Due tipi: Non-Volatili e Volatili

Memorie accesso casuale di tipo Non-Volatile

- ROM (Read-Only Memory)
  - Non scrivibile (sia vantaggio che svantaggio)

- EPROM (Erasable Programmable Read-Only Memory)
  - ✓ Scrivibile (ma una sola volta)
  - ✓ Cancellabile a livello di chip (raggi UV)
  - ☹ Necessario intervenire con apposito "kit di cancellazione"



- EEPROM (Electrically-Erasable Programmable Read-Only Memory)
  - ✓ Scrivibile (1 sola volta)
  - ✓ Cancellabile a livello di byte (elettricamente)
  - ☹ Dovendo scrivere molti byte si perde molto tempo nelle cancellazioni

- Memoria "Flash": Evoluzione della EEPROM
  - ✓ Scrivibile (1 sola volta)
  - ✓ Cancellabile a livello di blocco (elettricamente)

Memorie accesso casuale di tipo Volatile

- Conservano i dati solo se alimentate
- Principali tipi di implementazioni:
  - DRAM (Dynamic Random Access Memory)
    - ✓ Alta densita', basso consumo, basso costo
    - ☹ Lente, dinamiche → Necessita' di essere "rinfrescata" periodicamente
  - SRAM (Static Random Access Memory)
    - ✓ Veloci, statiche → basta alimentare il chip per non perdere i dati
    - ☹ Bassa densita', alto consumo, alto costo

Approfondimento: v. slide "Costi e Prestazioni delle Memorie"

Parametri delle Prestazioni della Memoria

t_a=tempo di accesso (Access Time)

- Intervallo fra l'inizio di una lettura (indirizzo su A-bus) e l'arrivo del PRIMO dato (su D-bus)

t_c=tempo di ciclo (Cycle Time)

- Intervallo fra l'inizio di una lettura e l'inizio della lettura successiva

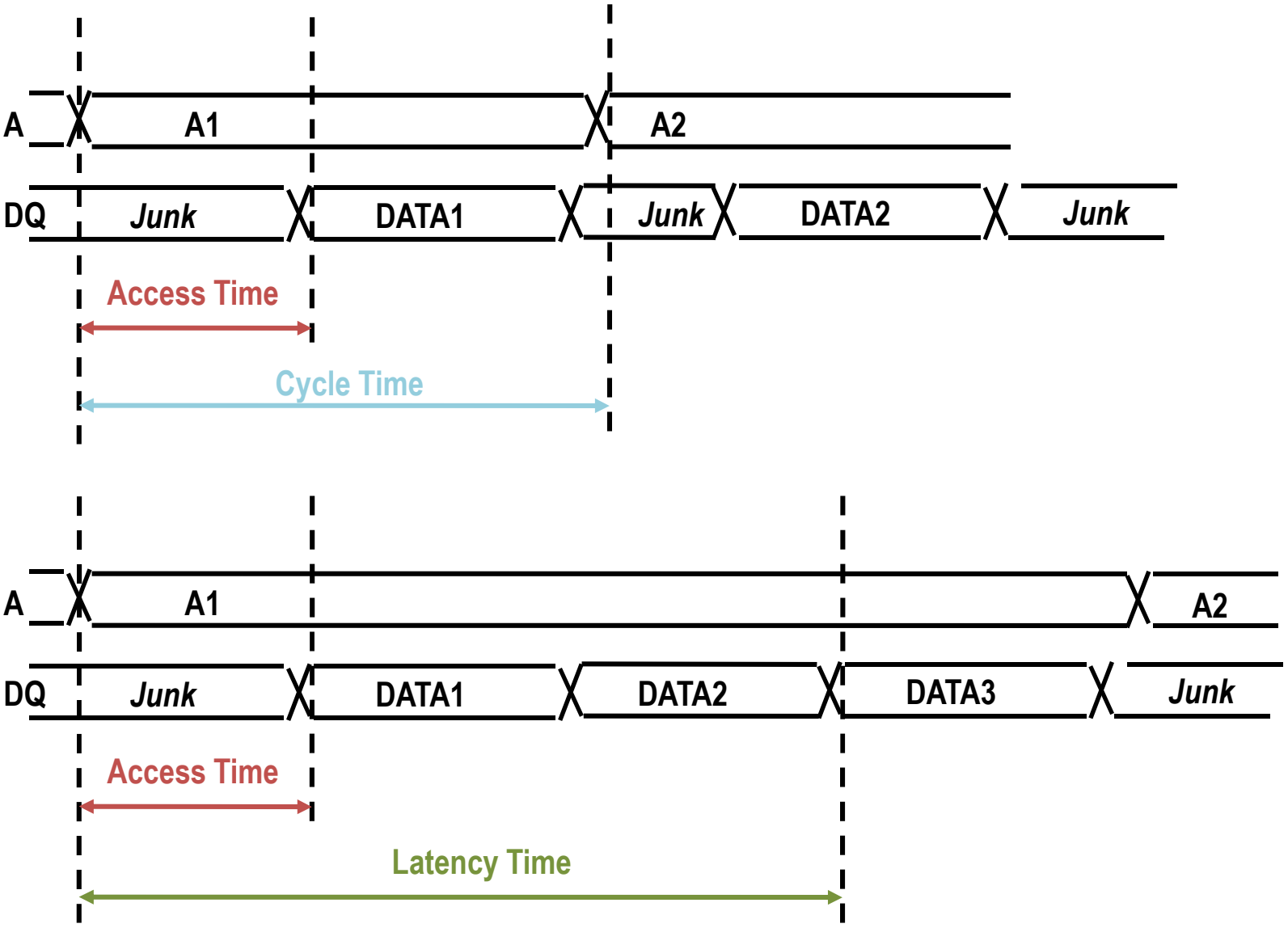
t_l=tempo di latenza (Latency Time)

- Tempo di accesso al DATO COMPLETO ovvero alla k-esima word nel caso in cui la memoria supporti trasferimenti a gruppi di k word (se k=1 $t_l=t_a$).
  Es. dischi, ma anche memorie RAM

ω=Banda (Bandwidth)

- Tasso di trasmissione dei byte (si misura in byte/s)

Tempi caratteristici delle memorie

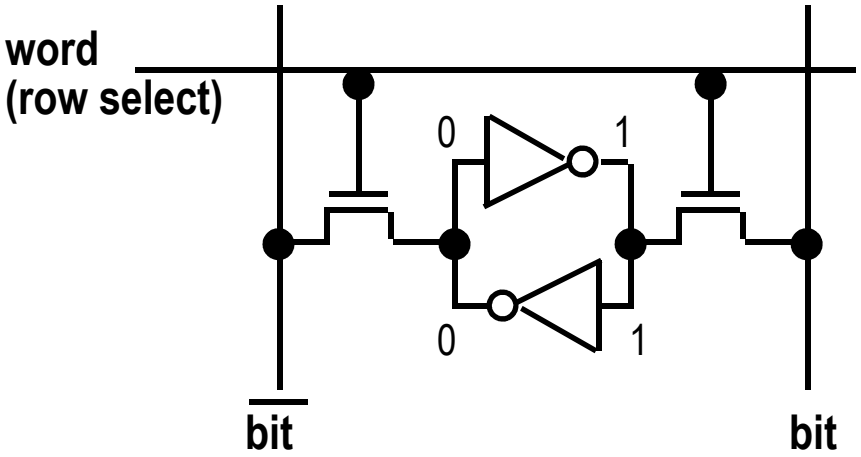


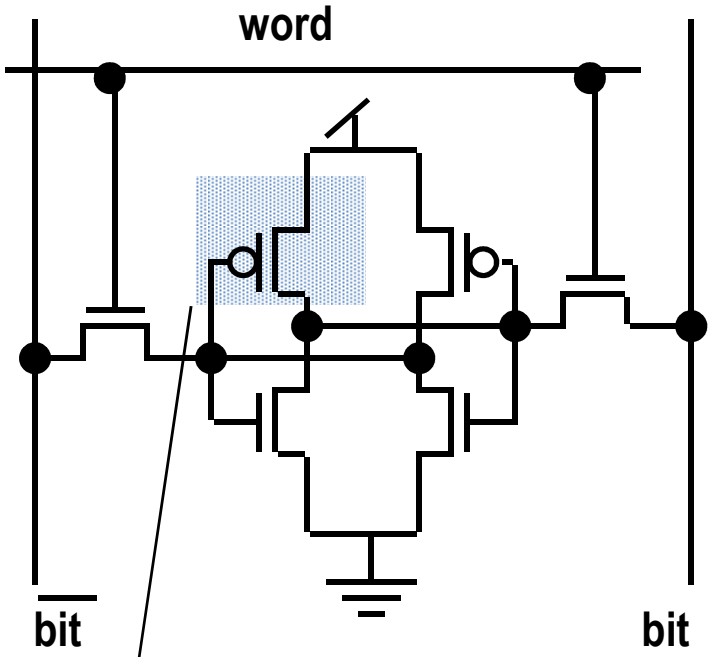
DRAM versus SRAM

- Per massimizzare il parametro prestazioni/costo
  - Si usa poca memoria SRAM (costosa e veloce) all'interno del chip (es. registri)
  - Si usa una ragionevole quantita' di DRAM (meno costosa e relativamente veloce) all'esterno del chip

Cella di RAM Statica (SRAM)

Cella a 6 Transistor





Puo' essere sostituito da una resistenza di pull-up per risparmiare area
→ Cella a 5 transistor

Scrittura

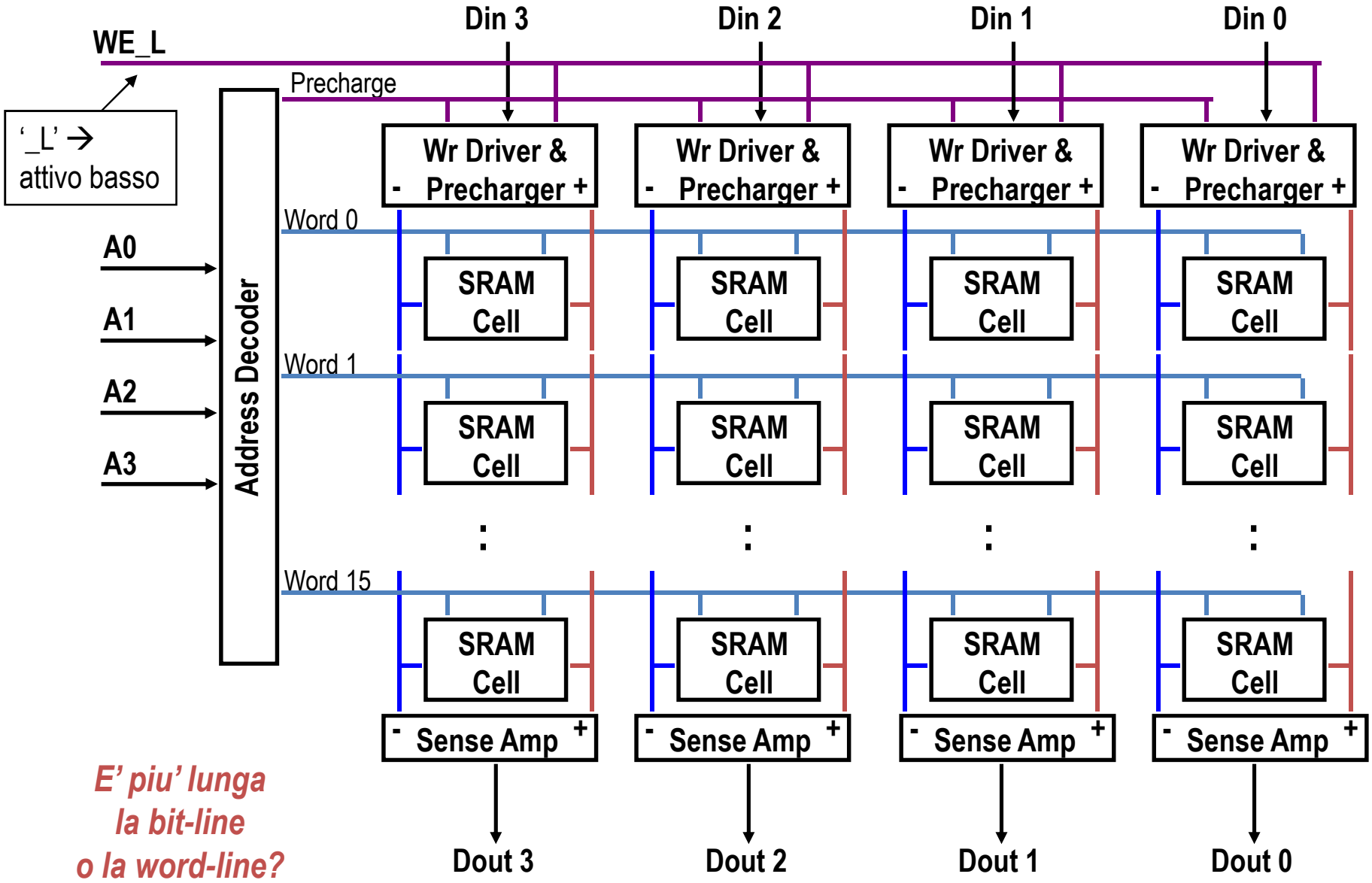
1. Preparare il dato sulle bit-line
2. Selezionare la riga (word-line)

Lettura

1. Precaricare bit e /bit a Vdd *
2. Selezionare la riga (word-line)
3. La cella preleva carica da bit o /bit
4. Il Sense-Amplifier in fondo alla colonna amplifica la differenza fra bit e /bit

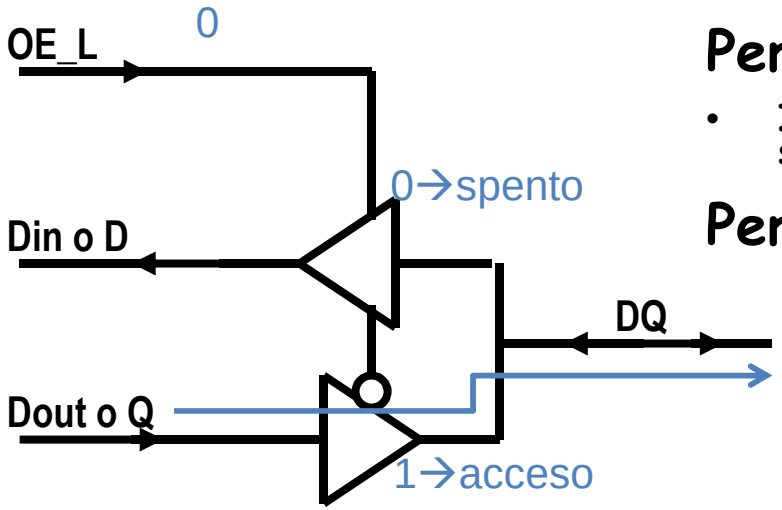
*in realta' la linee bit e /bit si caricherenno ad un valore un poco <Vdd

Organizzazione delle celle SRAM: 16-word x 4-bit



E' piu' lunga la bit-line o la word-line?

SRAM: Bus dati



Per Din e Dout si usa un unico filo

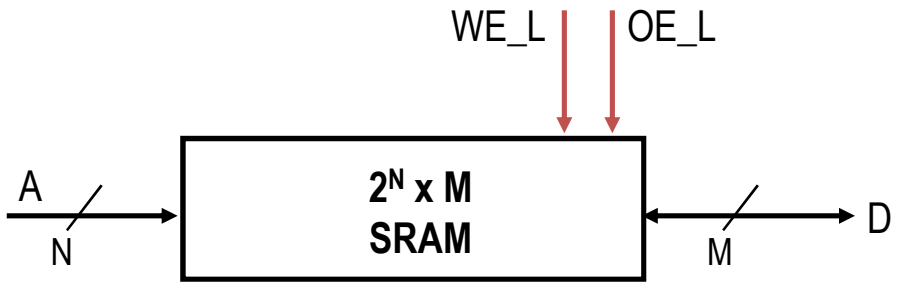
- Il motivo principale e' perche' sul bus dati posso avere scambio in entrambe le direzioni (lettura o scrittura)

Per decidere la direzione si usa /OE:

- OE_L → Output Enable (attivo basso), per controllare i buffer thristate:
  1 sul filo thristate → buffer attivo
  0 sul file thristate → buffer disattivo

- Le possibili combinazioni sono:
  - Scrittura:
    - WE_L attivo (basso), OE_L disattivo (alto) → DQ e' un ingresso
  - Lettura:
    - WE_L disattivo (alto), OE_L attivo (basso) → DQ e' un'uscita
  - Evitare di attivare contemporaneamente i segnali WE_L e OE_L
    - Il risultato risulta indeterminato
- Quando il chip non deve caricare il bus dati, si disabilita con
  - CE_L disattivo → alta impedenza

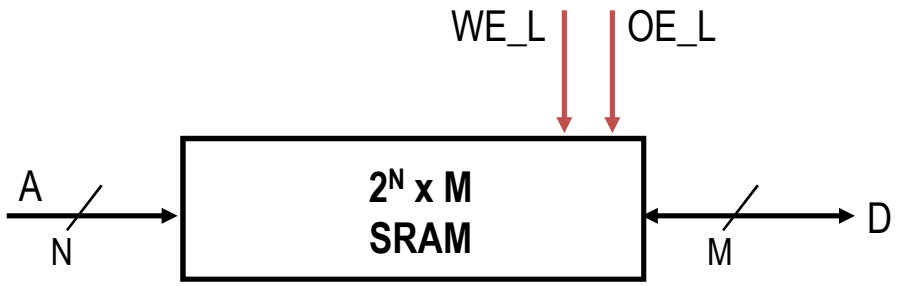
SRAM: Diagramma Logico

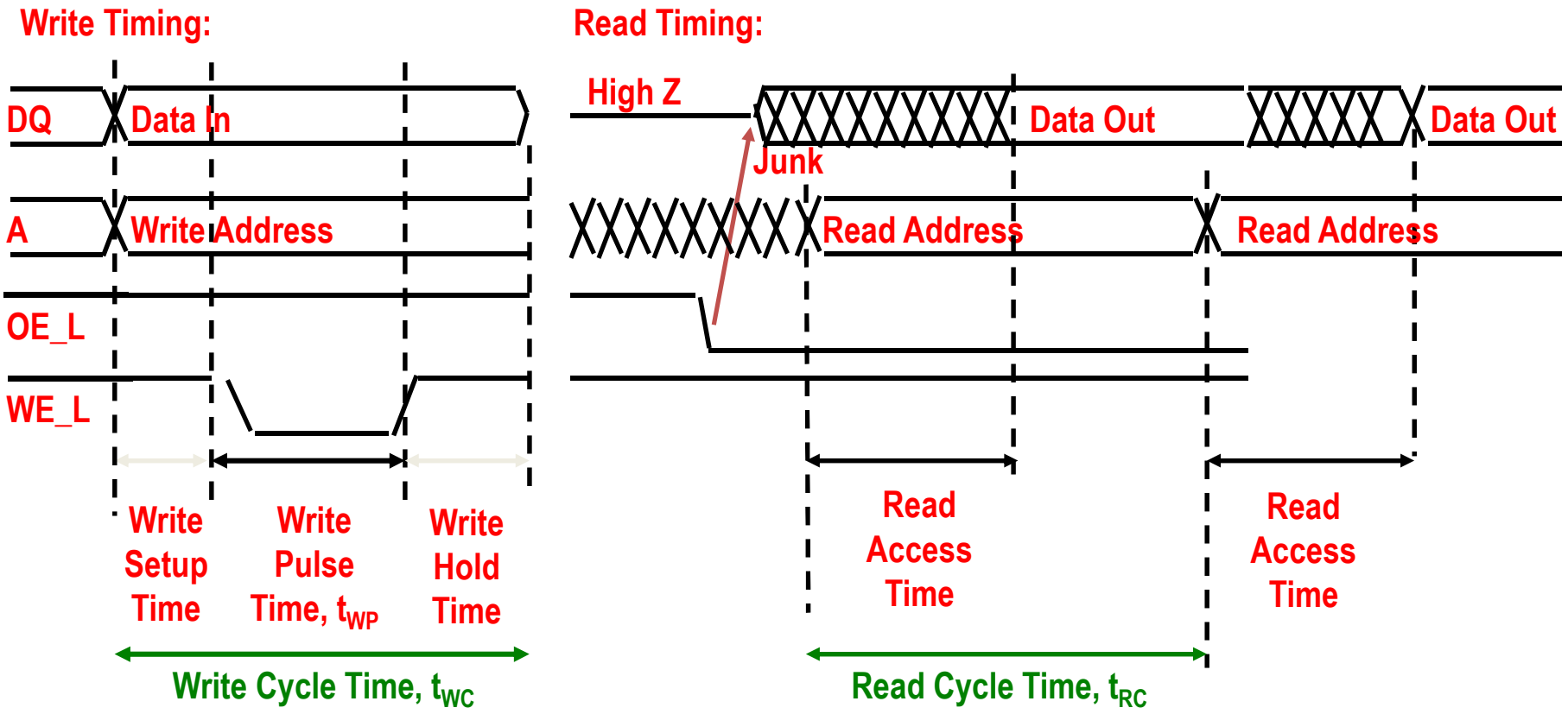


- La memoria ha una capacita' di 2^N word ad M-bit

- Altri segnali sempre presenti
  - CE_L → Chip Enable
  - Vcc, GND → Alimentazione

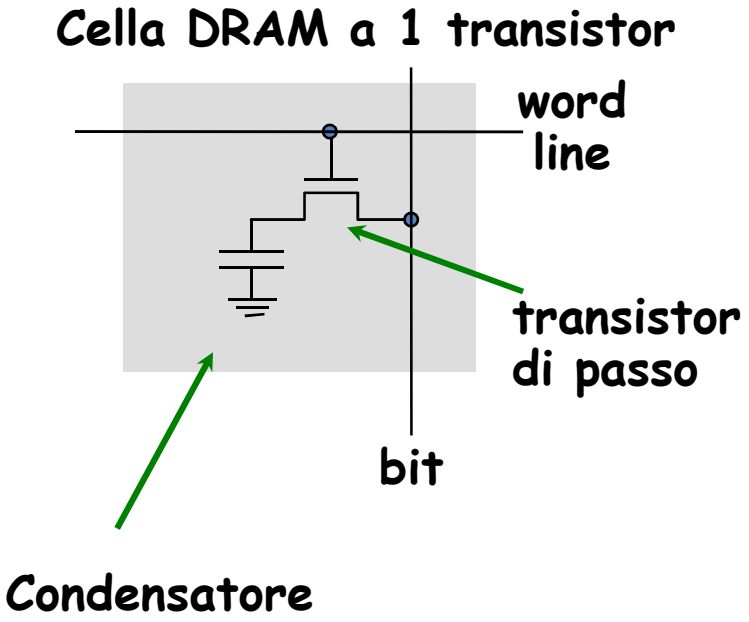
SRAM: Temporizzazioni



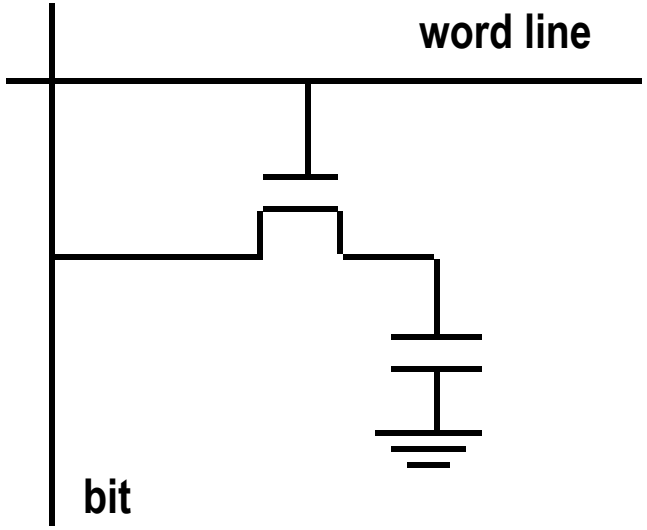


Cella di RAM dinamica (DRAM)

- La cella a 6 transistor per quanto possa sembrare piccola, ha un peso notevole sull'area totale di una RAM
- Qual e' il numero minimo di transistor che puo' essere usato per memorizzare un bit?
- Un solo transistor → l'elemento di memorizzazione puo' essere realizzato con un consensatore
  - Condensatore carico → 1 logico
  - Condensatore scarico → 0 logico
  - Il transistor consente la lettura e la scrittura



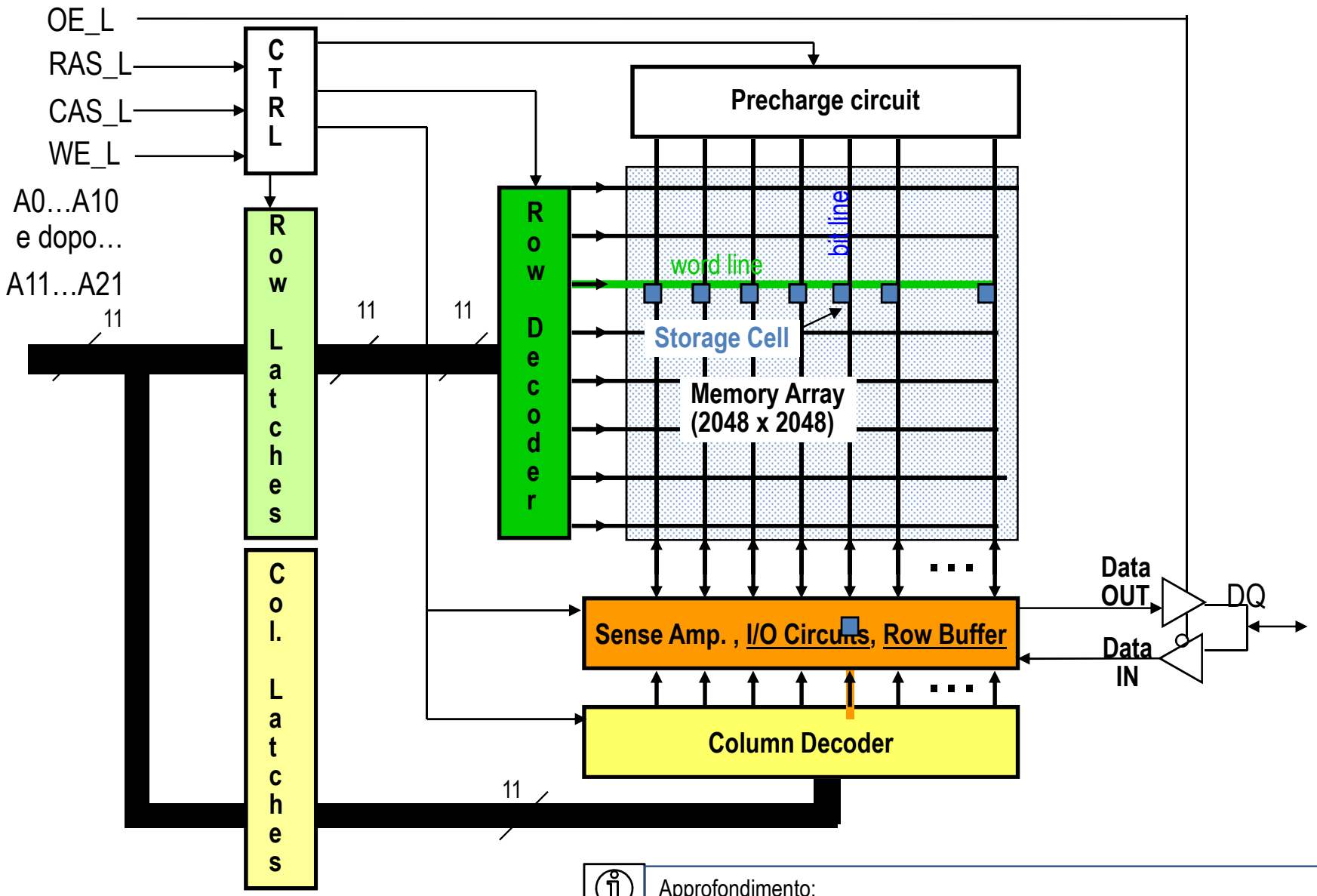
DRAM: Funzionamento della cella a 1 transistor



- **Scrittura:**
  1. Caricare la bit-line col dato
  2. Selezionare la riga (word-line)
- **Lettura:**
  1. Precaricare la bit-line a Vdd/2
  2. Selezionare la riga (word-line)
  3. Scambio di carica fra cella e bit-line
     - lievissima variazione di tensione sulla bit-line
  4. Rilevazione della variazione (Sense-Amplifier molto sofisticato)
     - l'amplificatore deve essere in grado di rilevare una variazione di carica di circa 1 milione di elettroni
  5. Caricare la bit-line col dato letto (scrittura !)
     - Ripristino il contenuto della cella
- **Refresh**
  1. Effettuo una "finta" lettura di <u>tutte</u> le celle
     - Ripristino il contenuto di <u>ogni</u> cella

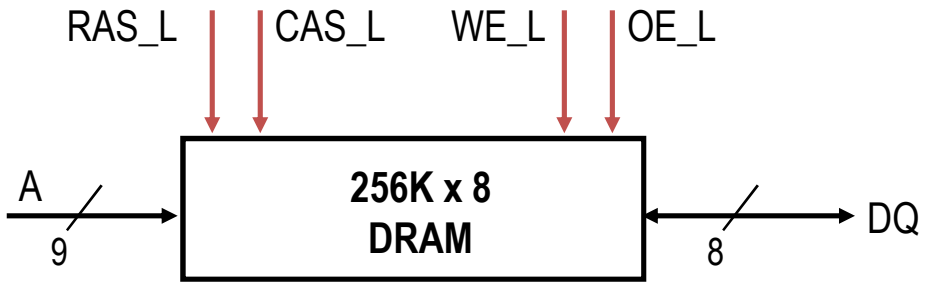
N.B. La cella ha
<u>un solo filo</u> di accesso
(word-line)
→ separazione ROW/COL

DRAM: Organizzazione logica di memoria a 4 Mbit



Approfondimento:
v. slide “DRAM: Organizzazione fisica di memoria a 4 Mbit”

DRAM: Diagramma Logico

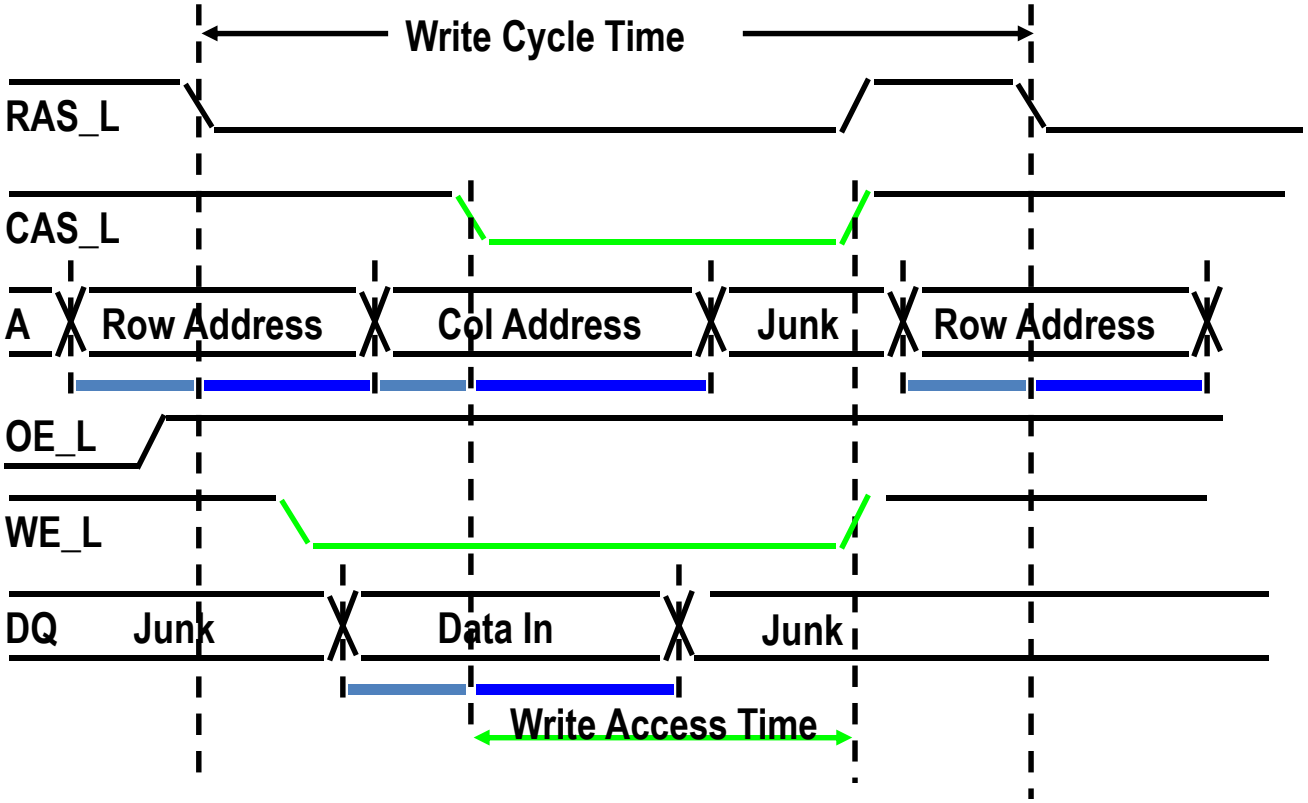


- Din e Dout sono multiplexati su DQ:
  - Scrittura:
    - WE_L attivo (basso), OE_L disattivo (alto) → DQ e' un ingresso
  - Lettura:
    - WE_L disattivo (alto), OE_L attivo (basso) → DQ e' un'uscita
- Gli indirizzi di Riga e di Colonna condividono i pin A
  - RAS_L va basso → i latch di riga memorizzano i pin A
  - CAS_L va basso → i latch di colonna memorizzano i pin A
  - Nota: RAS/CAS sono sensibili al fronte in discesa

Nota: In alcuni chip di DRAM il segnale OE non c'e' e invece di DQ ci sono Din e Dout

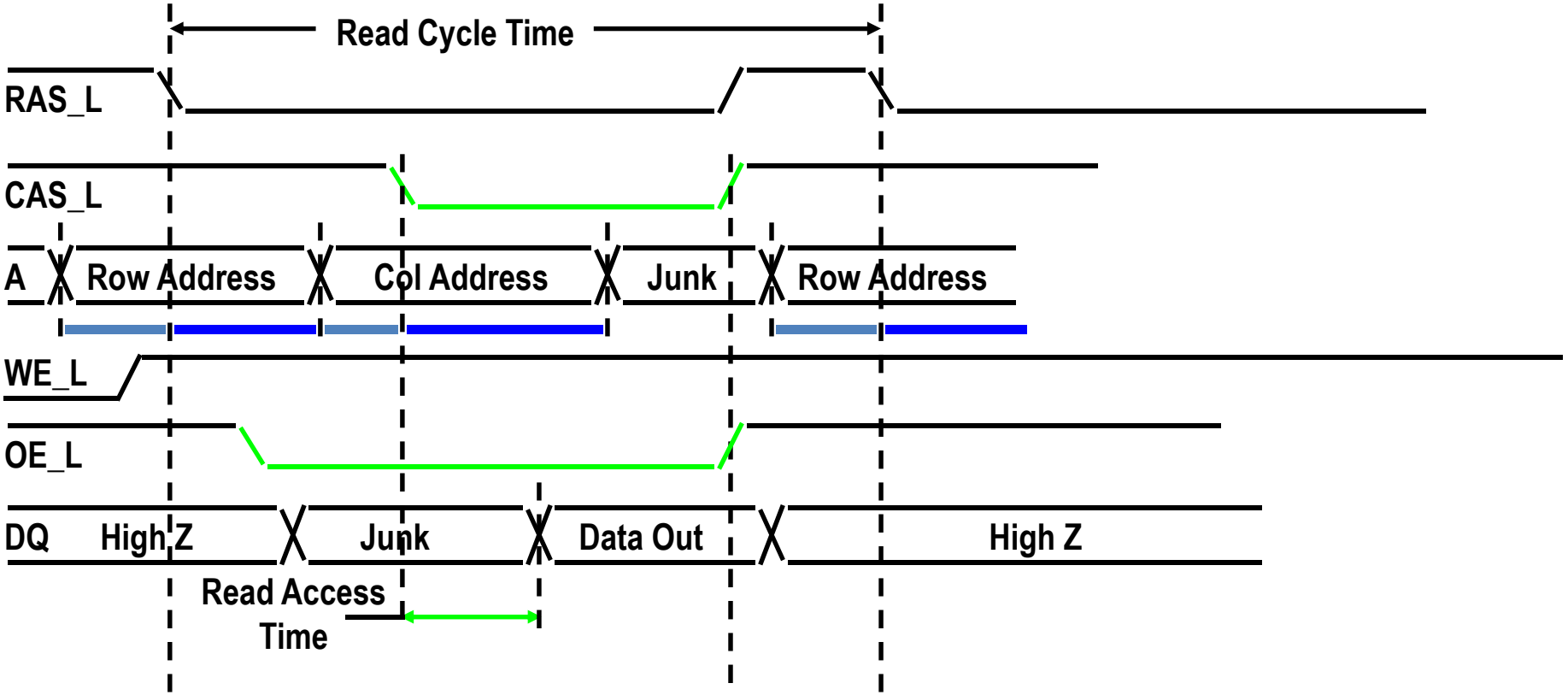
DRAM: Ciclo di scrittura

- L'accesso inizia con l'attivazione di RAS_L



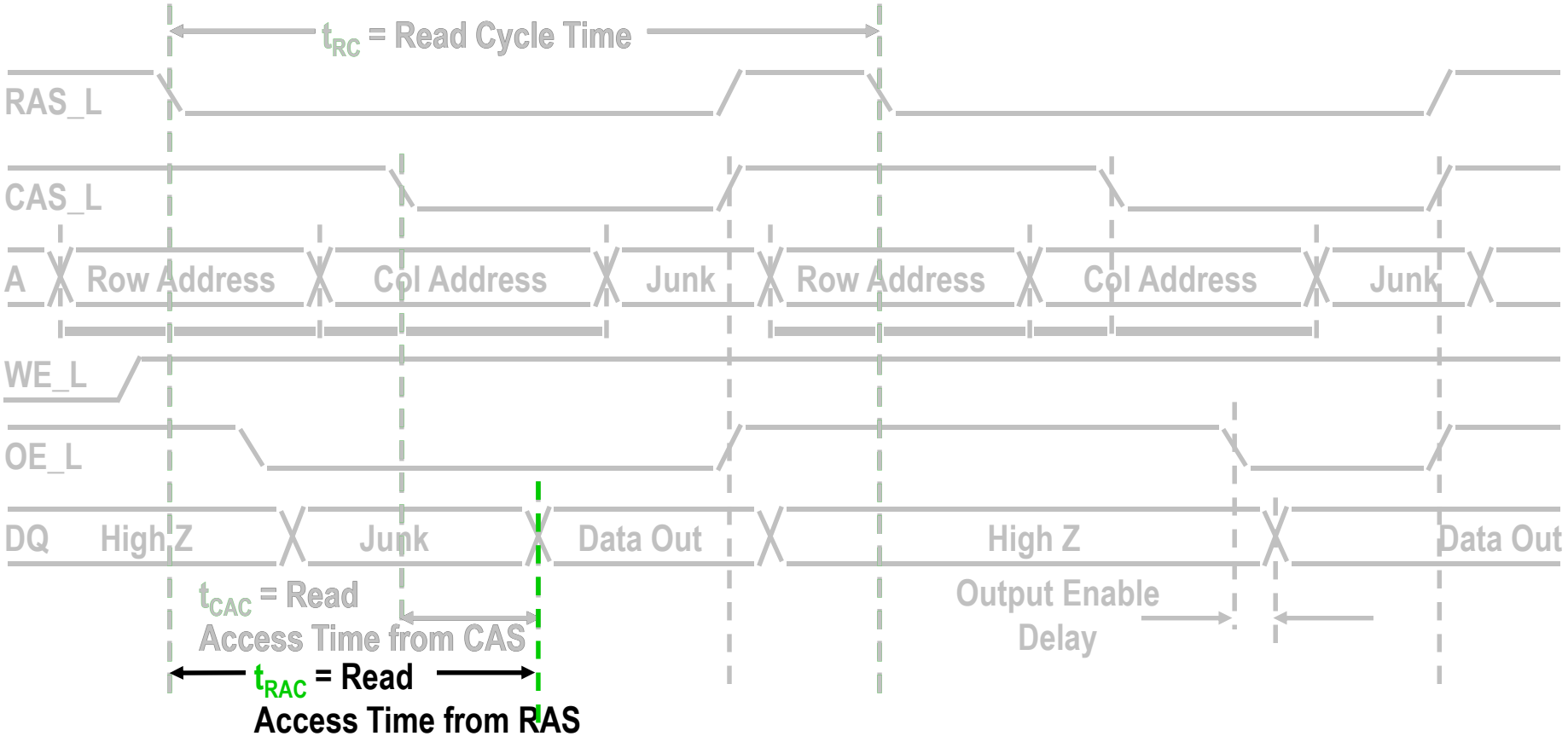
DRAM: Ciclo di lettura

- L'accesso inizia con l'attivazione di RAS_L



t_{RC} , t_{RAC} e t_{CAC}

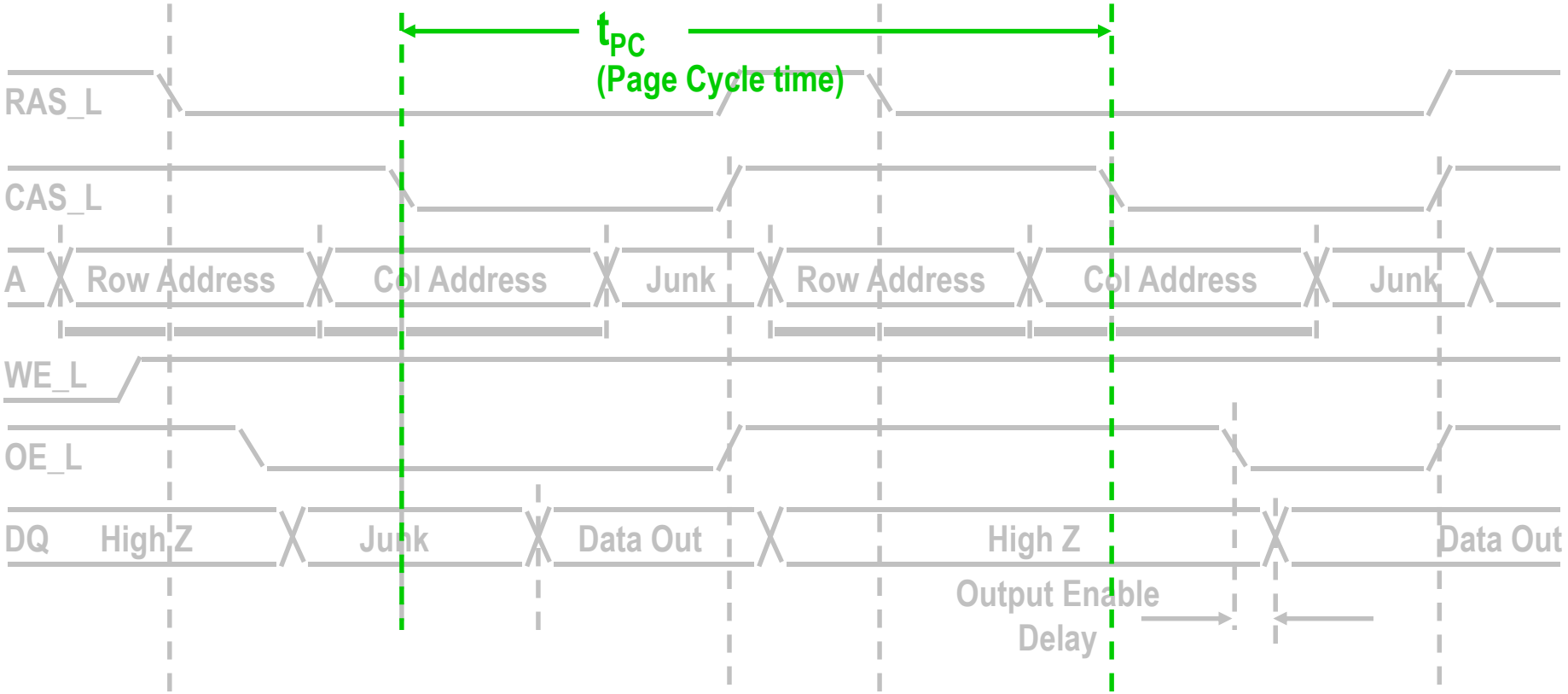
- t_{RC}: tempo minimo fra due accessi a righe
  - t_{RC} = 110 ns per una DRAM a 4Mbit
- t_{RAC} (e t_{CAC}): tempo minimo fra discesa di RAS (CAS) e dati validi
  - t_{RAC} e' data come "velocita' della DRAM" (t_{RAC}=60ns con t_{RC}=110ns)



t_{PC}

- t_{PC}: tempo minimo fra l'inizio dell'accesso a una colonna e l'inizio dell'accesso alla colonna successiva

  35 ns per una DRAM a 4Mbit e con t_{RAC} di 60 ns

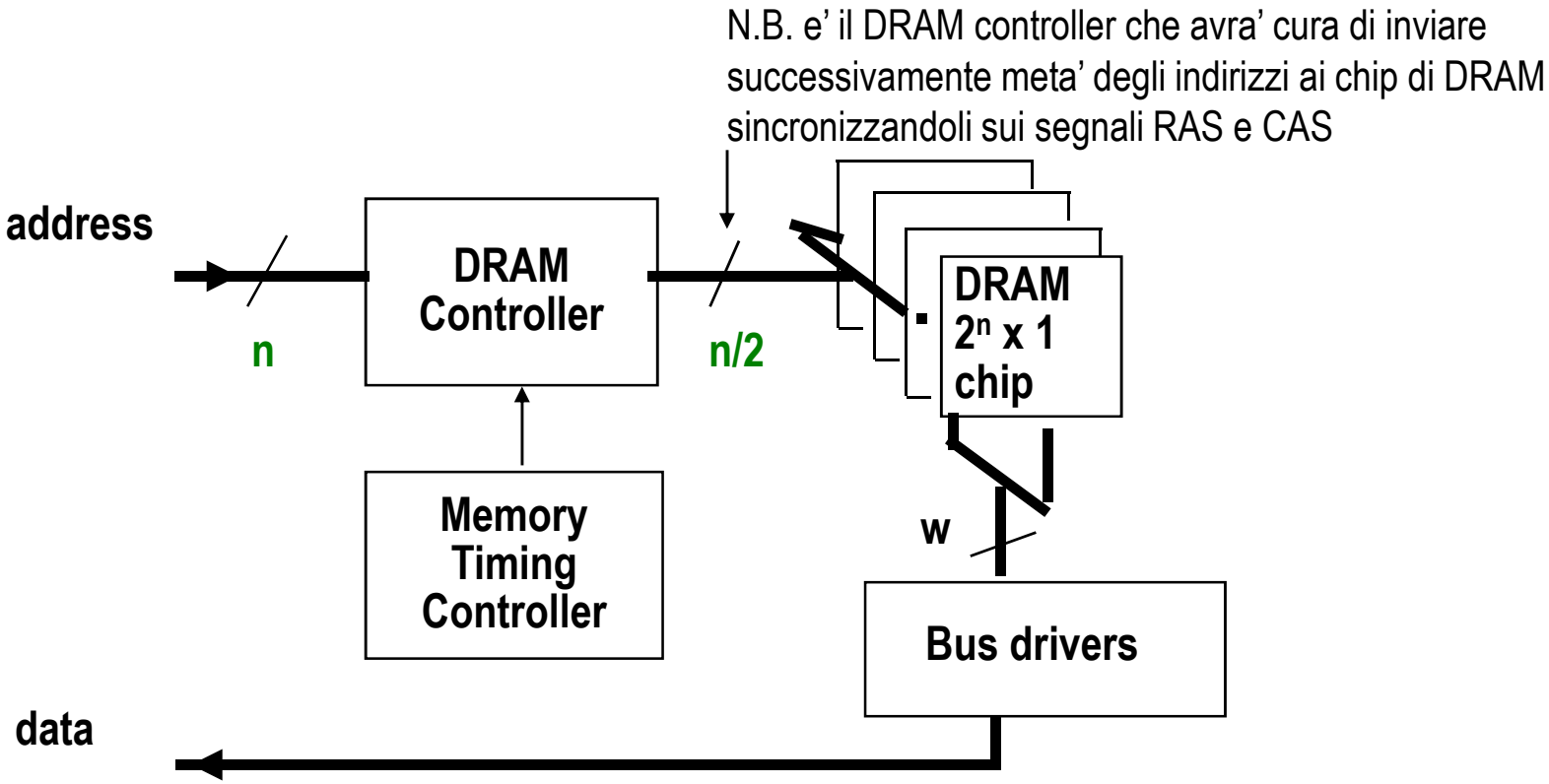


DRAM: Prestazioni

- Una DRAM da 60 ns (t_{RAC}) puo'
  - Effettuare un accesso a righe distinte ogni 110 ns (t_{RC})
  - Effettuare un accesso a colonna in (t_{CAC}) 15 ns, ma il tempo fra due accessi a colonna deve essere almeno 35 ns (t_{PC}).
    - I ritardi di indirizzamento esterni e tempi di "turnaround" sul bus fanno si che questo tempo salga a 40-50 ns

- Questi tempi non includono il tempo per inviare l'indirizzo fuori dal microprocessore (ritardo di indirizzamento) ne' il ritardo introdotto dal controller della memoria
  - Pilotaggio dei chip di DRAM, controller esterno, bus turnaround, modulo SIMM, pin, ...
  - Per una DRAM da "60 ns" (t_{RAC}) si puo' dire che e' gia' buono un tempo di accesso pari a 180-250 ns

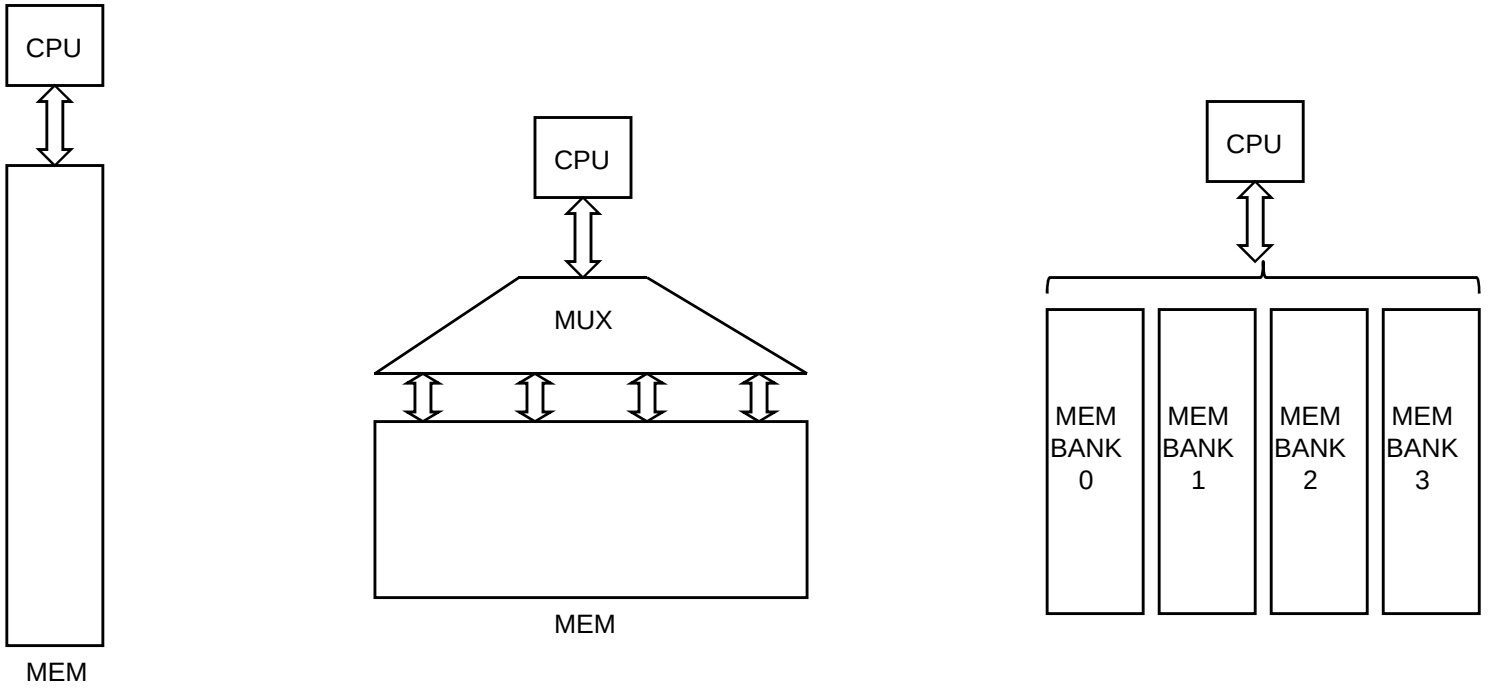
Sistema di Memoria



- Il tempo per ricevere un dato ("turnaround time") e' dato da

  Tc = Tcycle + Tcontroller + Tdriver

Tecniche per migliorare le prestazioni della memoria



Bus Semplice:

CPU, Bus, Memoria
in tutti i casi un bus
a larghezza fissa
(es. 32 bits)

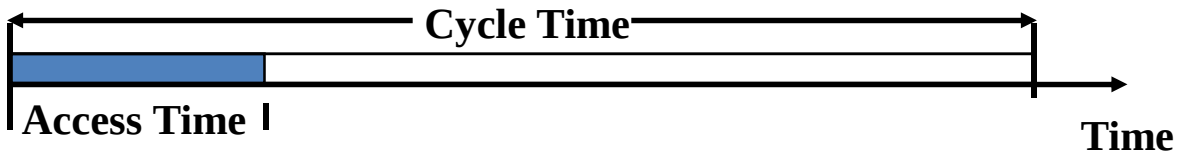
Bus largo:

CPU/Mux: 1 word
Mux/Memoria: N words
(es. 64 & 256 bits)

Interleaving:

CPU, Bus: 1 word
Memoria: N Banchi

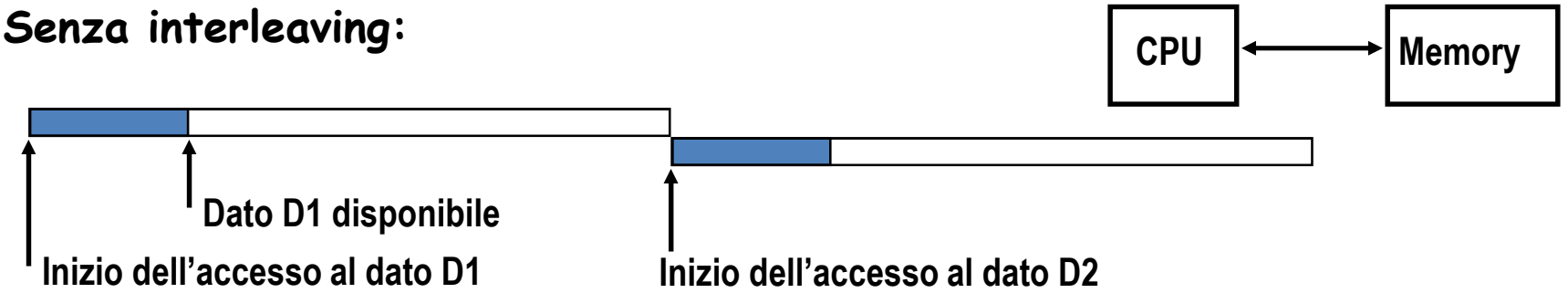
Interleaving degli accessi in memoria



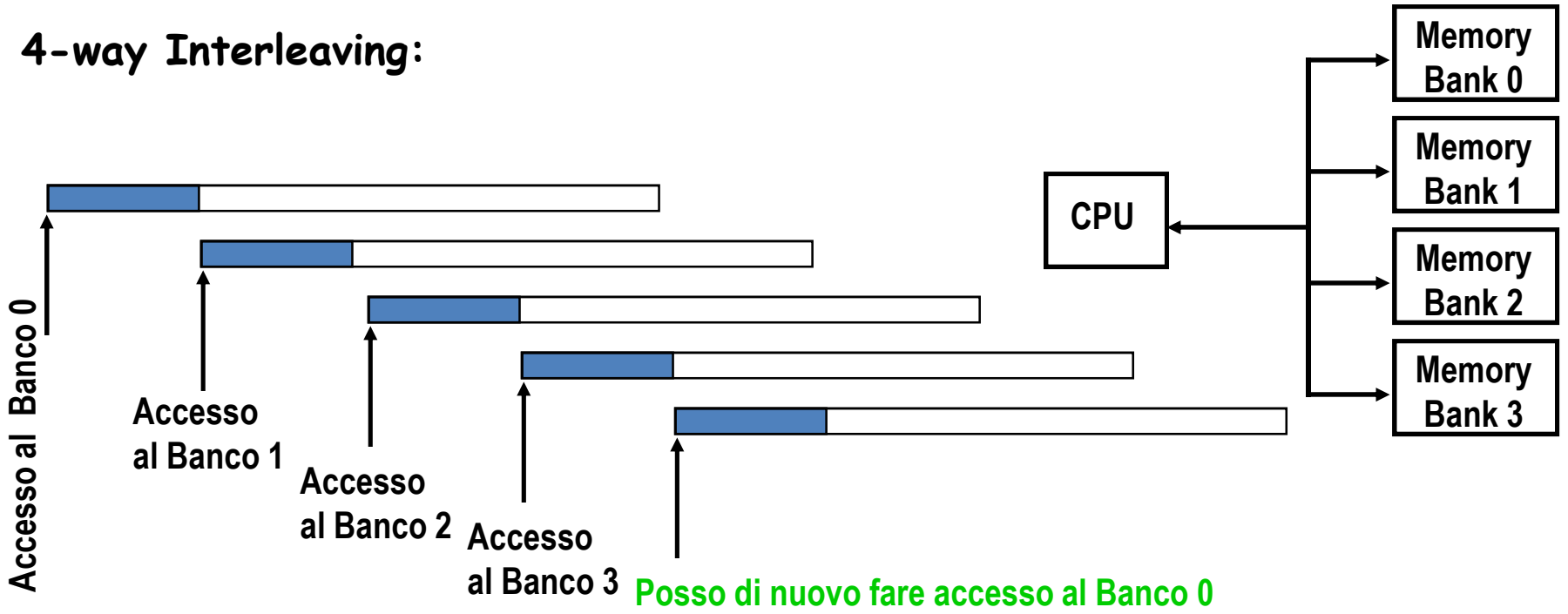
- Tempo di Ciclo della DRAM (Read/Write)
  - Indica quanto frequentemente posso fare accessi
- Tempo di Accesso della DRAM (Read/Write)
  - Indica quanto tempo occorre per ottenere il dato rispetto all'inizio dell'accesso
- Il tempo di Ciclo della DRAM (Read/Write) e' molto maggiore del suo Tempo di Accesso DRAM
  - Almeno di un fattore 2:1
- Perche' non iniziare un nuovo accesso mentre sto attendendo il dato precedente?

Increasing Bandwidth - Interleaving

Senza interleaving:



4-way Interleaving:



Confronto fra le tecniche viste

- Modello dei tempi di accesso
  - 1 ciclo per inviare l'indirizzo
  - 6 cicli per accedere al dato
  - 1 ciclo per inviare (indietro) il dato
  - Il blocco minimo di trasferimento sia pari a 4 word

• *Bus semplice*	= 4 x (1+6+1)	= 32 cicli
• *Bus largo*	= 1 + 6 + 1	= 8 cicli
• *Interleaving*	= 1 + 6 + 4x1	= 11 cicli

Scelta del numero dei banchi

- Il numero di banchi dovrebbe essere circa uguale al numero di periodi di clock per accedere ad una word di un banco
  - Questo consentirebbe di avere un dato ad ogni ciclo di clock

- Problema: la disponibilita' di DRAM a maggiore capacita' rende difficile avere piu' banchi

DRAM: diminuzione del numero di chip nel tempo

(da Pete MacWilliams, Intel)