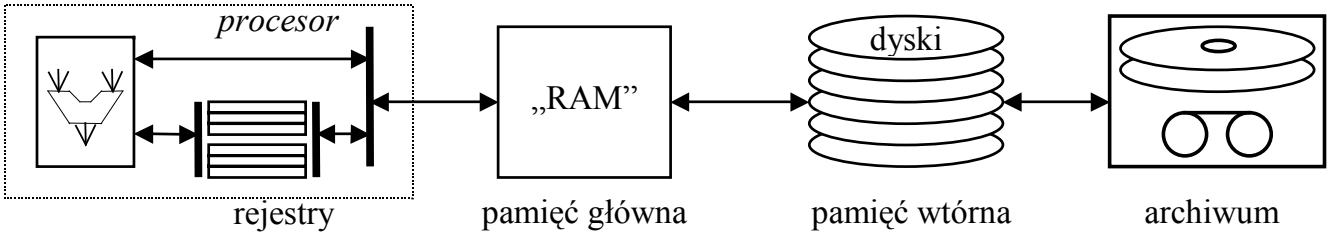
Hierarchia pamięci

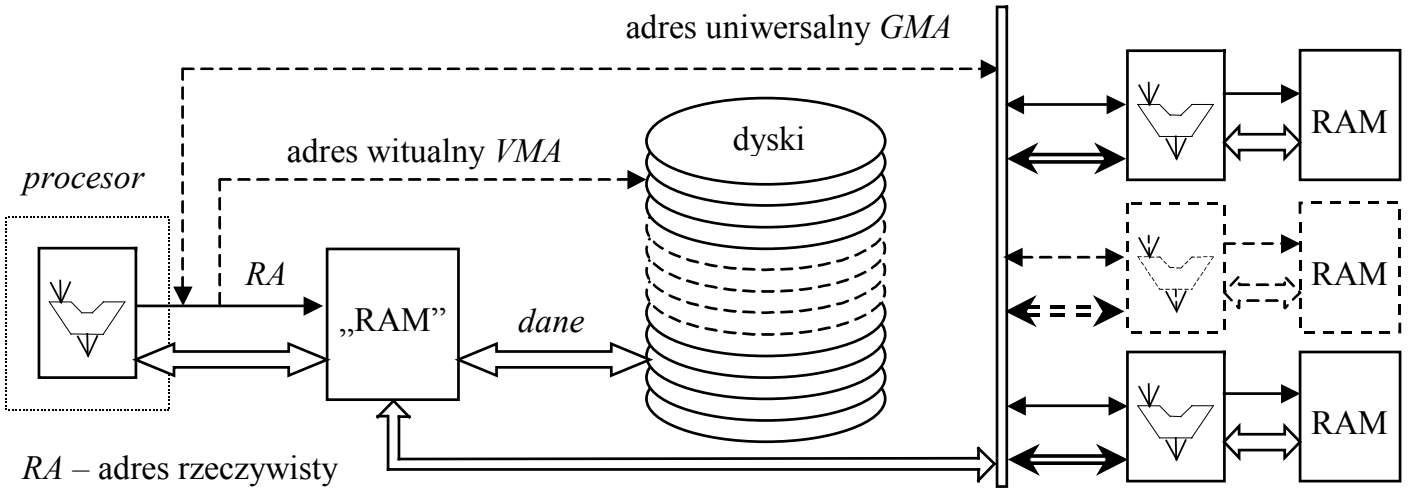
zasada działania komputera – pamięć: jedyne źródło danych
→ konieczna komunikacja procesora z pamięcią główną
→ przepustowość pamięci ogranicza szybkość przetwarzania (*memory bottleneck*)

szybki procesor & wolna pamięć główna
→ procesor wytwarza dodatkowe *cykle oczekiwania* (ang. *wait state*, WS)



czas dostępu	0,25–1 ns	5–20 ns	1–10 ms	—
pojemność	512B–4 kB	1GB–16GB	>400 GB	>> 500 GB
pobór mocy	duży	mały	bardzo mały	bardzo mały
	register	*memory*	*storage*	*archive*

Adresowanie pamięci



pamięć główna – adres rzeczywisty (*real address*, RA)

- nie wymaga etykiety (RAM – zbiór uporządkowany)

pamięć wtórna – adres wirtualny (*virtual memory address*, VMA)

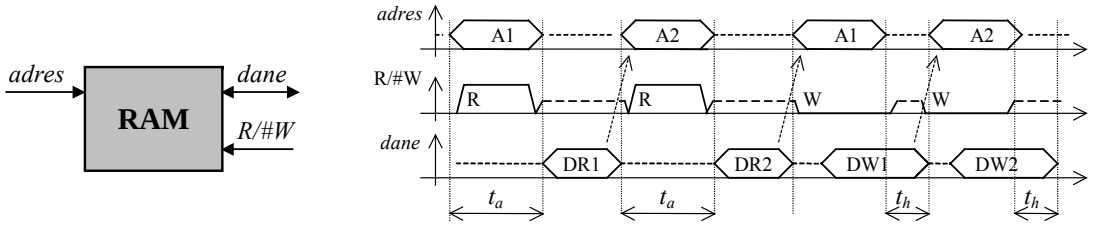
- konieczna etykieta identyfikująca blok oraz adres wewnątrz bloku

wszelkie pamięci świata – adres uniwersalny (*general memory address*, GMA)

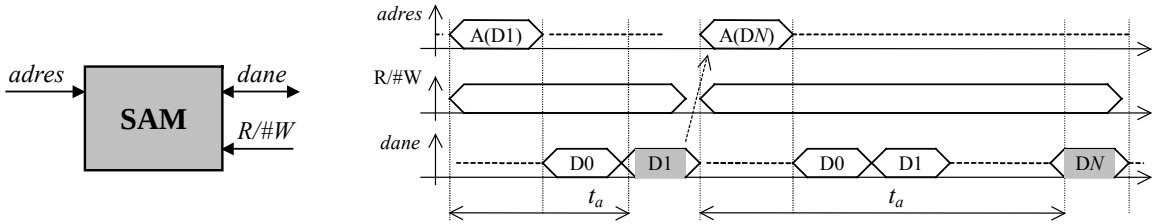
- konieczna etykieta identyfikująca obiekt i protokół transmisji

Organizacja i obsługa pamięci

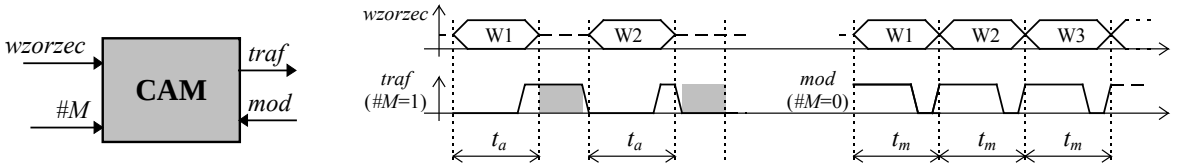
- pamięć o dostępie swobodnym (*random access memory*, RAM)



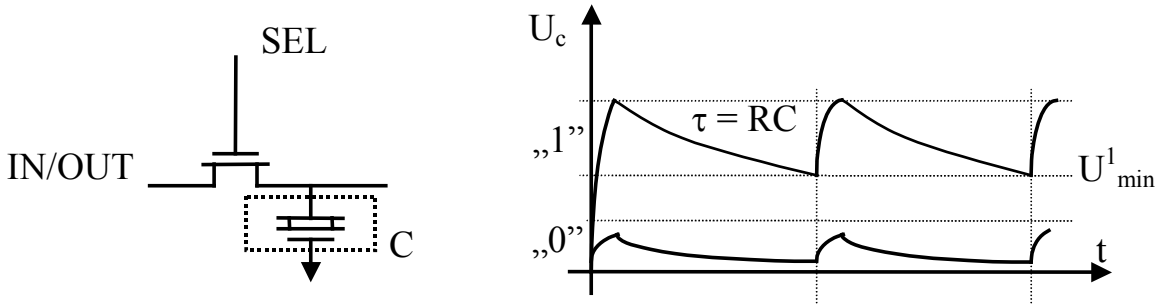
- pamięć o dostępie sekwencyjnym (*sequentially accessible memory*, SAM)



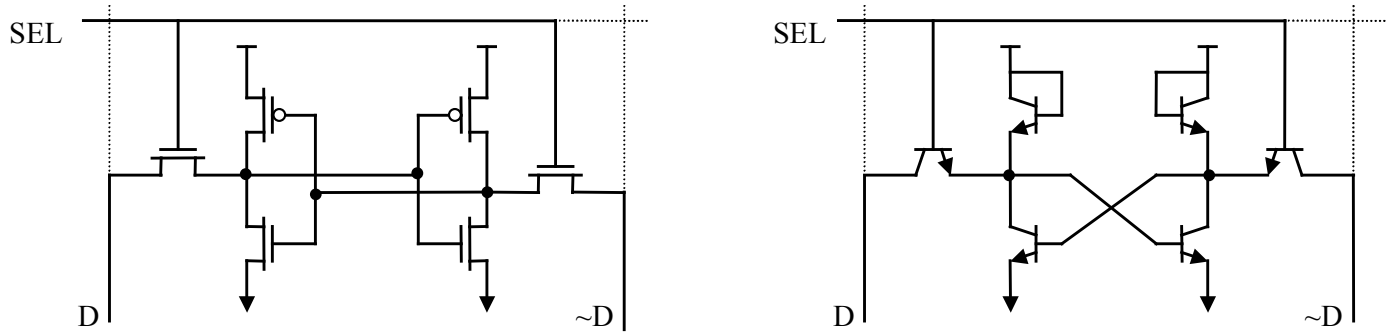
- pamięć adresowana zawartością (*contents addressable memory*, CAM)



Komórka pamięci półprzewodnikowej



Komórka pamięci dynamicznej (DRAM) i napięcie na pojemności C
(1 tranzystor/bit, mały pobór mocy – tylko podczas odświeżania)



Komórka pamięci statycznej CMOS i TTL (SRAM)
(4-6 tranzystorów/bit, duży pobór mocy – stale załączona para tranzystorów)

Pamięć o dostępie swobodnym

czas dostępu do danych *nie zależy od lokalizacji* (*random access*)

- zapisywalno-odczytywalne (RWM, *read-write memory*) – RAM
  - statyczne (SRAM)
  - dynamiczne (DRAM)
    - z utrzymaniem wyjść (EDO, *extended data output*)
    - synchroniczne (SDRAM, *synchronous DRAM*)
    - podwójnej szybkości (DDR, ..., DDR4, d*ouble data rate RAM*)
    - pseudo-dwuportowe (GDDR5) – do kart graficznych
  - nieulotne RAM (NVRAM, *non-volatile* RAM)
    - z podtrzymaniem (CMOS-RAM)
- tylko odczytywalne (ROM, *read-only memory*)
  - stałe – zawartość nadana podczas wytwarzania
  - reprogramowalne – zapisywane silnymi impulsami elektrycznymi
    - kasowane promieniowaniem ultrafioletowym (EPROM)
    - kasowane impulsem elektrycznym (EEPROM)
      - typu FLASH – programowalne w układzie (on-line)

Pamięć o dostępie sekwencyjnym

czas dostępu do danych *zależy od lokalizacji*

- magnetyczna
  - taśmowa (*streamer*)
  - dyskowa
    - dyski elastyczne (*floppy*)
    - zestawy dysków „sztywnych" (*hard disk assembly*)
- optyczna
  - stała CD-ROM (*compact disk ROM*)
  - archiwalna CD-WR, WORM (*write once read many*)
  - magnetooptyczna CD-RW, CD-RAM, WREM (*write read erase memory*)
  - uniwersalna DVD (*digital versatile disk*)
  - dużej gęstości BD (*Blue-ray Disk*) – niebieski laser (405 nm), do 200 GB

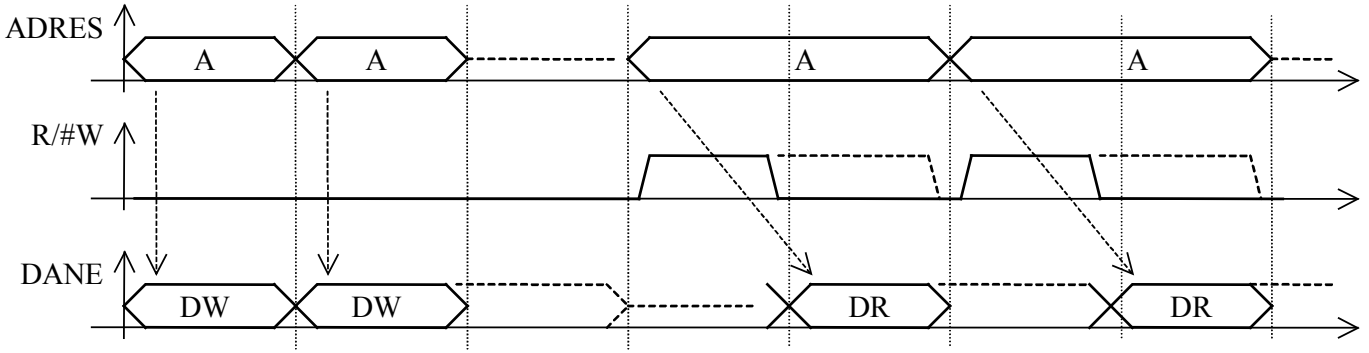
→ **pamięć o dostępie bezpośrednim** DAM (*direct access memory*)

dane z nośnika sekwencyjnego → bufor SIPO → odczyt równoległy

blok danych → bufor PISO → zapis sekwencyjny

Organizacja i obsługa statycznej pamięci RAM

- odczyt – dwa cykle pamięci: stabilny adres → transfer danych
- zapis – jeden cykl pamięci: stabilny adres & transfer danych



Cykle pamięci statycznej

dostęp seryjny do danych w jednym wierszu

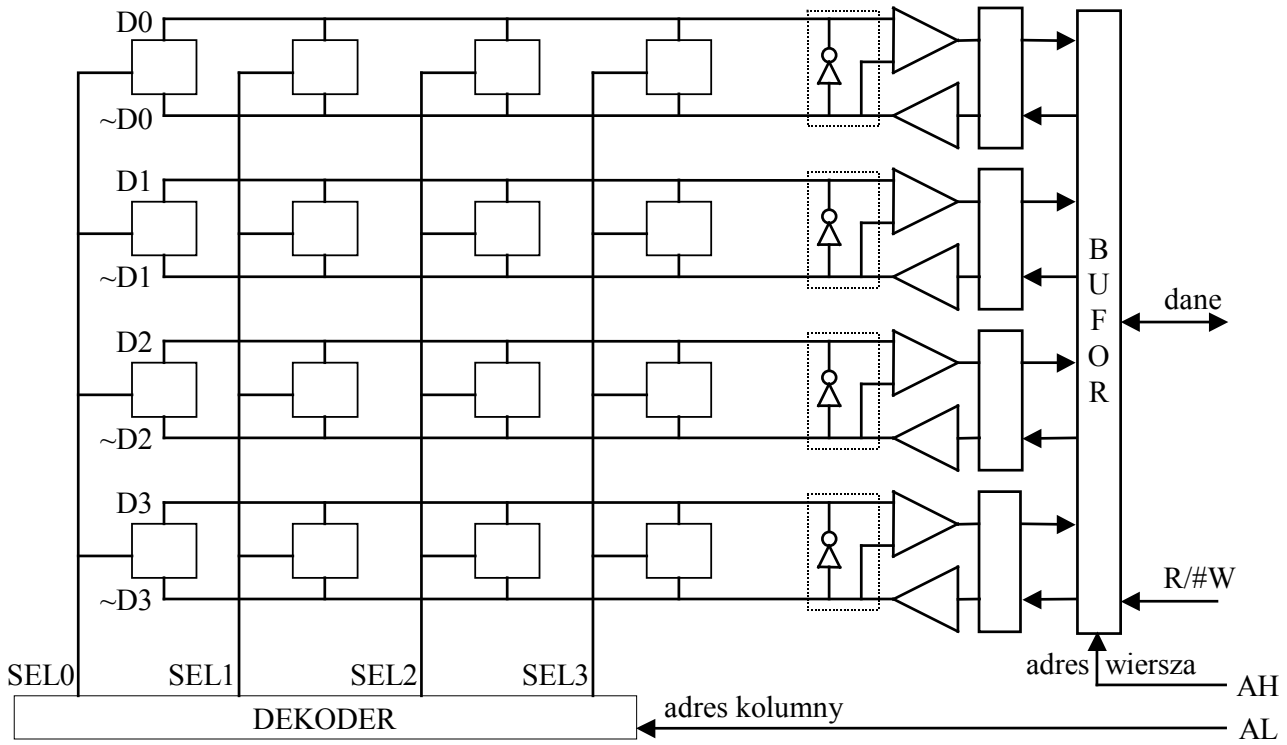
→ zatrzaśnięcie adresu na czas transferu i buforowanie linii danych.

skrócenie czasu dostępu (EDO RAM)

→ utrzymanie danej po zaniku adresu → szybsze zmiany adresów

→ seryjny dostęp do danych w różnych modułach

Matryca pamięci statycznej



Organizacja i obsługa pamięci statycznej

Organizacja i obsługa pamięci dynamicznej (DRAM)

odświeżanie – okresowo (τ=RC=0,25–1ms / f_R= 1–4k/s)

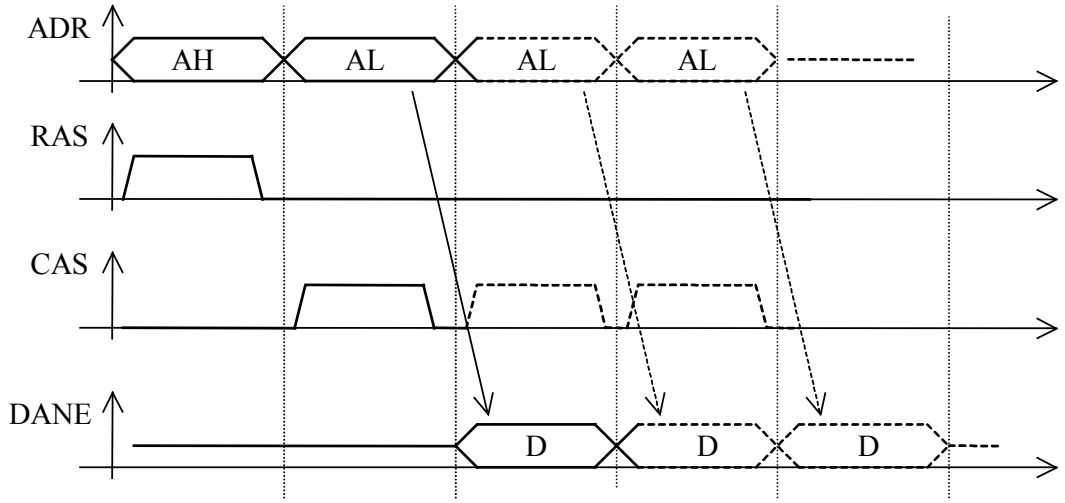
rozładowanie (odczyt) →wzmocnienie→przeładowanie (zapis)

odczyt

rozładowanie (**odczyt**) →wzmocnienie→przeładowanie (zapis)

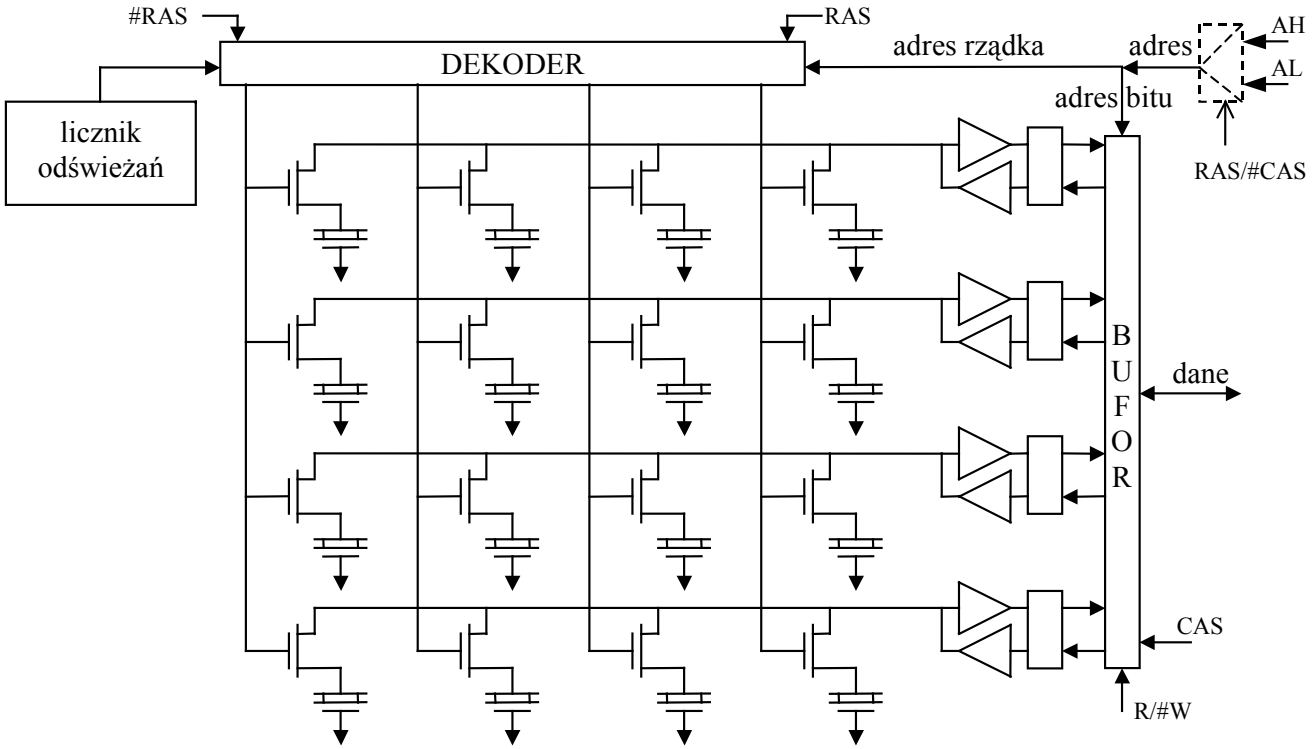
zapis

rozładowanie (odczyt) →**wymuszenie**→przeładowanie (**zapis**)



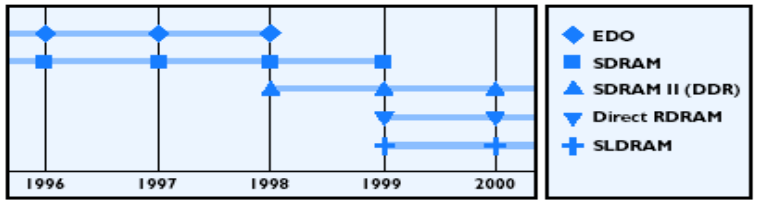
Cykle pamięci dynamicznej: AH – adres rządka, AL – adres kolumny

Matryca pamięci dynamicznej



Organizacja pamięci dynamicznej.
RAS – strob adresu rządka, CAS – strob adresu danej

Ewolucja architektury i technologii pamięci



Source: Toshiba, Intel, and Rambus

Standard JEDEC (Joint Electronic Devices Engineering Council)

EDO – Extended Data Out memory, możliwość adresowania kolejnej lokacji przed zakończeniem poprzedniego transferu

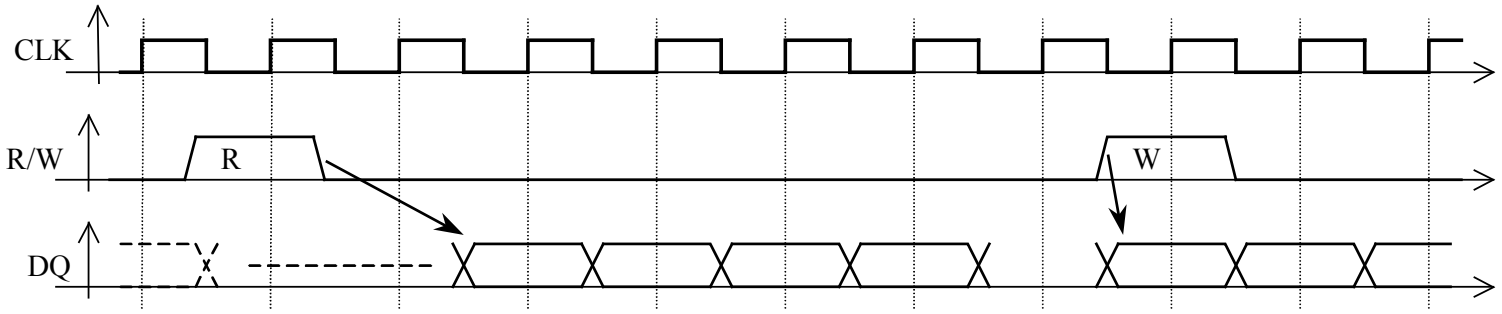
SDRAM (synchronous DRAM) – synchronizacja wejścia i wyjścia, 4-banki pamięci

DDR (double-data rate SDRAM), **SDRAM II** – szybsza wersja SDRAM umożliwiająca odczyt danych na obu zboczach CLK z synchronizowanych banków,

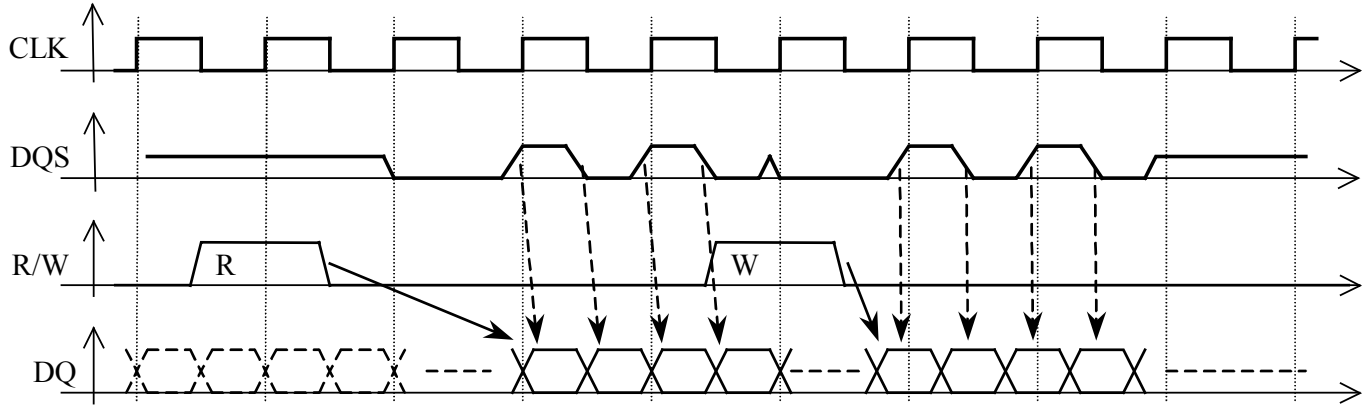
RDRAM® (Rambus™ DRAM) – zwiększona przepustowość wewnętrzna
SLDRAM (Synclink DRAM) – 16 banków pamięci, nowy interface i logika sterująca
QBM – 2x**DDR** z synchronizacją przesuniętą o ćwierć okresu

DDR3 – DDR 3. generacji, 1,5/1,35V, 0,8-2,4 Gb/s, moduł DIM max 16GB
DDR4 – DDR 4. generacji, 1,2V, 1,6-3,2 Gb/s, moduł DIM max 128GB

Pamięci SDRAM i DDR

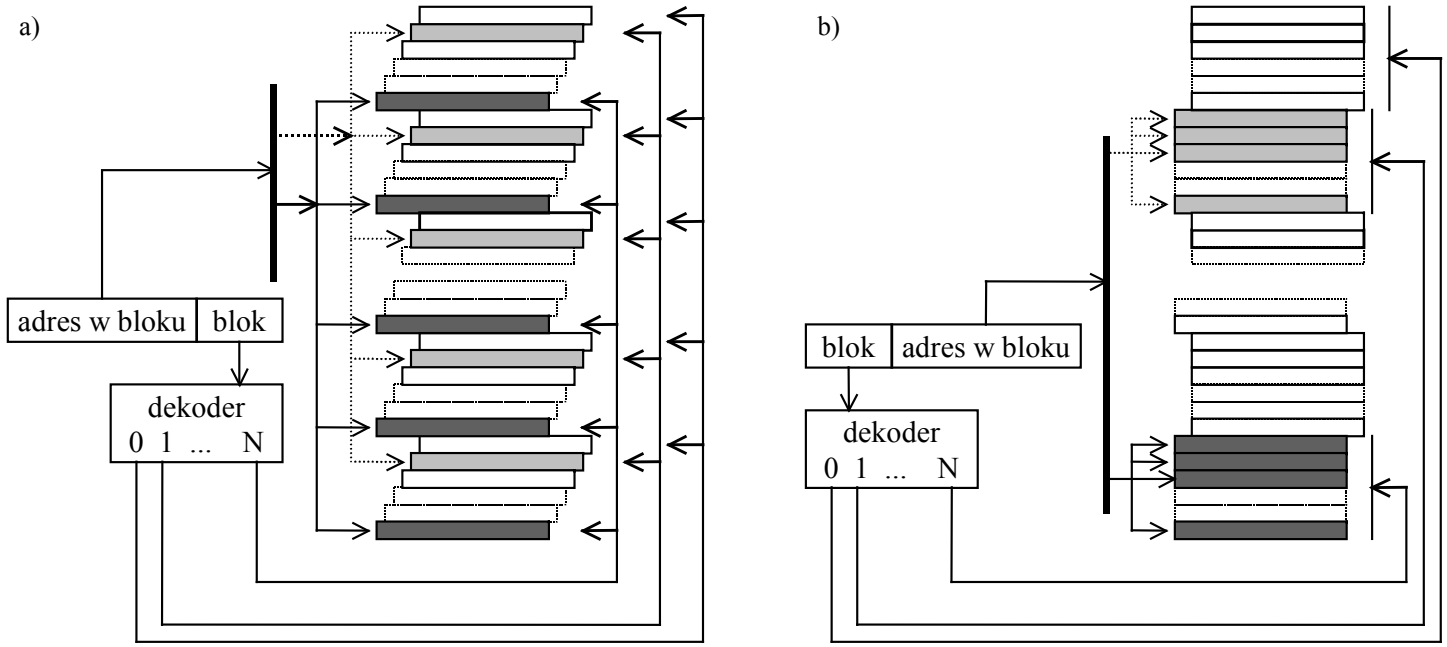


Działanie pamięci SDRAM



Działanie pamięci DDR (DQS – strob danych)

Organizacja pamięci głównej



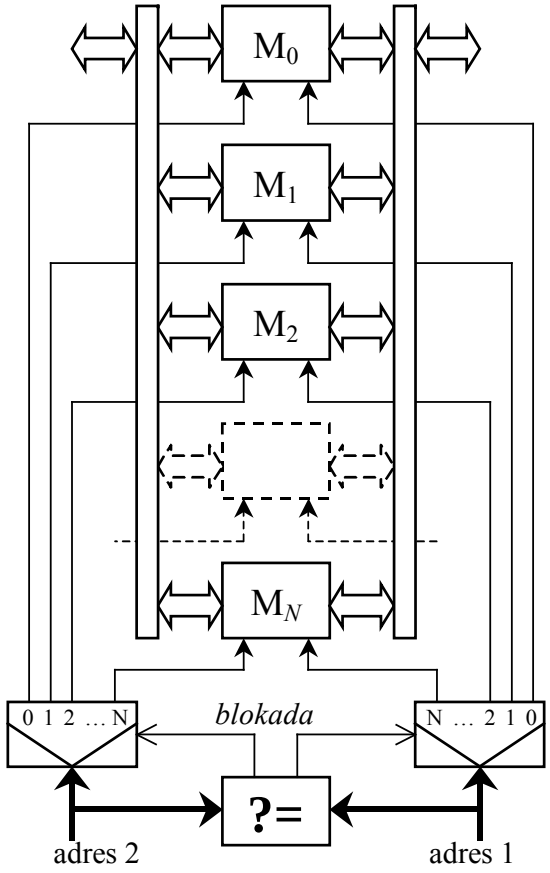
przeplot bloków (*low order interleave*)

- dostęp do słów bloku *jednoczesny*
- trudna kontrola bloku
  - bity parzystości bajtów/słów
  - dodatkowy blok kontrolny

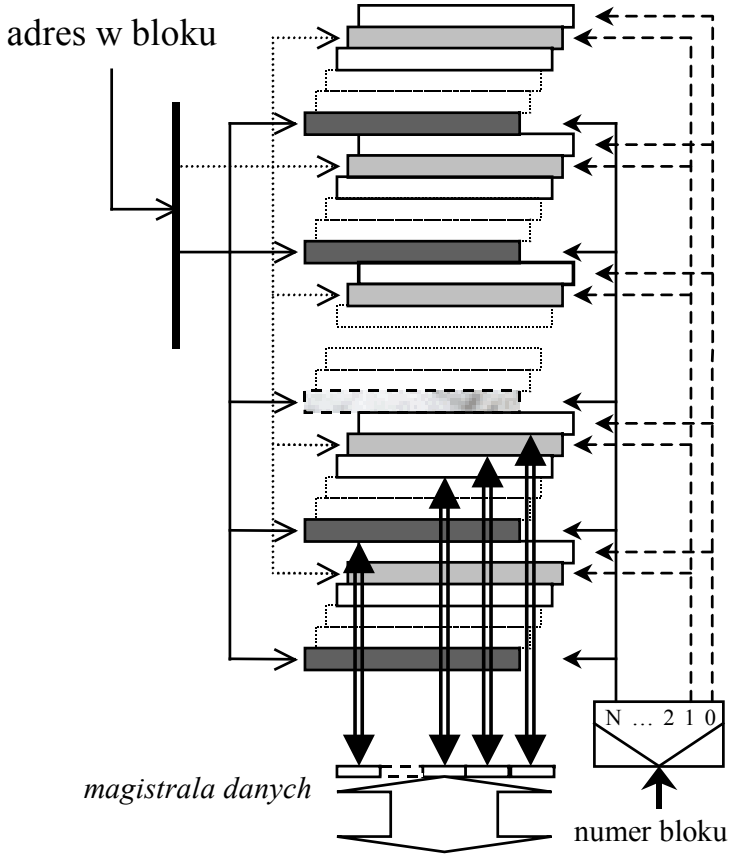
przeplot słów (*high order interleave*)

- dostęp do słów bloku *sekwencyjny*
- łatwa kontrola bloku
  - bity parzystości bajtów/słów
  - (parzystość skrośna – kod prostokątny

Współbieżny dostęp do pamięci

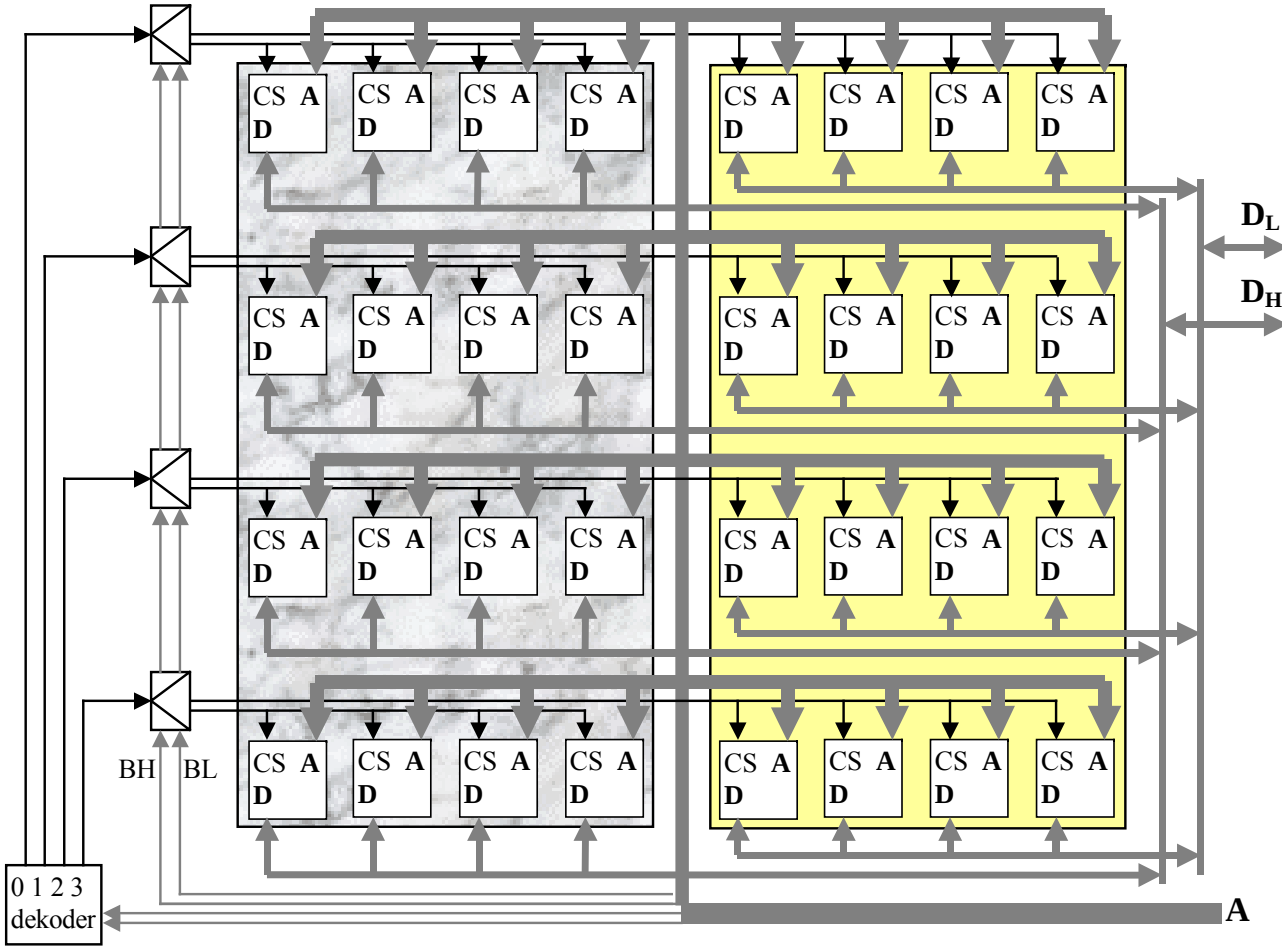


pamięć dwudrożna

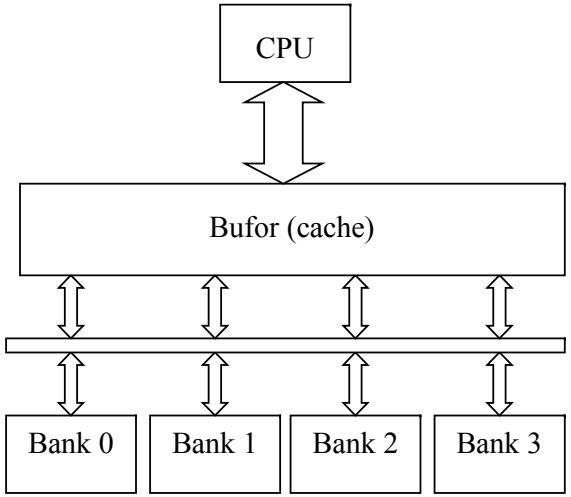


dostęp jednoczesny

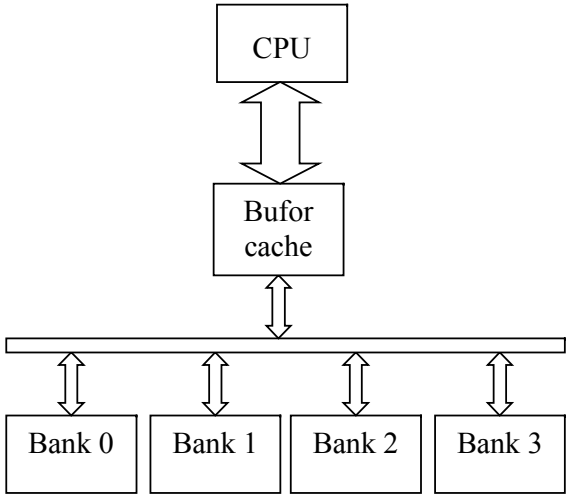
Banki pamięci głównej



Banki pamięci – organizacja dostępu



szeroka magistrala



transfery nakładane

- większy koszt
- większa przepustowość
- łatwa kontrola

- mniejszy koszt
- mniejsza przepustowość
- trudniejsza kontrola

Pamięć o dostępie sekwencyjnym

nośnik informacji

- pamięć magnetyczna
  domena magnetyczna (mini-dipol)

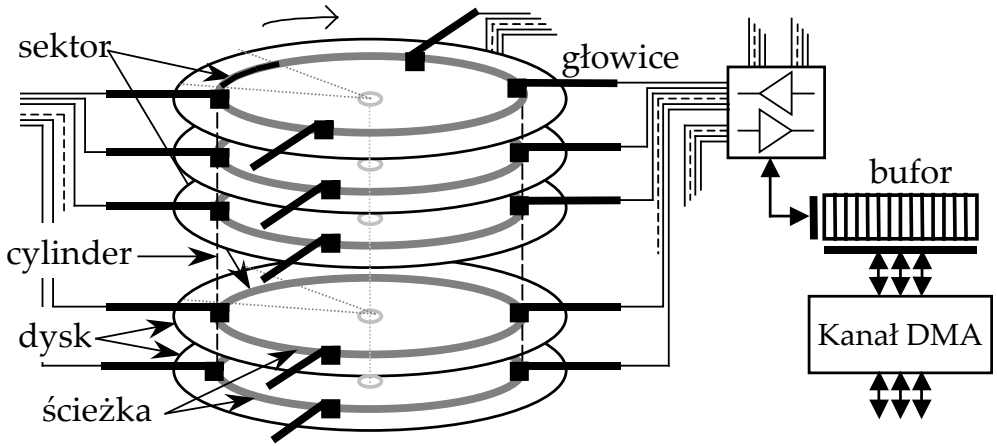
- pamięć optyczna
  polaryzacja odbitej wiązki monochromatycznej

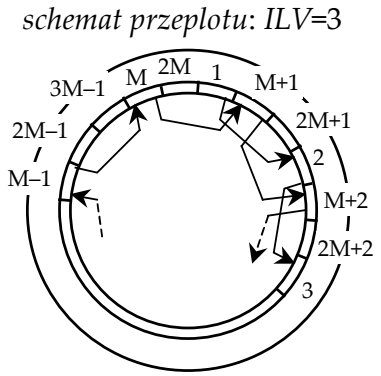
- pamięć magnetooptyczna zapisywalna
  efekt Kerra – zależność polaryzacji światła od kierunku namagnesowania
  zapis: namagnesowanie plamki w temperaturze > T_{Curie}

problemy:

- wiarygodność danych → kody korekcyjne CRC (*cyclic redundancy check*)
- czas dostępu
  przeplot – fragmenty łańcucha w oddalonych sektorach ścieżki
  buforowanie → pamięć o dostępie bezpośrednim (DAM)

Pamięć buforowana o dostępie sekwencyjnym – dysk magnetyczny



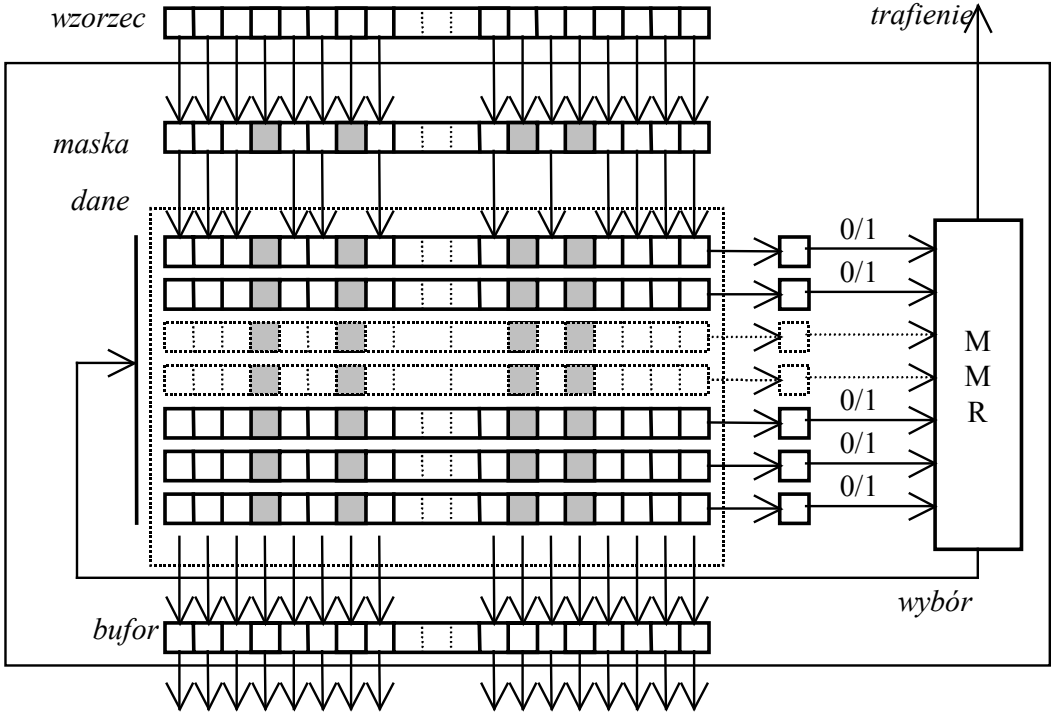


Organizacja pamięci dyskowej

- przyśpieszenie dostępu:
  - → współbieżny dostęp do sektorów na tej samej ścieżce (wiele głowic)
    - → dystrybucja plików pomiędzy dyskami
  - → bufor dysku (*disk cache*) – pojemność = (sektor), ścieżka, dysk, cylinder
  - → przeplot sektorów (*interleave*) – części łańcucha w oddalonych sektorach
    - → liczba sektorów ścieżki = $(ILV+1)M+1$ (ILV – współczynnik przeplotu)

Pamięć skojarzeniowa (*associative memory*)

(…) – asocjacyjna, adresowalna przez zawartość (*content-addressable*, CAM),



Organizacja pamięci skojarzeniowej

Pamięć o dostępie bezpośrednim

- buforowanie nośnika sekwencyjnego
- prognozowanie zapotrzebowania na dane (lokacje)