Pamięć

dr hab. inż. Krzysztof Patan, prof. PWSZ

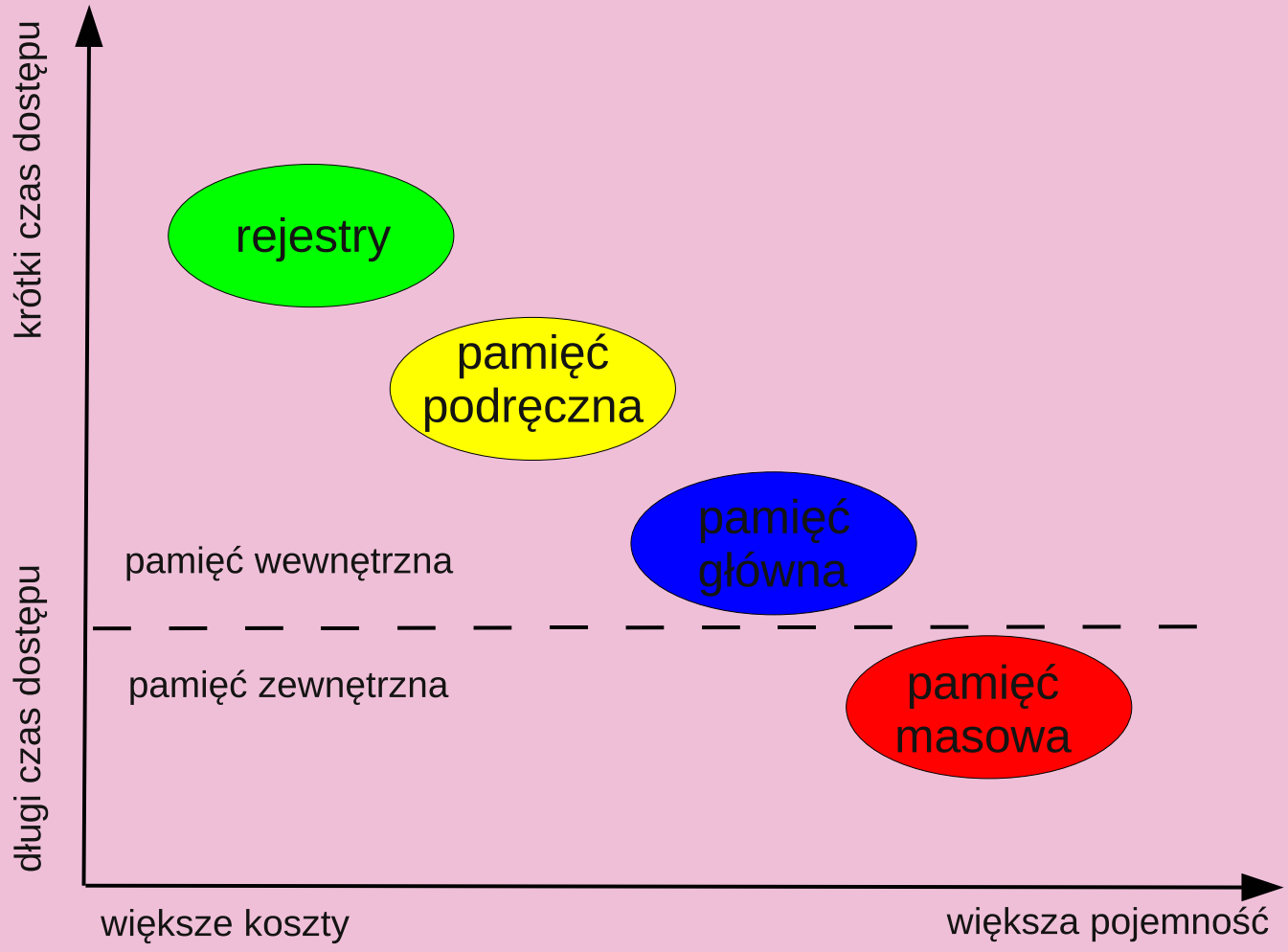
Instytut Politechniczny
Państwowa Wyższa Szkoła Zawodowa w Głogowie
k.patan@issi.uz.zgora.pl

Hierarchia pamięci

- Wiele programów, aby wykonać zadanie, potrzebuje znacznych zasobów pamięci oraz szybkiego dostępu do danych
- Pomysł – użycie ogromnych szybkich pamięci
  - ✗ szybkie pamięci (statyczne) zajmują duże obszary układów VLSI, wymagają dużej mocy – rozwiązanie jest niepraktyczne
  - ✗ problem czasu dostępu do pamięci – im większa pamięć tym czas dostępu wolniejszy
- Rozwiązanie – wykorzystanie zasady lokalności (ang. *principle of locality*)
  1. odniesienie tymczasowe – komórki raz użyte wkrótce będą użyte ponownie
  2. odniesienie przestrzenne – komórki pamięci umieszczone blisko komórki użytej także będą wkrótce użyte
  3. nie trzeba posiadać ogromnych rozmiarów pamięci, aby zapewnić szybki dostęp do danych

- Można zbudować pamięć o bardzo krótkim czasie dostępu na tyle dużą, aby mogła pomieścić blok danych na których program aktualnie pracuje – pierwszy poziom hierarchii pamięci (rejestry procesora)
- Blok danych, który będzie użyty jako następny jest przechowywany w kolejnym poziomie hierarchii – pamięć podręczna
- Kolejne poziomy hierarchii przechowują coraz większe bloki danych – poziom 3 to pamięć główna i poziom 4 to pamięć masowa
- Rozmiar bloku danych rośnie wraz ze schodzeniem w dół hierarchii pamięci
- Czas dostępu się wydłuża wraz ze schodzeniem w dół hierarchii pamięci

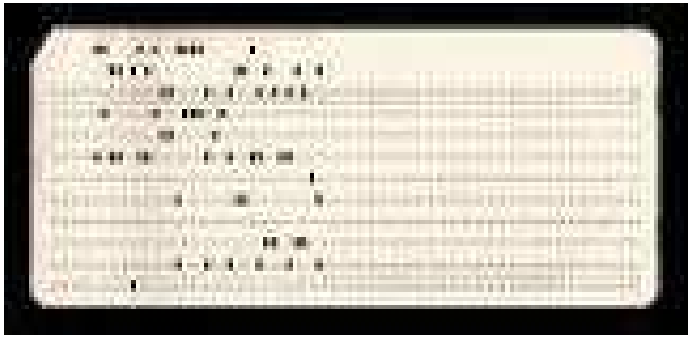
Hierarchia pamięci



Trochę historii

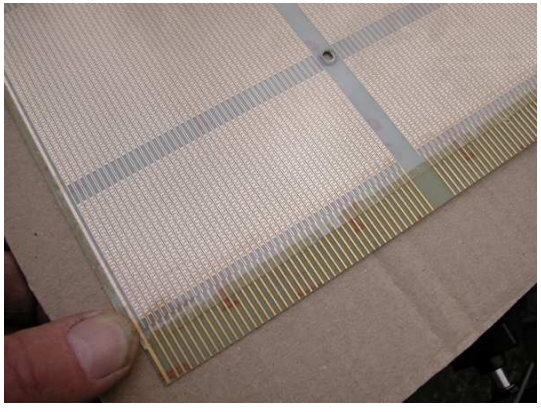
Pamięć tylko do odczytu, ROM (ang. Read Only Memory)

karty dziurkowane – Babagge 1700

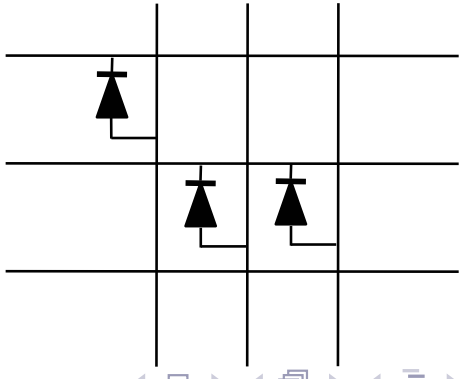


taśma dziurkowana, strumień danych – Harvard MK1

pamięć zrównoważona – IBM

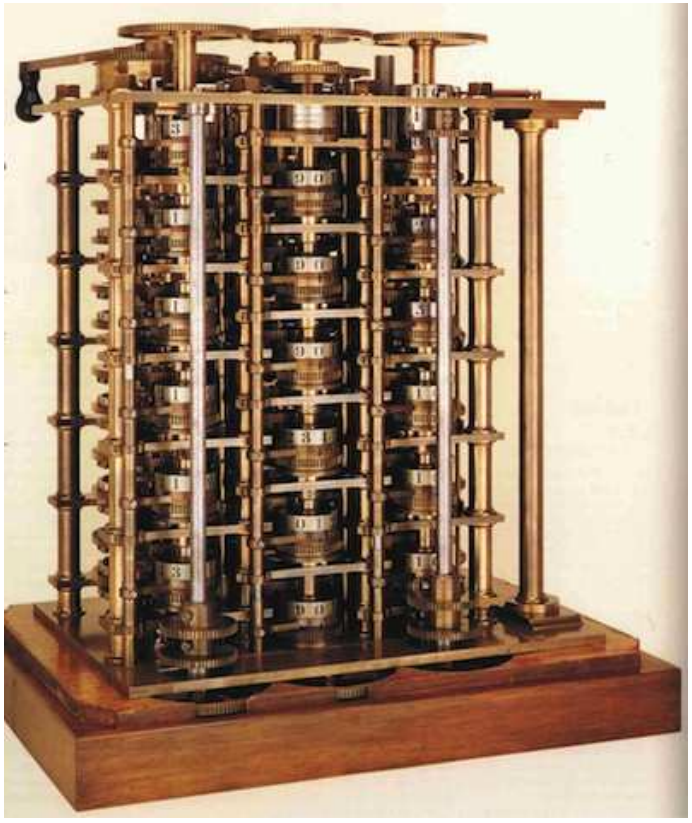


macierz diodowa – EDVAC

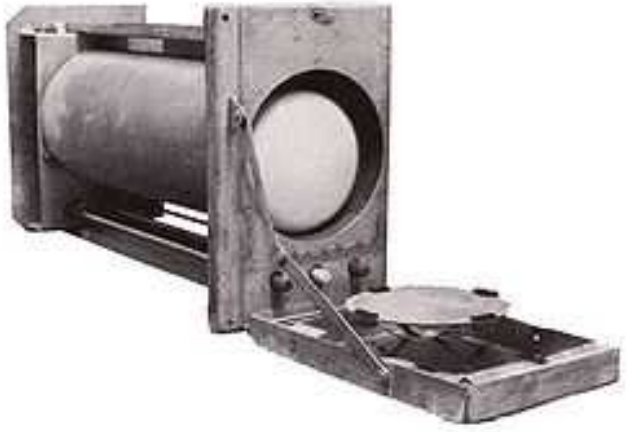


Wczesne rozwiązania pamięci do odczytu i zapisu (ang. Read/Write Memory)

Babagge, 1800, liczby przechowywane za pomocą krążków mechanicznych



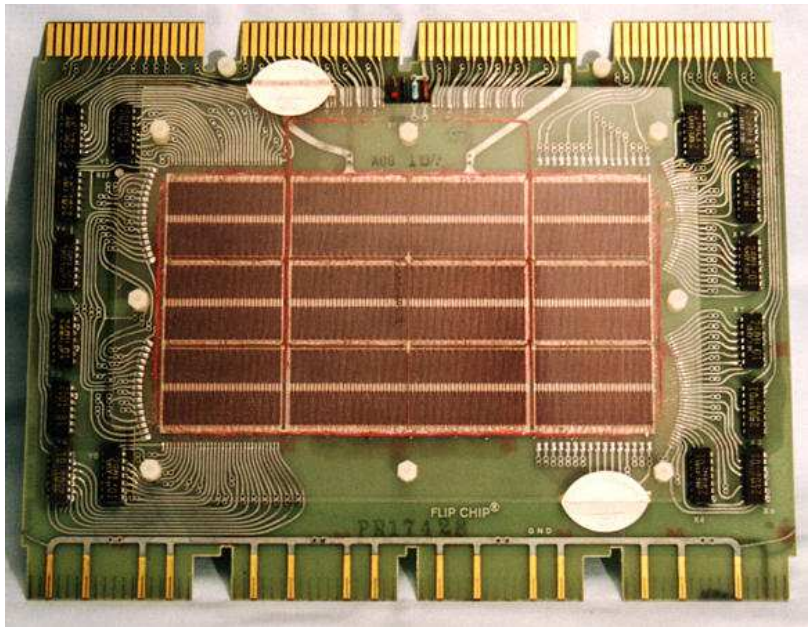
tuba Williamsa, Manchester Mark 1, 1947



linia opóźniająca Mercury, UNIVAC 1, 1951

Pamięć rdzeniowa

- Pierwsza realizacja pamięci dużej skali – Forester, MIT przełom lat 40 i 50 XX wieku
- Bity przechowywane są w postaci polaryzacji namagnesowanych małych rdzeni ferrytowych rozmieszczonych na 2 wymiarowej siatce przewodów
- Zbieżne pulsy na odpowiednich przewodach X i Y powodują zapisanie komórki
- Odporna nieulotna pamięć
- Używana w komputerach pokładowych wahadłowców
- Czas dostępu – $1\mu s$
- Rdzenie były montowane ręcznie
- DEC PDP-8, $4k$ słów $\times$ 12bitów



Pamięci półprzewodnikowe

- Pierwsze pamięci półprzewodnikowe pojawiły się na początku lat 70 XX wieku
  - stworzona przez firmę Intel
  - pamięci były statyczne (Static Random Access Memory, SRAM)
- Pierwszą pamięcią dynamiczną (Dynamic RAM, DRAM) był układ Intel 1103
  - 1kb pamięci na pojedynczym układzie scalonym
  - wykorzystano kondensator do zapamiętania wartości
- Pamięci półprzewodnikowe są ulotne
- Pamięci półprzewodnikowe szybko zastąpiły pamięci rdzeniowe

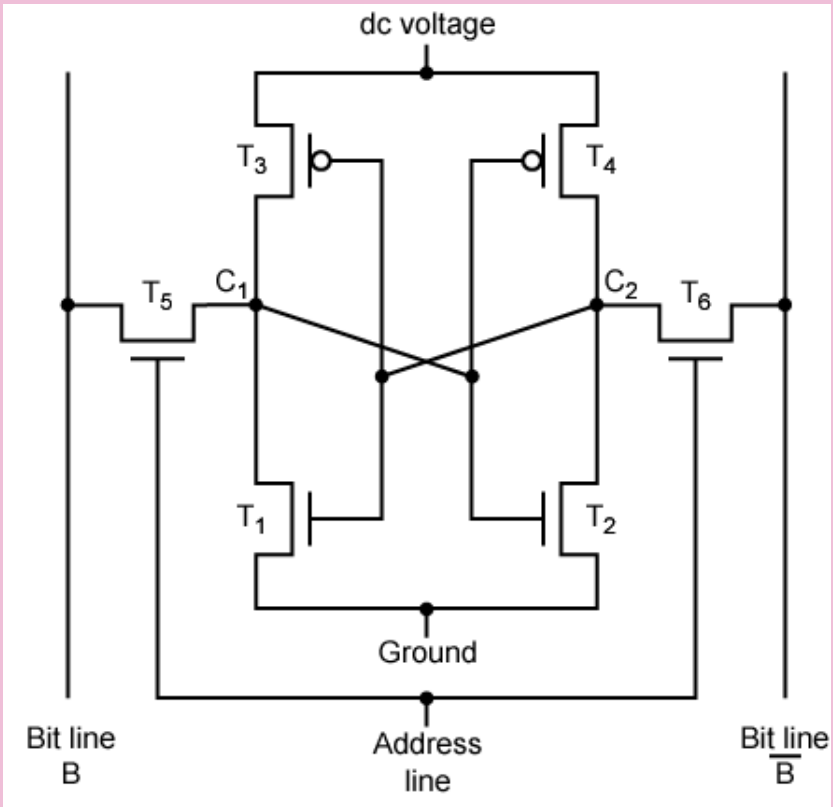
Pamięć ROM

- Pamięć nieulotna
- Typy pamięci ROM:
  1. zapisywane podczas produkcji – bardzo drogie
  2. programowalne (tylko raz)
     - PROM (ang. *Programmable ROM*)
     - wymagany sprzęt specjalistyczny do zaprogramowania
  3. wielokrotnie programowalne
     - EPROM (ang. *Erasable PROM*) – usuwanie danych promieniami UV
     - EEPROM (ang. *Electrically Erasable PROM*) – usuwanie danych elektrycznie, zapis trwa dużo dłużej jak odczyt
     - pamięci typu flash – usuwanie danych elektrycznie

Pamięć SRAM

Struktura komórki pamięci

- Układ tranzystorów unipolarnych tworzy przerzutnik dwustabilny
- Stan 1: C_1 wysokie, C_2 niskie, T_1 i T_4 wyłączone, T_2 i T_3 włączone
- Stan 0: C_2 wysokie, C_1 niskie, T_2 i T_3 wyłączone, T_1 i T_4 włączone
- T_5 i T_6 – tranzystory sterowane linią adresową
- zapis – podaj wartość na B i $\bar{B}$
- odczyt – wartość jest na linii B



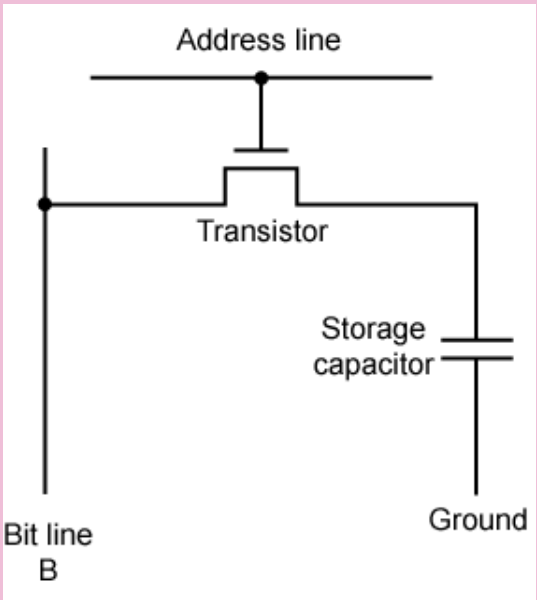
Pamięć SRAM – podsumowanie

- Bity przechowywane na zasadzie przełącznika wł./wył. (1/0)
- Wymagane stałe zasilanie przełącznika – duże zużycie energii
- Nie wymagane odświeżanie stanu – brak kondensatorów
- Skomplikowana konstrukcja, duże rozmiary na jeden bit realizacji
- Drogie rozwiązanie
- Szybki dostęp do komórek (czas dostępu $20 - 200ns$)
- Zastosowania – pamięć podręczna
- Bateryjne podtrzymanie zasilania – pamięci trwałe

Pamięć DRAM

Struktura komórki pamięci

- Kiedy bit jest czytany lub pisany na linię adresową podawany jest sygnał i przez tranzystor zaczyna płynąć prąd
- Operacja zapisu
  - podajemy napięcie na linię bitową – wysokie dla 1 i niskie dla 0
  - następnie podajemy sygnał na linię adresową
  - następuje przepływ prądu, który ładuje kondensator
- Operacja odczytu
  - podajemy sygnał na linię adresową – tranzystor zaczyna pracować
  - kondensator rozładowuje się poprzez linię bitową do czujnika
  - następuje porównanie odczytanej wielkości ze wzorcową, aby określić wartość bitu (0 lub 1)
  - stopień naładowania kondensatora musi zostać odtworzony



Pamięć DRAM – podsumowanie

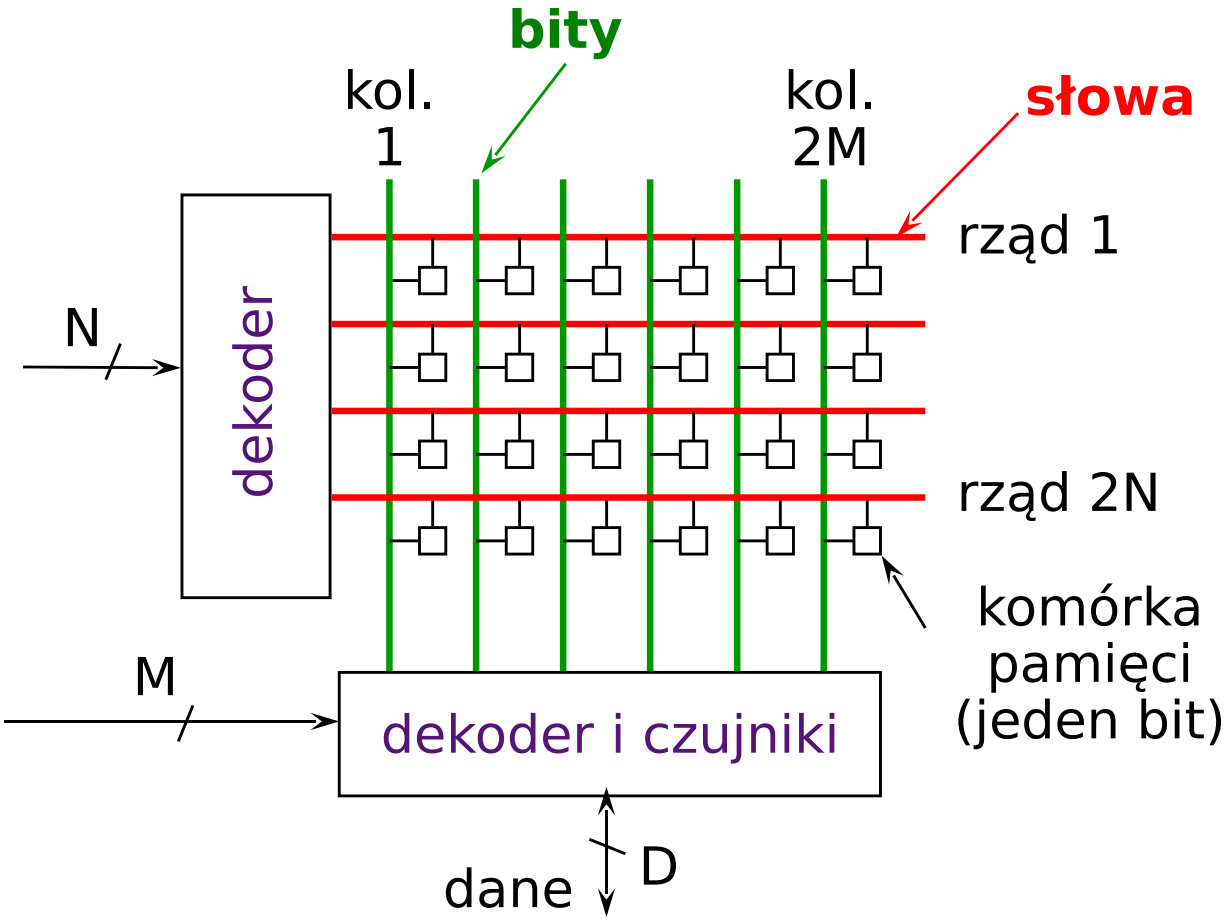
- Odczyt komórki pamięci jest operacją niszczącą powodującą rozładowanie kondensatora
- Wewnętrzny układ sterujący zapewnia regenerację (odtwarzanie stanu)
- Kondensatory komórek nieużywanych przez pewien czas rozładowują się stopniowo poprzez pasożytnicze upoływności układu
- Konieczność okresowego odświeżania zwartości pamięci

Porównanie pamięci SRAM i DRAM

- oba typy to pamięci ulotne
- zasilanie niezbędne do przechowywania danych
- komórka dynamiczna
  - łatwiejsza w budowie, mniejszych rozmiarów
  - tańsza w konstrukcji
  - wymaga odświeżania stanu (rozładowanie kondensatorów)
  - umożliwia budowę pamięci o dużych rozmiarach (RAM)
  - mały pobór energii
- komórka statyczna
  - szybsza w działaniu
  - duże rozmiary
  - droga w wytworzeniu
  - duży pobór energii
  - pamięci podręczne

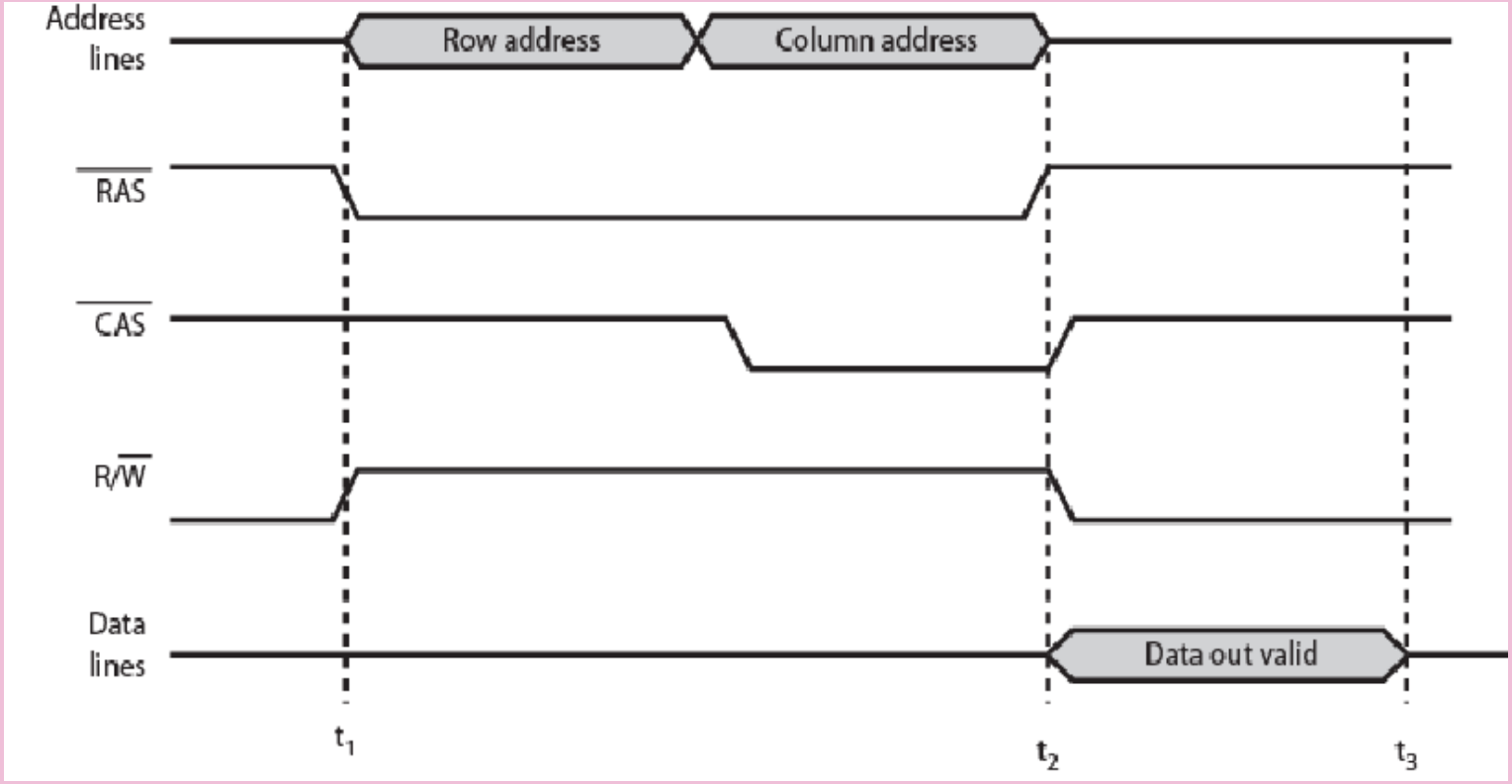
Nowoczesna architektura pamięci DRAM

Wartości bitów przechowywane są w dwuwymiarowej tablicy zrealizowanej w postaci układu scalonego



- Pamięci DRAM pracują w trzech fazach
  - dostęp do rzędu RAS (ang. *Row Access Select*) – dekodowanie adresu rzędu, udostępnienie rzędu, odświeżenie komórek pamięci
  - dostęp do kolumny CAS (ang. *Column Access Select*)– dekodowanie adresu kolumny, czytanie – przesyłanie odpowiednich bitów na wyjścia układu scalonego, zapis – ładowanie do komórek pamięci porządanych wartości
  - ładowanie – ładowanie lini bitowych do znanych wartości, wymagane przed ponownym dostępem do rzędu
  - każdy krok trwa ok. 15-20ns

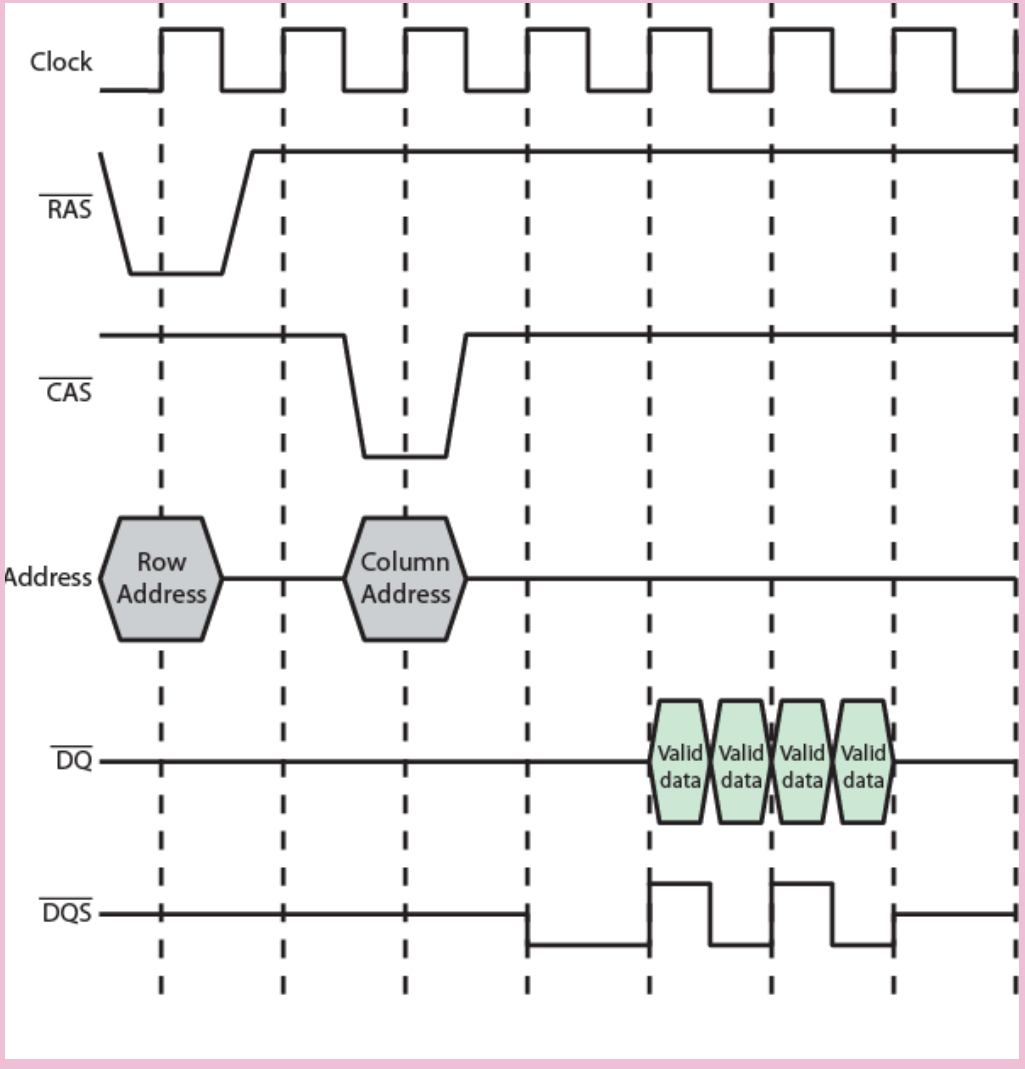
Uproszczony schemat dostępu do danych pamięci DRAM



Pamięci DRAM synchroniczne

- Dostęp jest synchronizowany za pomocą zewnętrznego zegara
- W tradycyjnej pamięci DRAM, po podaniu adresu kiedy dane są wyszukiwane w pamięci, CPU czeka na pobranie tych danych nie wykonując żadnych czynności
- W pamięci SDRAM, dostęp do danych taktowany jest zegarem, CPU wie kiedy dane będą dostępne więc może wykonywać inne zadania
- Efektywniejsza praca systemu komputerowego
- Double Data Rate SDRAM (DDR SDRAM) pobiera dane dwukrotnie w czasie jednego cyklu zegarowego (zbocze narastające i opadające)

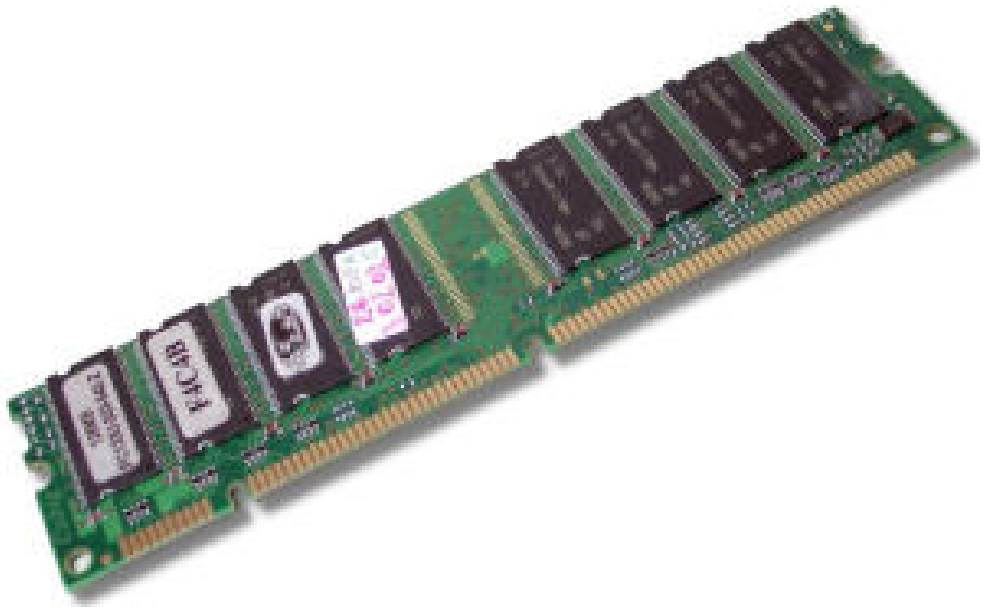
Schemat dostępu do danych pamięci DDR SDRAM



Podział pamięci RAM

SDR SDRAM

- Single Data Rate Synchronous Dynamic Random Access Memory
- Taktowana częstotliwościami 66, 100 i 133 MHz
- Produkowane układy 32, 64, 128, 256 i 512 MB
- Produkcja została zaprzestana z powodu pojawienia się pamięci DDR – szybszych i wydajniejszych



DDR SDRAM

- Double Data Rate Synchronous Dynamic RAM
- Produkcję rozpoczęto w 1999 roku.
- Dane przesyłane są w czasie trwania zarówno rosnącego jak i opadającego zbocza zegara – dwa razy większa przepustowość
- Układy zasilane są napięciem 2,5 V, a nie 3,3 V – znaczące ograniczenie poboru mocy
- Oznaczenia układów: PC-x (PC-y)
  gdzie x – częstotliwość pracy, y –przepustowość

- PC-200 (PC-1600): 64 bity $\cdot$ 2 $\cdot$ 100 MHz = 1600 MB/s
- Produkowane typy: PC-200, PC-266, PC-333, PC-400

DDR2 SDRAM

- Double Data Rate 2 Synchronous Dynamic RAM
- Stosowana jest wyższa częstotliwość taktowania: 533, 667, 800, 1066 MHz
- Niższy pobór prądu w stosunku do DDR SDRAM
- Temperatura pracy do 70^oC
- Moduły pamięci DDR2 nie są kompatybilne z modułami DDR
- Napięcie zasilania 1,8V
- Identyczny sposób oznaczania jak przy DDR
- Produkowane typy: PC2-3200, PC2-4200, PC2-5200, PC2-6400, PC2-8000

DDR3 SDRAM

- Double Data Rate 3 Synchronous Dynamic RAM
- Wykonana w technologii 90 nm – zastosowanie niższego napięcia – 1,5 V
- Zmniejszony pobór mocy o około 40% w stosunku do pamięci DDR2
- Większa przepustowość w porównaniu do DDR2 i DDR
- Pamięci DDR3 nie są kompatybilne wstecz, tzn. nie będą współpracowały z chipsetami obsługującymi DDR i DDR2
- Identyczny sposób oznaczania jak przy DDR
- Moduły DDR3: PC3-6400, PC3-8500, PC3-10600, PC3-12700

Architektura Dual Channel

- Technologia stosowana w płytach głównych do wydajniejszej obsługi pamięci DDR SDRAM
- Podwojenie przepustowości przesyłu danych przez magistralę łączącą pamięć RAM z mostkiem północnym (ang. northbridge), pełniącego rolę kontrolera pamięci
- Wykorzystuje dwa 64-bitowe kanały, co razem daje kanał szerokości 128 bitów dla przesyłu danych pomiędzy pamięcią RAM, a procesorem
- Wymaga umieszczania kości pamięci parami w skorelowanych ze sobą gniazdach oznaczonych odpowiednimi kolorami Dual Channel

Pamięć podręczna

- Poprawia komunikację pomiędzy pamięcią operacyjną, a procesorem
- Pamięć o krótkim czasie dostępu, niewielkich rozmiarów
- Wykorzystuje zasadę lokalności
  1. lokalność tymczasowa – pamiętanie zawartości ostatnio używanych obszarów
  2. lokalność przestrzenna – pobranie bloków danych umieszczonych w pobliżu ostatnio używanych obszarów

Umiejscowienie pamięci podręcznej

Zasada działania pamięci podręcznej – operacja czytania

- Pobranie adresu od procesora
- Przeszukanie pamięci podręcznej w celu znalezienia pasujących wpisów
- Przypadek 1 – znalezienie wpisu (dopasowanie)
  - pobranie danych z pamięci podręcznej i przekazanie do CPU
- Przypadek 2 – brak wpisu (chybienie)
  - przeczytanie danych z pamięci RAM
  - przekazanie danych do CPU
  - aktualizacja zawartości pamięci podręcznej

Współpraca pamięci podręcznej z pamięcią operacyjną

- Algorytmy wymiany zawartości pamięci podręcznej
  - najdawniej używany LRU (ang. *Least Recently Used LRU*)
  - kolejka FIFO (ang. *First-In-First-Out*)
- Spójność pamięci podręcznej
  - zapis przez (ang. *write throught*)
  - zapis z opóźnieniem (ang. *write back*)
- Średni czas dostępu do pamięci:

$$t_{av} = ht_c + (1 - h)(t_m + t_c)$$

gdzie h – współczynnik trafień, t_c – czas dostępu do pamięci podręcznej, t_m – czas dostępu do pamięci operacyjnej