

CALCOLATORI

La gerarchia di memoria

Marco Roveri

marco.roveri@unitn.it

*Lezione basata su materiale preparato
dai Prof. Luigi Palopoli e Giovanni Iacca*



UNIVERSITÀ DEGLI STUDI DI TRENTO

Dipartimento di Ingegneria
e Scienza dell'Informazione

Organizzazione gerarchica della memoria

- Un elaboratore senza memoria non funziona ...
- La memoria negli elaboratori non è tutta uguale
- Per decenni il sogno di ogni programmatore è stato quello di avere **tanta** memoria con accessi ultrarapidi
- Esistono compromessi tra costo, prestazioni e dimensione della memoria

Basics ...

- La memoria serve a contenere dati, bisogna poter leggere e scrivere in memoria...
- La memoria indirizzata direttamente (memoria principale, memoria cache):
 - è di tipo volatile, cioè il suo contenuto viene perso se si spegne l'elaboratore
 - è limitata dallo spazio di indirizzamento del processore
- La memoria indirizzata in modo indiretto (memoria periferica):
 - è di tipo permanente: mantiene il suo contenuto anche senza alimentazione
 - ha uno spazio di indirizzamento “software” non limitato dal processore

Basics ...

- Le informazioni nella memoria principale (indirizzamento diretto) sono accessibili al processore in qualsiasi momento
- Le informazioni nella memoria periferica (indirizzamento indiretto) devono prima essere trasferite nella memoria principale
- Il trasferimento dell'informazione tra memoria principale e memoria periferica è mediato dal software (tipicamente il Sistema Operativo)

Terminologia

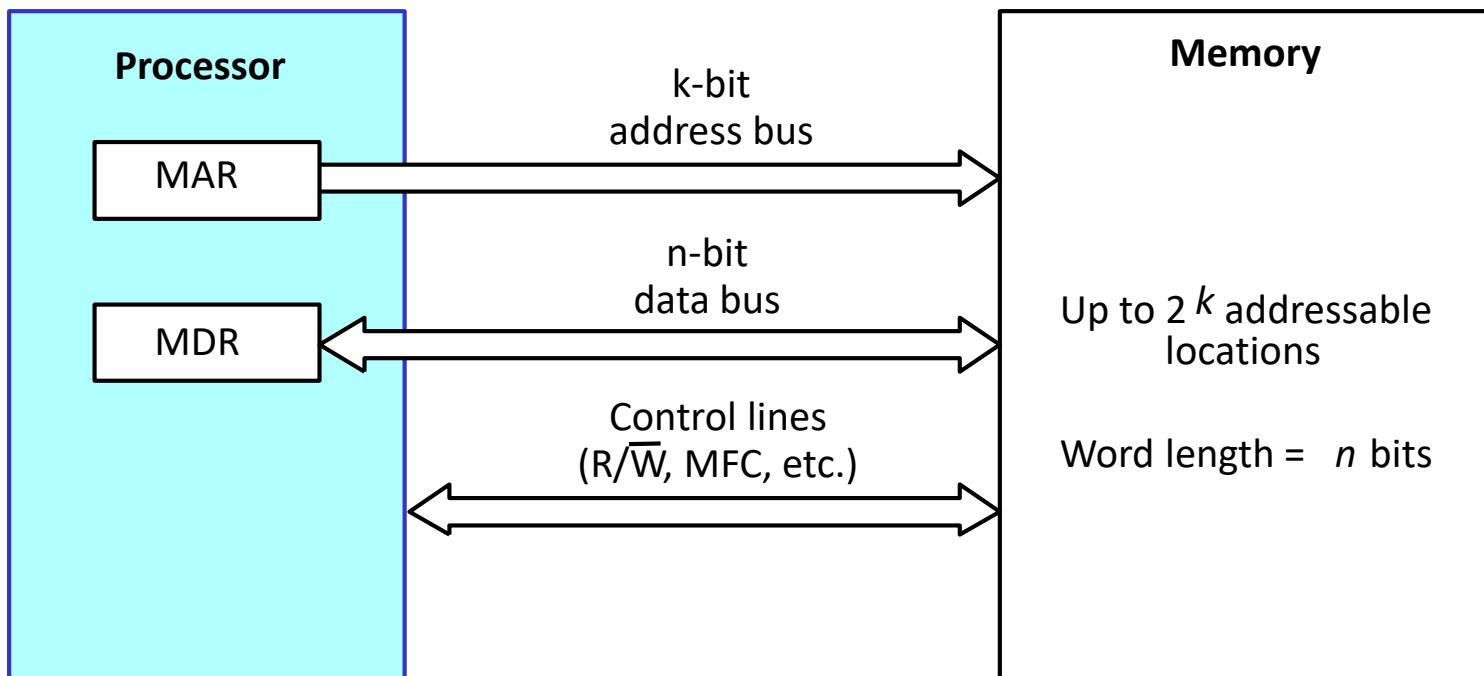
- **Tempo di accesso:** tempo richiesto per *una* operazione di lettura/scrittura nella memoria
- **Tempo di ciclo:** tempo che intercorre tra l'inizio di due operazioni consecutive (es. due read) tra locazioni diverse; in genere leggermente superiore al tempo di accesso
- **Accesso Casuale:**
 - non vi è alcuna relazione o ordine nei dati memorizzati
 - tipico delle memorie a semiconduttori

Terminologia

- **Accesso Sequenziale:**
 - l'accesso alla memoria è ordinato o semi-ordinato
 - il tempo di accesso dipende dalla posizione
 - tipico dei dischi e dei nastri
- **RAM: Random Access Memory**
 - memoria scrivibile/leggibile a semiconduttori
 - tempo di accesso indipendente dalla posizione dell'informazione
- **ROM: Read Only Memory**
 - memoria a semiconduttori in sola lettura
 - accesso casuale o sequenziale

Memoria principale

- Connessione “logica” con il processore



MAR: Memory Address Register

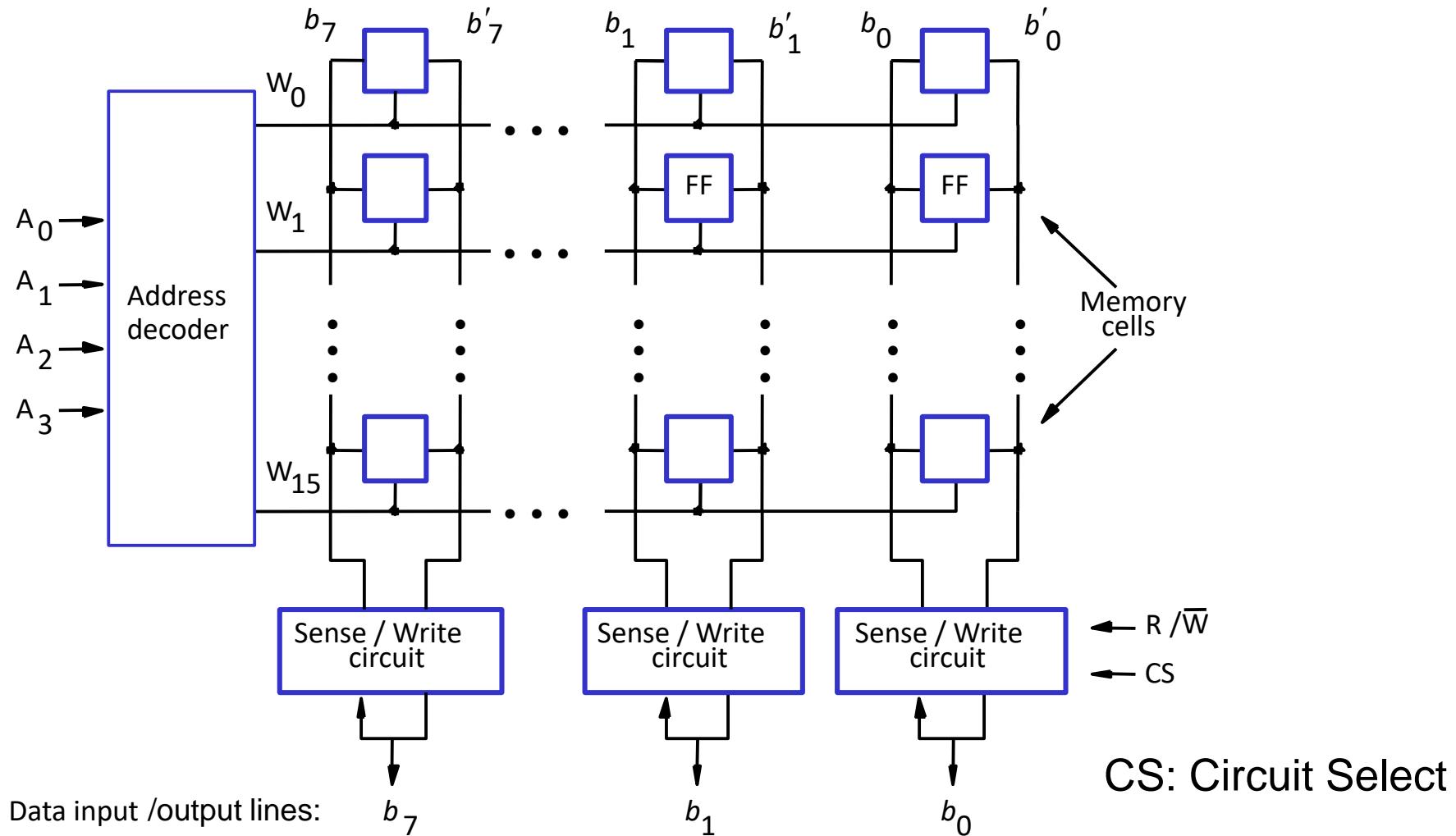
MDR: Memory Data Register

MFC: Memory Function Completed

Memorie RAM a semiconduttori

- Memorizzano singoli bit, normalmente organizzati in byte e/o word
- Data una capacità N (es. 512 Kbit) la memoria può essere organizzata in diversi modi a seconda del parallelismo P (es. 1, 4, 8)
 - 512K X 1
 - 128K X 4
 - 64K X 8
- L'organizzazione influenza il numero di pin di I/O del circuito integrato (banco) che implementa la memoria

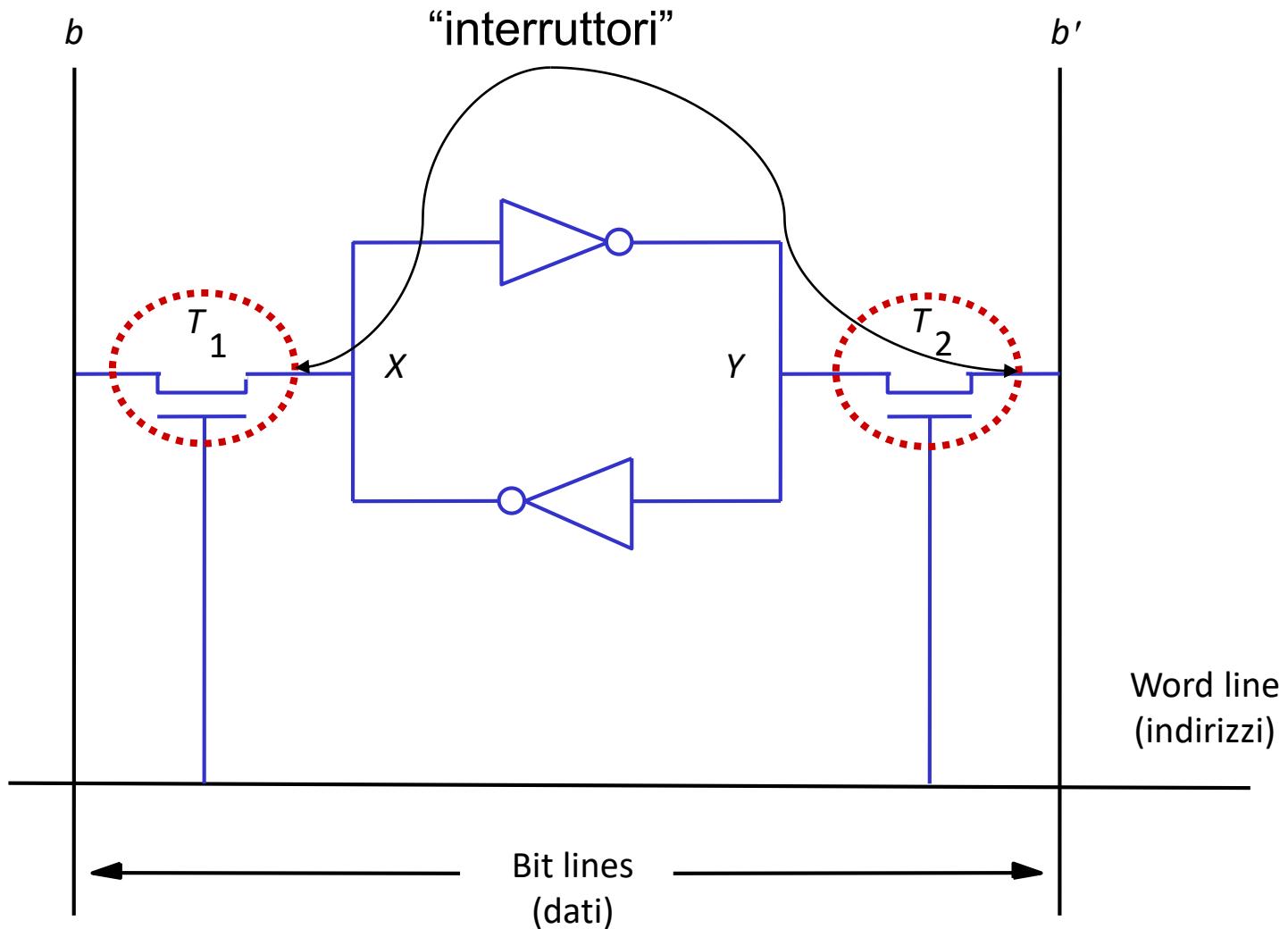
Organizzazione dei bit in un banco di memoria 16 X 8



Memorie Statiche (SRAM)

- Sono memorie in cui i bit possono essere tenuti indefinitamente (posto che non manchi l'alimentazione)
- Estremamente veloci (tempo di accesso di pochi ns)
- Consumano poca corrente (e quindi non scaldano)
- Costano care perché hanno molti componenti per ciascuna cella di memorizzazione

SRAM: cella di memoria



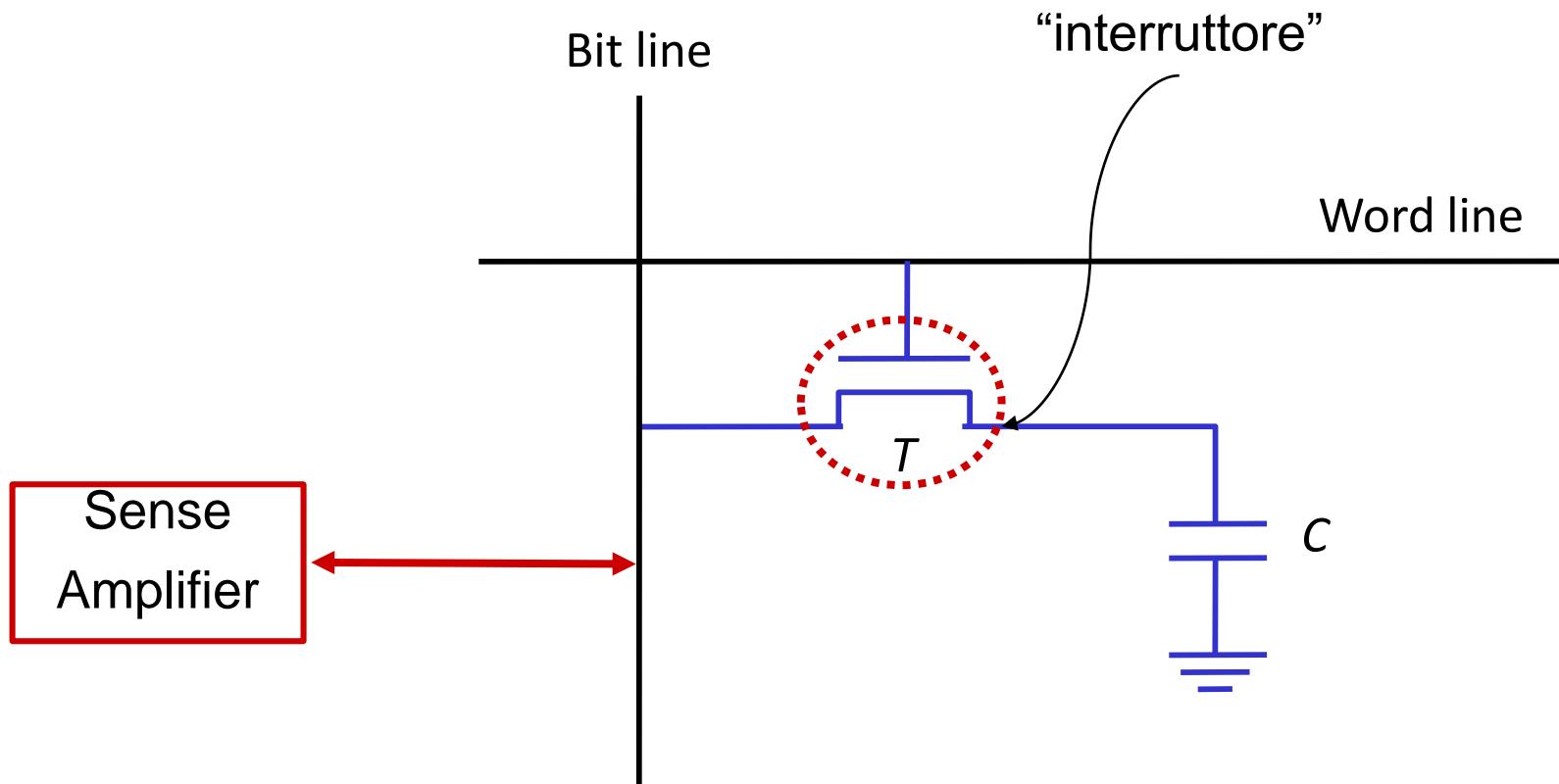
SRAM: lettura e scrittura

- $b' = \text{NOT}(b)$: i circuiti di terminazione della linea di bit (sense/write circuit) interfacciano il mondo esterno che non accede mai direttamente alle celle
- La presenza contemporanea di b e $\text{NOT}(b)$ consente un minor tasso di errori
- **Scrittura:** la linea di word è alta e chiude T_1 e T_2 ; il valore presente su b e b' , che funzionano da linee di pilotaggio, viene memorizzato nel latch a doppio NOT
- **Lettura:** la linea di word è alta e chiude T_1 e T_2 , le linee b e b' sono tenute in stato di alta impedenza: il valore nei punti X e Y viene “copiato” su b e b'
- Se la linea di word è bassa, T_1 e T_2 sono interruttori aperti: il consumo è praticamente nullo

RAM dinamiche (DRAM)

- Sono le memorie più diffuse nei PC e simili
- Economiche e a densità elevatissima (in pratica 1 solo componente per ogni cella)
 - la memoria viene ottenuta sotto forma di carica di un condensatore
- Hanno bisogno di un aggiornamento (*refresh*) continuo del proprio contenuto che altrimenti “svanisce” a causa delle correnti parassite
- Consumi elevati a causa del rinfresco continuo

DRAM: cella di memoria



DRAM: lettura e scrittura

- **Scrittura:** la linea di word è alta e chiude T, il valore presente su b viene copiato sul condensatore C (carica il transistor)
- **Lettura:** la linea di word è alta e chiude T, un apposito circuito (**sense amplifier**) misura la tensione su C
 - se è sopra una certa soglia data, pilota la linea b alla tensione nominale di alimentazione, ricaricando C
 - se è sotto la soglia data, mette a terra la linea b scaricando completamente il condensatore C

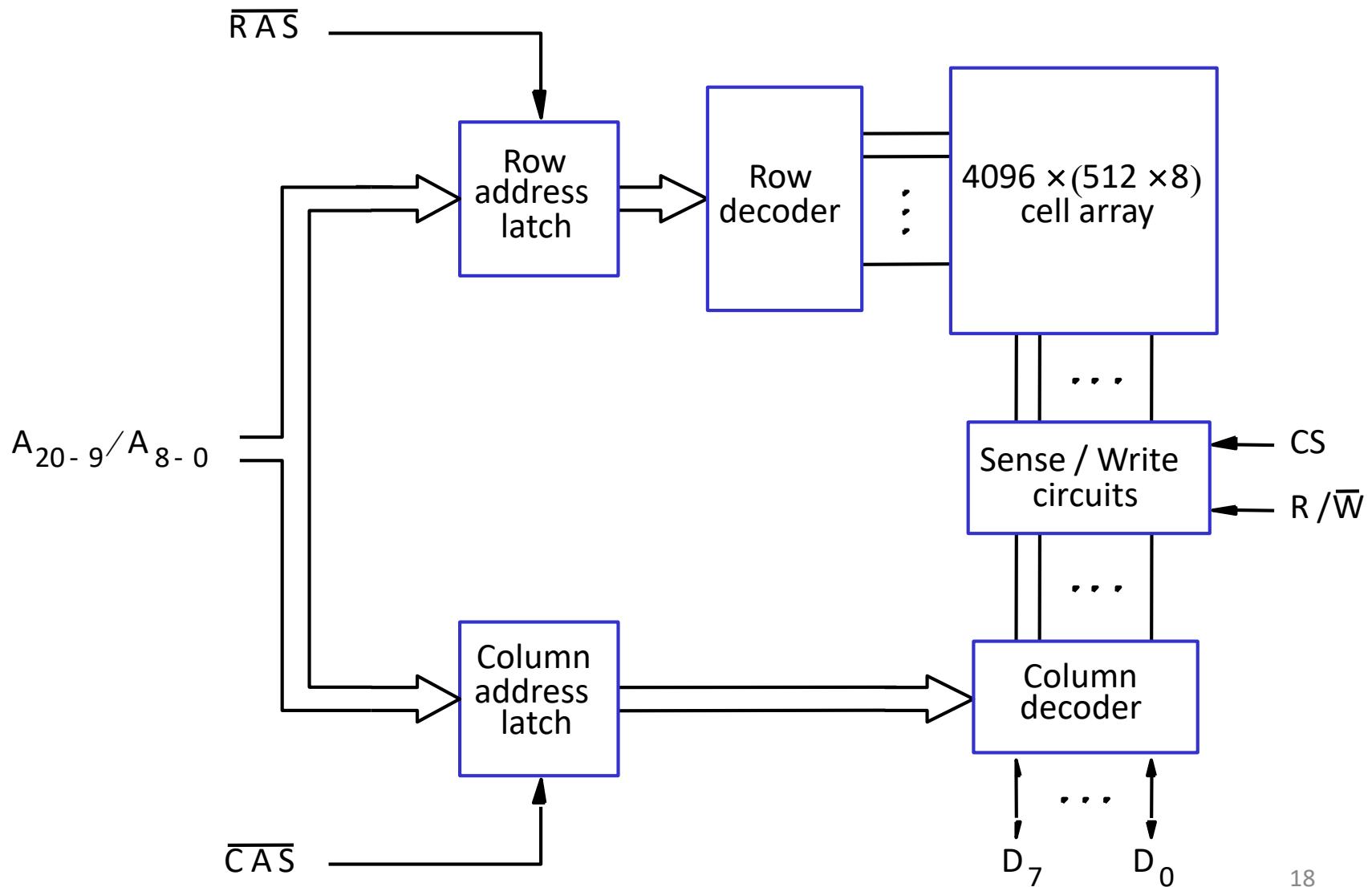
DRAM: tempi di rinfresco

- Nel momento in cui T viene aperto, il condensatore C comincia a scaricarsi (o caricarsi, anche se più lentamente, a causa delle resistenze parassite dei semiconduttori)
- E' necessario “rinfrescare” la memoria prima che i dati “spariscano”
basta fare un ciclo di lettura
- In genere il chip di memoria contiene un circuito per il refresh (lettura periodica di tutta la memoria); l’utente non si deve preoccupare del problema

DRAM: multiplazione degli indirizzi

- Data l'elevata integrazione delle DRAM il numero di pin di I/O è un problema
- E' usuale multiplicare nel tempo l'indirizzo delle righe e delle colonne negli stessi fili
- Normalmente le memorie non sono indirizzabili al bit, per cui righe e colonne si riferiscono a byte e non a bit
- Es. una memoria 2M X 8 (21 bit di indirizzo) può essere organizzata in 4096 righe (12bit di indirizzo) per 512 colonne (9bit di indirizzo) di 8 bit ciascuno

Organizzazione di una DRAM 2M X 8



DRAM: modo di accesso veloce

- Spesso i trasferimenti da/per la memoria avvengono a blocchi (o pagine)
- Nello schema appena visto, vengono selezionati prima 4096 bytes e poi tra questi viene scelto quello richiesto
- E' possibile migliorare le prestazioni semplicemente evitando di "riselezionare" la riga ad ogni accesso se le posizioni sono consecutive
- Questo viene chiamato "fast page mode" (FPM) e l'incremento di prestazioni può essere significativo