



# Architektura komputerów

sem. zimowy 2024/2025

cz. 5

Tomasz Dziubich

# Tematyka wykładu

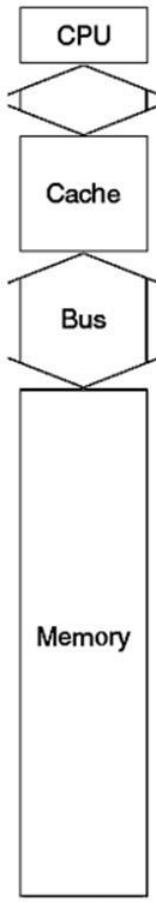
- Podstawy
  - Zasady adresacji sygnałowej pamięci
  - Budowa komórki pamięci
  - Parametry układów pamięci
  - Odświeżanie pamięci
- Klasyfikacja układów pamięci
  - Hierarchiczność pamięci
  - Zasada lokalności
- Rodzaje pamięci
  - Pamięć RAM
  - Pamięć podręczna - cache
  - Pamięć ROM



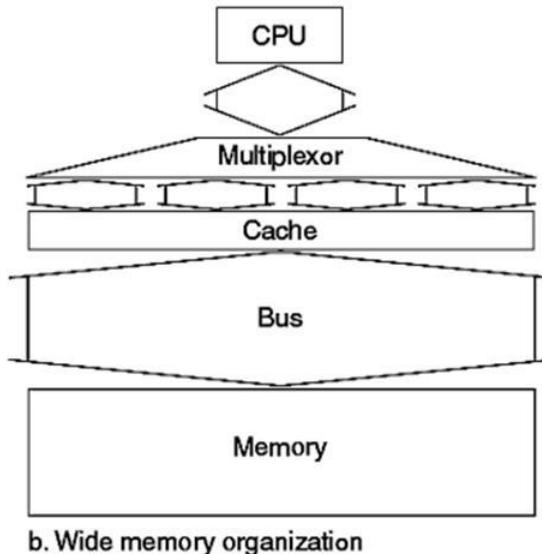
# Cel stosowania pamięci operacyjnej

- Procesor dysponuje mocą przetwarzania programów, ale nie ma zdolności ich zapamiętywania
- Zapamiętanie programów i danych jest zadaniem dla pamięci
- Czas dostępu do pamięci może być stały (o tzw. dostępie swobodnym) lub zmienny
- Pamięć operacyjna o dostępie swobodnym może być trwała lub nietrwała
- Pamięci operacyjna to zazwyczaj pamięć RAM (*random access memory*) lub ROM (*read only memory*)

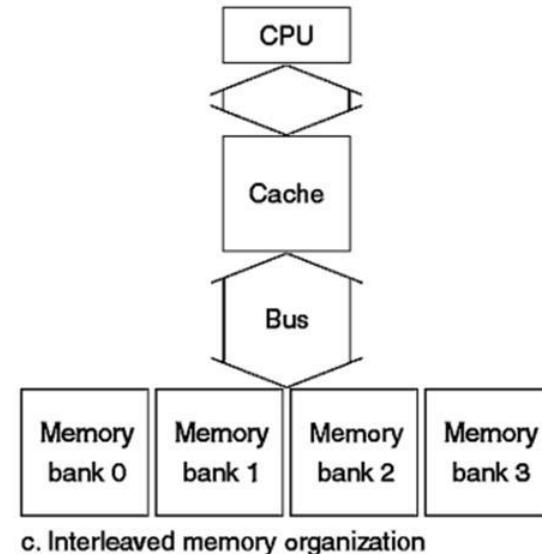
# Sposoby komunikacji procesor – pamięć



a. One-word-wide  
memory organization

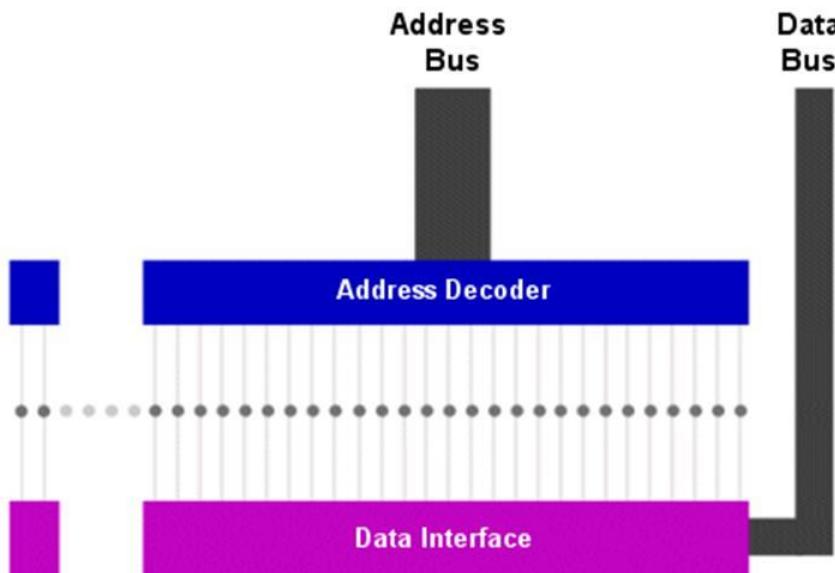


b. Wide memory organization

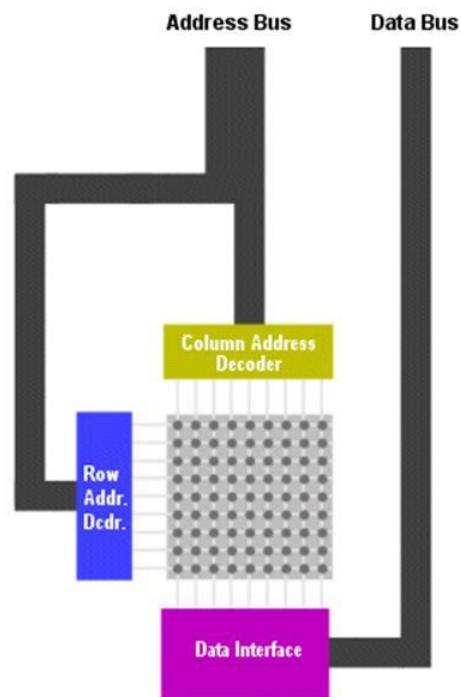


c. Interleaved memory organization

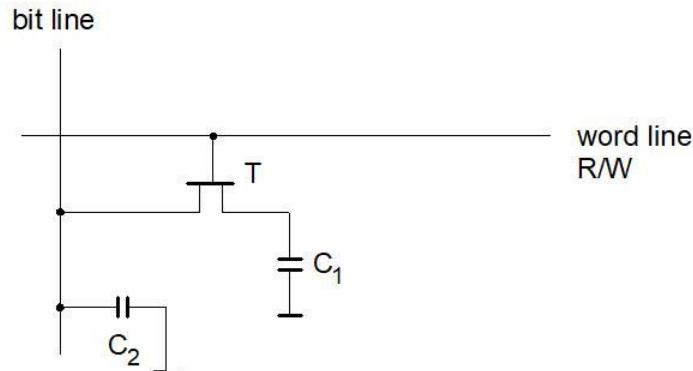
## Schemat uproszczony



## Schemat fizyczny



## Pamięć dynamiczna DRAM



- Tranzystor i kondensator

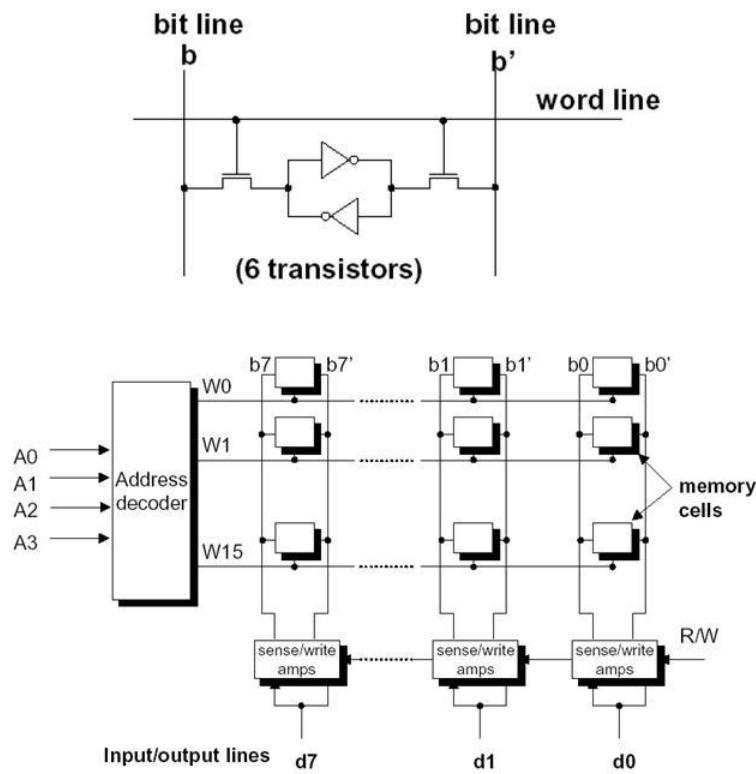
### Zalety

- Niski pobór mocy
- Duży stopień scalenia
- Niski koszt wykonania

### Wady

- Czas dostępu ~60 ns
- konieczność odświeżania zawartości
- dłuższy cykl odczytu
- Niska odporność na zakłócenia

## Pamięć statyczna SRAM



- Przerzutnik dwustanowy

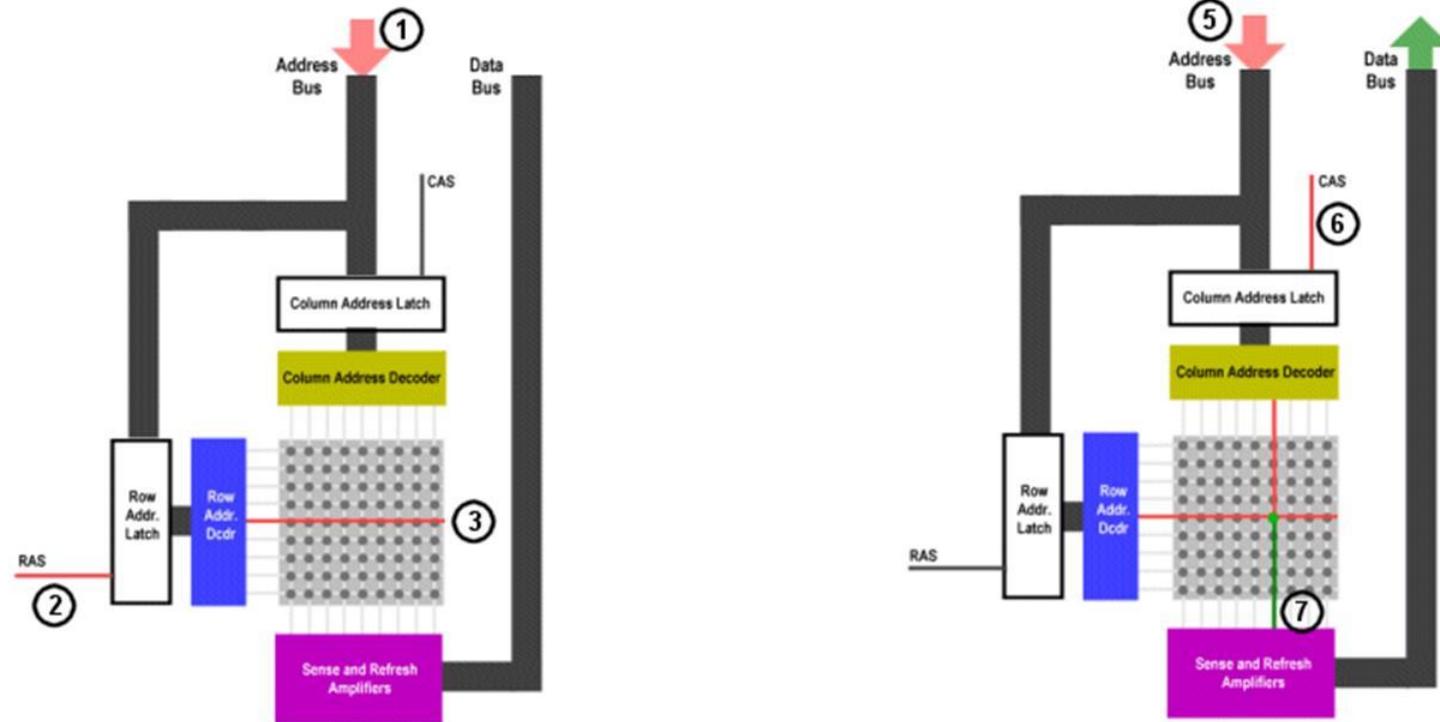
## Zalety

- Brak potrzeby odświeżania zawartości
- Krótki czas dostępu ~10 ns (5-10 cykli)
- Krótszy cykl odczytu
- Wysoka odporność na zakłócenia

## Wady

- Duży pobór mocy
- Mały stopień scalenia

# Operacja odczytu danych

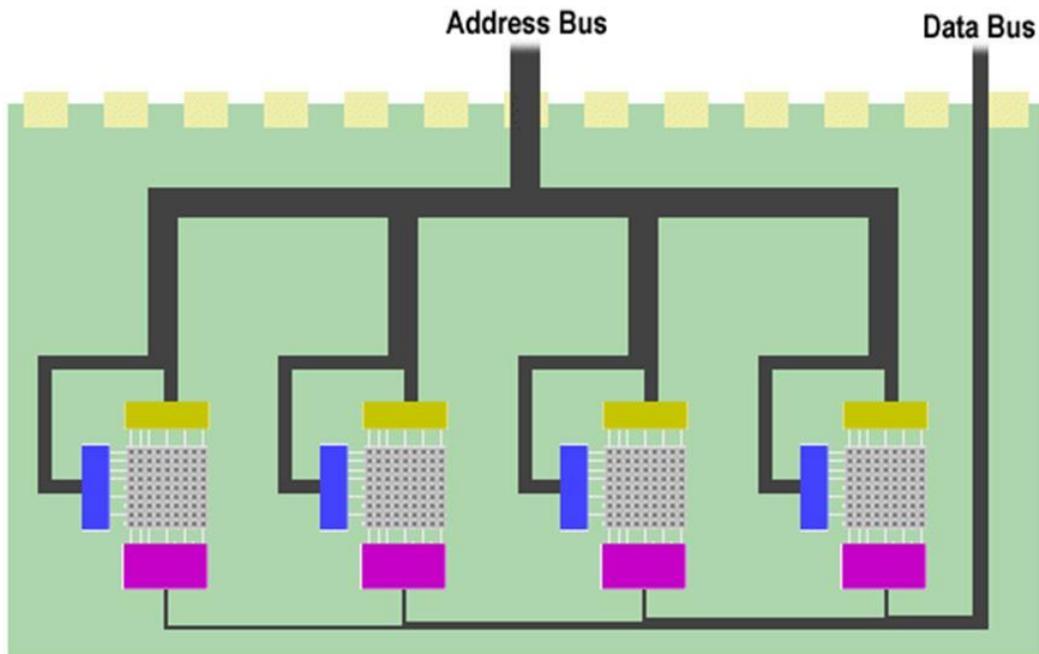


- Szerokość AB (Address Bus) = 22 bity ( $2 * 11$ )
- Szerokość DB (Data Bus) = 8 bitów
- Każdy chip zapamiętuje  $2^{22} = 4194304$  bitów (4 Mb)
- 8 chipów połączonych razem (każdy produkuje 1 bit wyjściowy) daje 4 MB
- Pamięć taką oznacza się jako 4M x 8
- Szerokość AB w Pentium IV (Address Bus) = 36 bitów ( $2 * 18$ ), w i7 Core = 64 bity

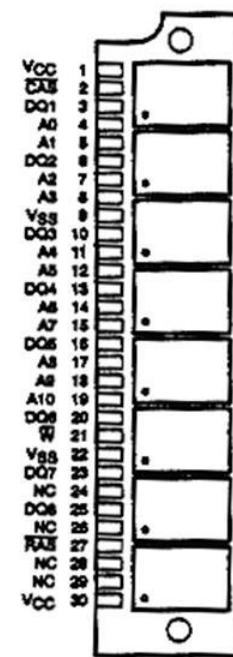
SINGLE IN-LINE MODULE  
(TOP VIEW)

PIN NOMENCLATURE	
A0-A10	Address Input
CAS	Column-Address Strobe
DQ1-DQ8	Data In/Data Out
NC	No Internal Connection
RAS	Row-Address Strobe
Vcc	5-V Supply
Vss	Ground
W	Write Enable

# Fizyczna budowa układu



SINGLE IN-LINE MODULE  
(TOP VIEW)



PIN NOMENCLATURE

A0-A10	Address Inputs
CAS	Column-Address Strobe
DQ1-DQ8	Data In/Data Out
NC	No Internal Connection
RAS	Row-Address Strobe
V <sub>CC</sub>	3-V Supply
V <sub>SS</sub>	Ground
W	Write Enable



- Kondensator – będący nośnikiem informacji – ulega rozładowaniu (samoistnemu) – stąd konieczność odświeżania jego zawartości
- Proces odświeżania występuje co 2 - 4 ms, polega na odczytaniu zawartości komórki i powtórnym jej zapisaniu
- Podczas operacji odczytu następuje przepływ ładunku z kondensatora C1 do linii bitu, co pociąga za sobą wymazanie informacji
- Konieczne jest więc ładowanie pojemności (odświeżenie zawartości) także bezpośrednio po odczycie (pogorszenie parametrów)
- Dwie metody odświeżania:
  - asynchronicznie — łatwe w realizacji, mało wydajne (procesor jest blokowany na kilkadziesiąt mikrosekund)
  - synchronicznie — trudniejsze i kosztowniejsze w realizacji (odświeżanie występuje w czasie wyznaczonych odcinków zegara co kilkanaście mikrosekund)



- FPM DRAM (*Fast Page Mode DRAM*) — w tego typu pamięciach przy dostępie do zawartości wiersza stosuje się sekwencję sygnałów strobujących [RAS, CAS, CAS, CAS, CAS] zamiast [(RAS, CAS), (RAS, CAS), (RAS, CAS), (RAS, CAS)]
- EDO DRAM (*Extended Data Out DRAM*) — ulepszony FPM DRAM z bardziej precyzyjnie generowanymi sygnałami CAS
- SDRAM (*Synchronous DRAM*) — sterowanie synchroniczne impulsami zegarowymi
- DDR SDRAM (*Double Data-Rate Synchronous DRAM*) — ulepszona wersja SDRAM, w której używane oba zbocza impulsu zegarowego jako sygnału sterującego



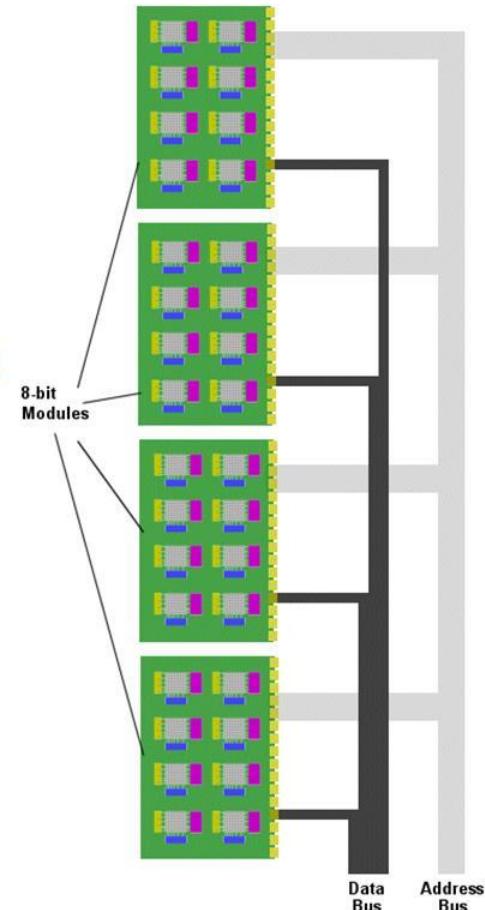
## Czasy dostępu i ceny pamięci w różnych technologiach

Memory technology	Typical access time	Price per GiB in 2012
SRAM semiconductor memory	0.5–2.5 ns	\$500–\$1000
DRAM semiconductor memory	50–70 ns	\$10–\$20
Flash semiconductor memory	5,000–50,000 ns	\$0.75–\$1.00
Magnetic disk	5,000,000–20,000,000 ns	\$0.05–\$0.10

D.A. Patterson and J.L. Hennessy. *Computer Organization and Design: The Hardware/Software Interface*. The Morgan Kaufmann Series in Computer Architecture and Design. Elsevier Science, 2013. ISBN 9780124078864.

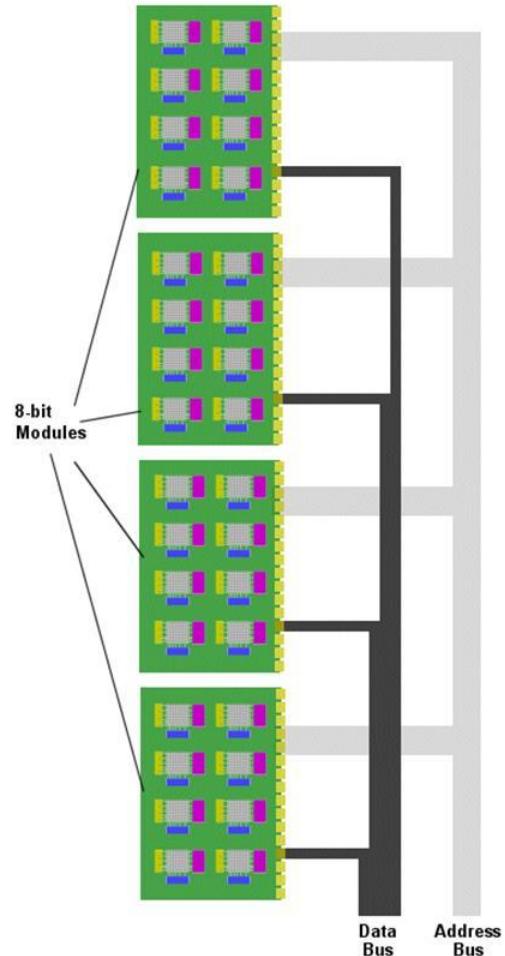
# Sposoby zwiększania wydajności pamięci DRAM

- Zwiększenie szerokości odczytywanych danych (problem wyrównywania)
- Manipulacja sygnałami sterującymi
- Wprowadzenie synchronicznego odczytu
- Odczyt na opadającym i narastającym zboczu zegara
- Bankowanie układów pamięci

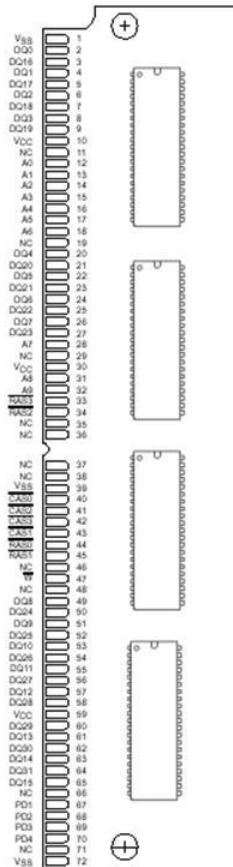


- Zwiększenie pamięci w systemie odby sposoby:

Wydłużenie długości słowa - szersza magistra  
lub wydłużenie czasu odczytu - połączenie kostek par



- SIMM
- 72 SIMM (32 bity danych)
- 168 DIMM (64 bity danych)

BJ SINGLE IN-LINE MEMORY MODULE  
(TOP VIEW)

(SIDE VIEW)

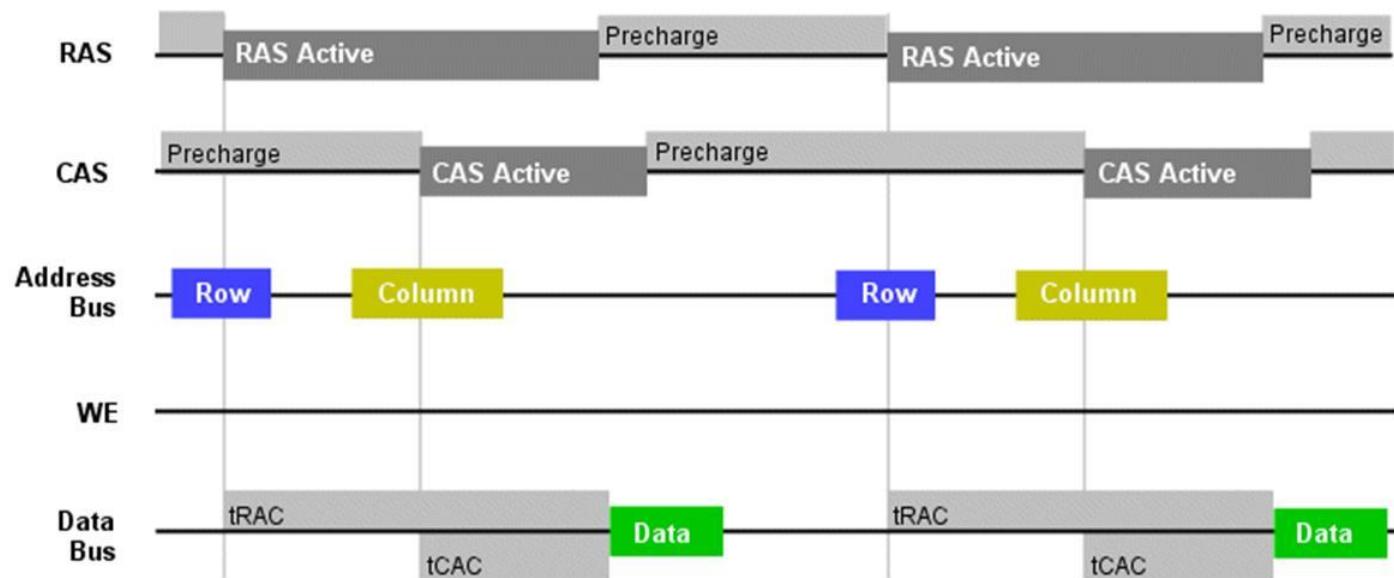


PIN NOMENCLATURE	
A0–A9	Address Inputs
CAS0–CAS3	Column-Address Strobe
DQ0–DQ31	Data In/Data Out
NC	No Connection
PD1–PD4	Presence Detects
RAS0–RAS3	Row-Address Strobe
VCC	5-V Supply
Vss	Ground
W	Write Enable

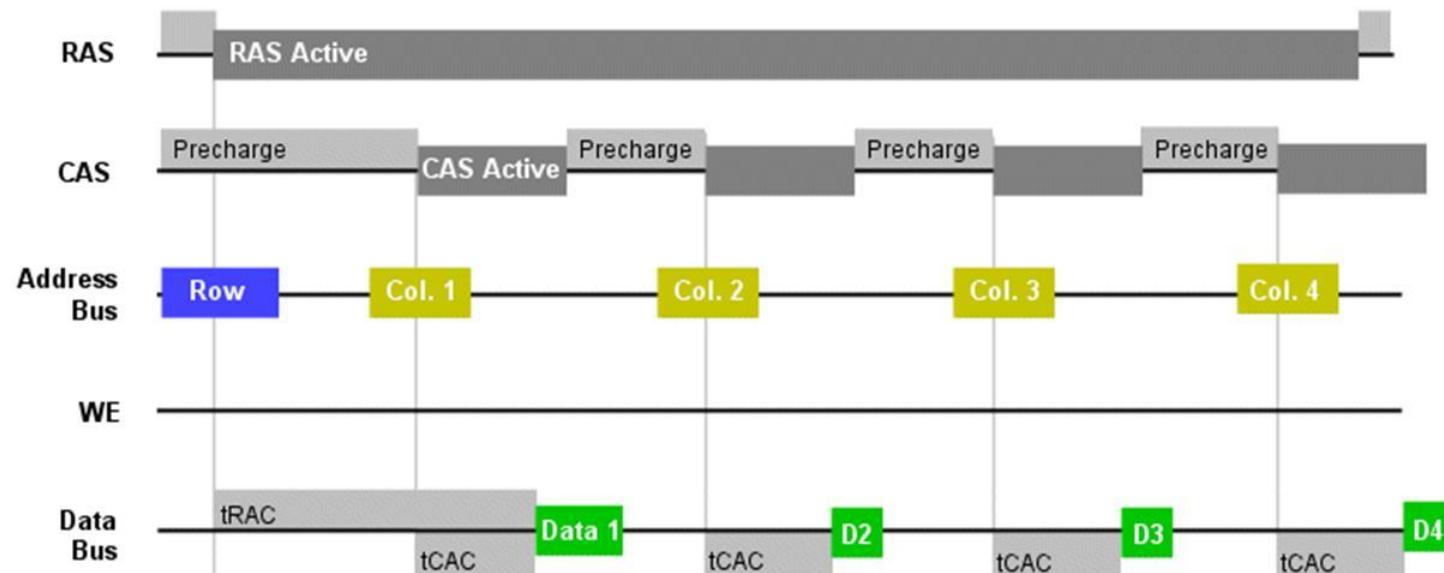


- czas dostępu (*access time*)
- czas cyklu (*cycle time*)
- opóźnienie CL, CAS Latency
- x-y-z (dla pamięci asynchronicznych)
- u-x-y-z (dla pamięci synchronicznych)

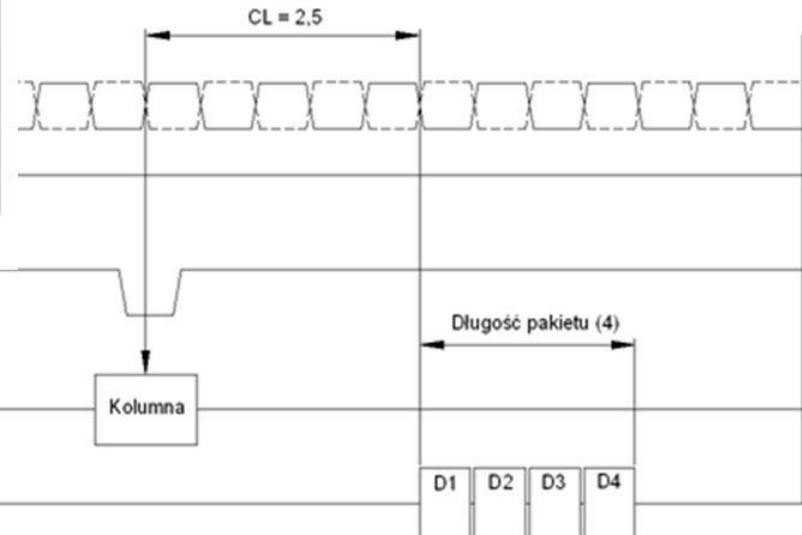
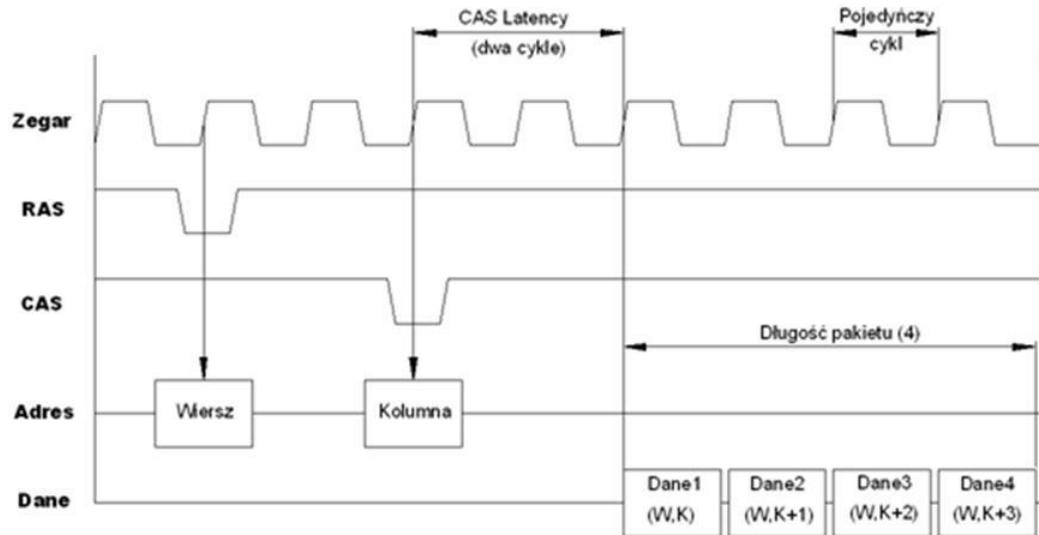
## DRAM Read



## Fast Page Mode Read



# Parametry układów pamięci



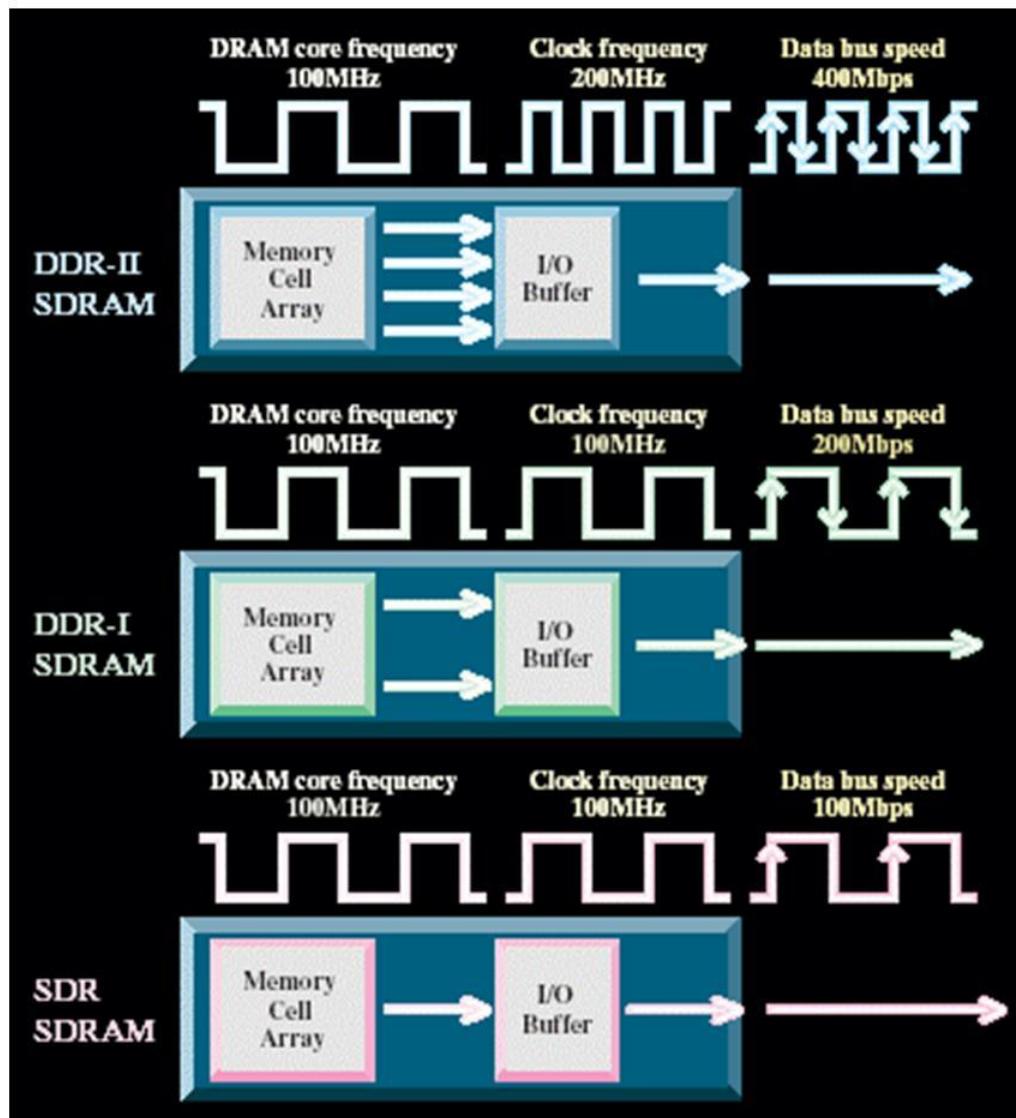
TRUTH TABLE 1 – Commands and DQM Operation

(Notes: 1)

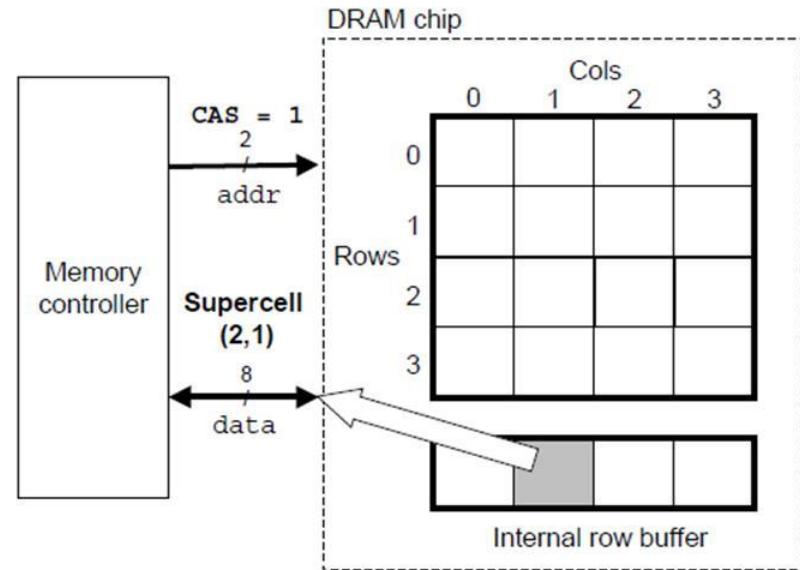
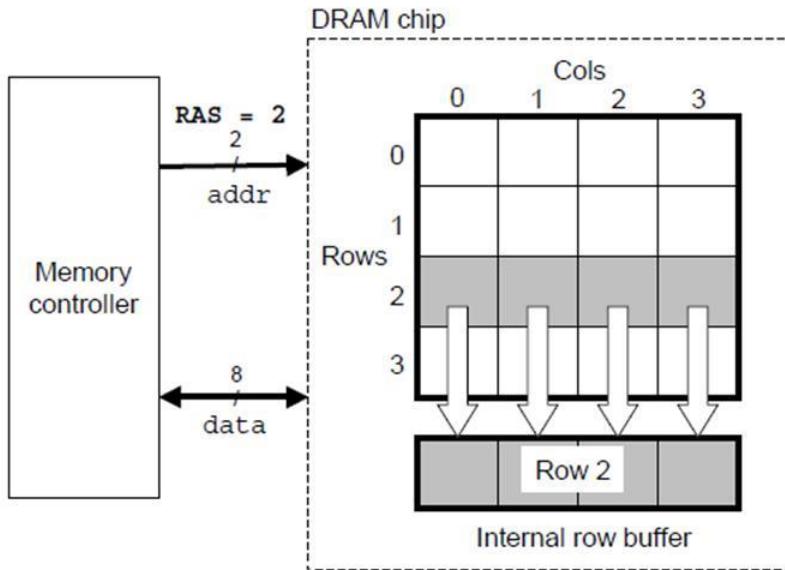
NAME (FUNCTION)	CS#	RAS#	CAS#	WE#	DOM	ADDR	DOs	NOTES
COMMAND INHIBIT (NOP)	H	X	X	X	X	X	X	
NO OPERATION (NOP)	L	H	H	H	X	X	X	
ACTIVE (Select bank and activate row)	L	L	H	H	X	Bank/Row	X	3
READ (Select bank and column and start READ burst)	L	H	L	H	X	Bank/Col	X	4
WRITE (Select bank and column and start WRITE burst)	L	H	L	L	X	Bank/Col	Valid	4
BURST TERMINATE	L	H	H	L	X	X	Active	
PRECHARGE (Deactivate row in bank or banks)	L	L	H	L	X	Code	X	5
AUTO REFRESH or SELF REFRESH (Enter self refresh mode)	L	L	L	H	X	X	X	6, 7
LOAD MODE REGISTER	L	L	L	L	X	Op-code	X	2
Write Enable/Output Enable	-	-	-	-	L	-	Active	8
Write Inhibit/Output High-Z	-	-	-	-	H	-	High-Z	8

- NOTE: 1. CKE is HIGH for all commands shown except SELF REFRESH.  
 2. A0-A10 and BA define the op-code written to the Mode Register.  
 3. A0-A10 provide row address, and BA determines which bank is made active (BA LOW = Bank 0; BA HIGH = Bank 1).  
 4. A0-A9 (A9 is a "Don't Care" for x8) provide column address; A10 HIGH enables the auto precharge feature (nonpersistent), while A10 LOW disables the auto precharge feature; BA determines which bank is being read from or written to (BA LOW = Bank 0; BA HIGH = Bank 1).  
 5. For A10 LOW, BA determines which bank is being precharged (BA LOW = Bank 0; BA HIGH = Bank 1); for A10 HIGH, both banks are precharged and BA is a "Don't Care".  
 6. This command is AUTO REFRESH if CKE is HIGH, SELF REFRESH if CKE is LOW.  
 7. Internal refresh counter controls row addressing; all inputs and I/Os are "Don't Care" except for CKE.  
 8. Activates or deactivates the DOs during WRITES (zero-clock delay) and READS (two-clock delay).

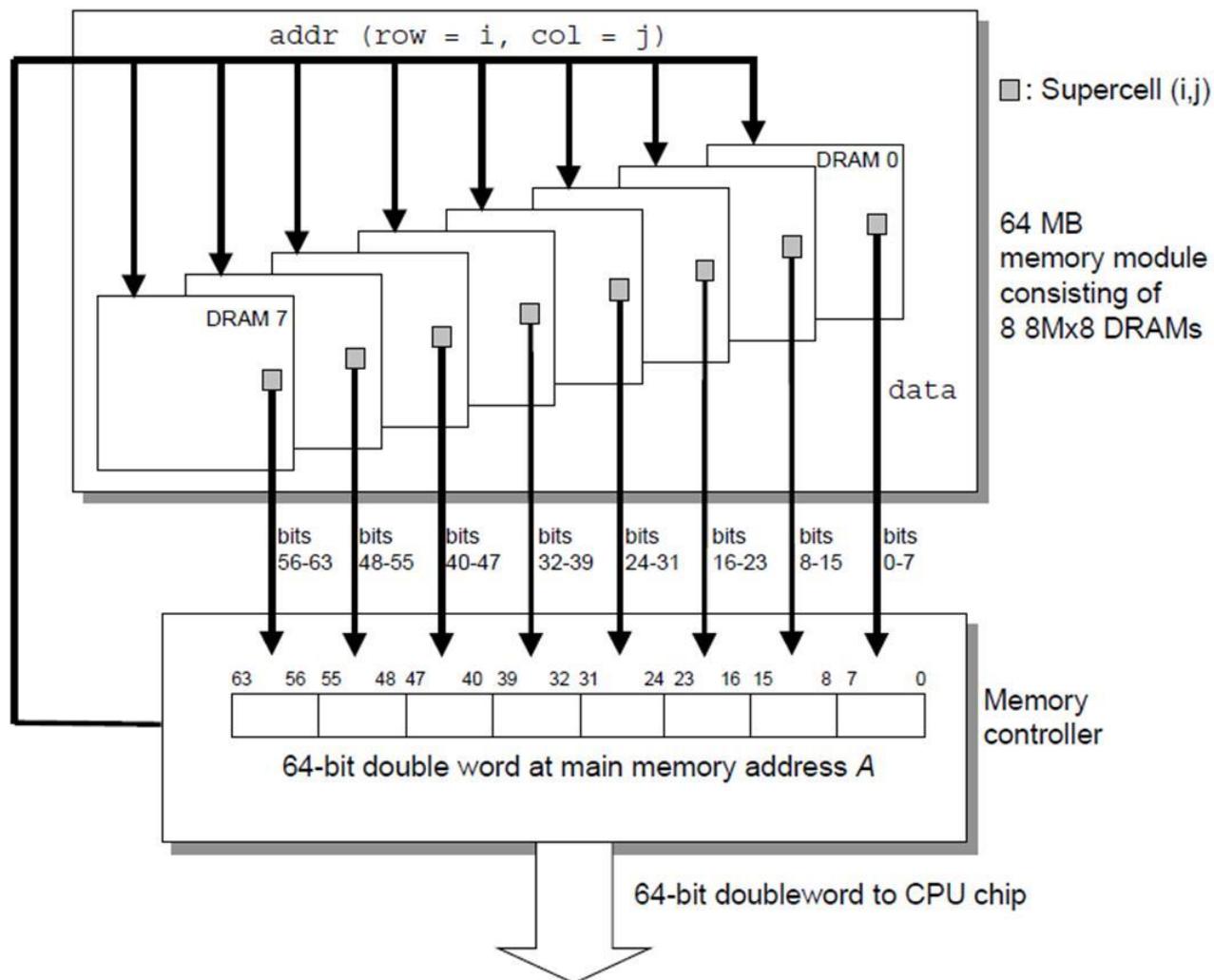
# SDR – DDR – DDR2



# Odczyt zawartości z „superkomórki” DRAM



# Odczytanie zawartości modułu pamięci





Year of introduction	Chip size	Row access strobe (RAS)		Column access strobe (CAS)/data transfer time (ns)	Cycle time (ns)
		Slowest DRAM (ns)	Fastest DRAM (ns)		
1980	64K bit	180	150	75	250
1983	256K bit	150	120	50	220
1986	1M bit	120	100	25	190
1989	4M bit	100	80	20	165
1992	16M bit	80	60	15	120
1996	64M bit	70	50	12	110
1998	128M bit	70	50	10	100
2000	256M bit	65	45	7	90
2002	512M bit	60	40	5	80
2004	1G bit	55	35	5	70
2006	2G bit	50	30	2.5	60

Cycle time is the sum of RAS and CAS times.



- w pamięciach stosuje się niekiedy dodatkowy bit (*parity bit*)
- w komputerach pełniących bardziej odpowiedzialne funkcje stosuje się moduły pamięci wyposażone w kilka bitów parzystości (detekcja i korekcja)- układy i pamięć ECC (*error checking and correction*)
- Pamięci ECC zalecane są w systemach wyposażonych w więcej niż 1 GB RAM
- W rozwiązańach klastrowych stosuje się rozwiązanie Chipkill



## Inne zagadnienia związane z DRAM

- DDR 3, 4 i 5
- Pamięci GDDR (większa szerokość, sposób podłączenia)
- Redukcja zużytej energii – wprowadzenie trybu uśpienia (*power down mode*)

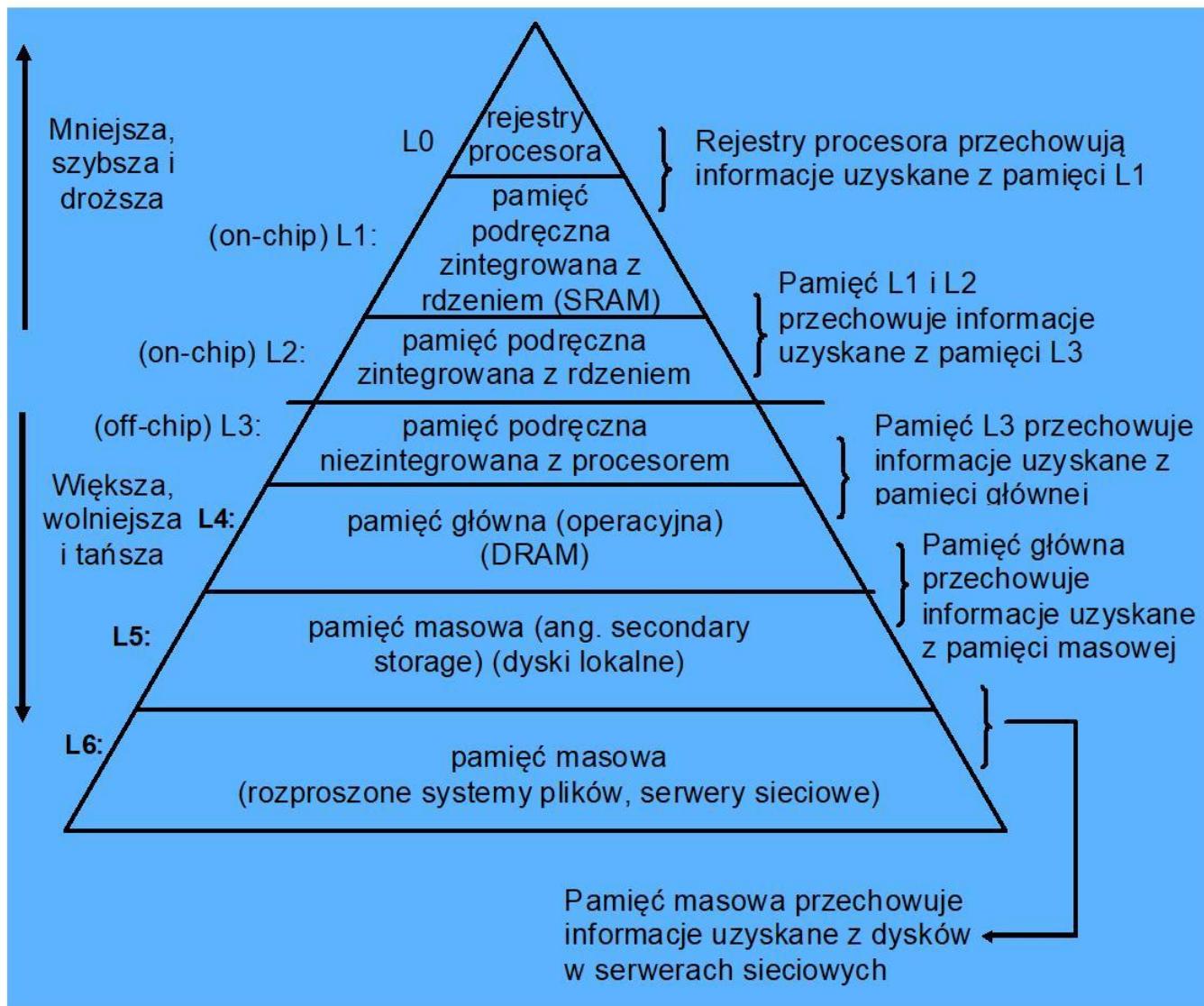


W ciągu krótkiego odcinka czasu procesor odwołuje się do wąskiego zakresu komórek pamięci z danymi bądź rozkazami

```
sum = 0;  
for (i = 0; i < n; i++) sum += a[i];  
return sum;
```

- lokalność czasowa (*temporary locality*) - wielokrotne odwoływanie do tych samych komórek pamięci, nie koniecznie o zbliżonych adresach (np. wykonywanie pętli złożonej z kilku rozproszonych fragmentów kodu)
- lokalność przestrzenna (*spatial locality*) - odwoływanie do komórek pamięci znajdujących się w sąsiedztwie

# Hierarchia układów pamięci



## Serwery:

Reg (1000 bajtów/300 ps)

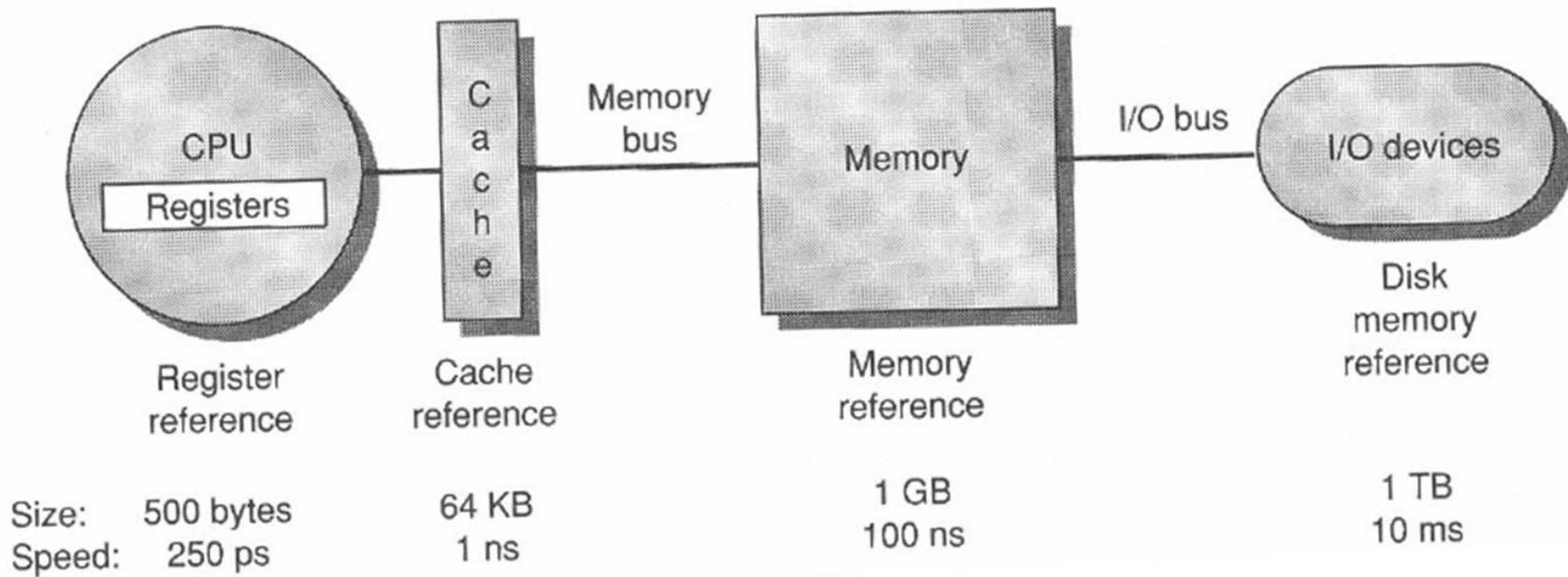
L1 (64 KB/1 ns)      L2 (256 KB/3-10 ns)

L3 (2-4 MB/10-20 ns)

Mem (4-16 GB/50 – 100 ns)

Flash (4-8 GB 25-50 us)

Disk (4-16 TB/ 5-10 ms)





- pamięci dynamiczne są zbyt wolne dla współczesnych procesorów, a przy dostępie występują stany oczekiwania procesora
- pamięć podręczna zawiera pewną liczbę obszarów (tzw. wierszami lub liniami), które służą do przechowywania bloków kopiowanych z pamięci głównej
- typowy blok zawiera 4 - 64 bajtów
- pamięć podręczna może być używana do przechowywania rozkazów i danych;



- Gdzie umieścić blok w pamięci podręcznej?
- Jak odnaleźć blok jeśli jest w pamięci?
- Który blok wymienić?
- Co zrobić przy zapisie w pamięci podręcznej?



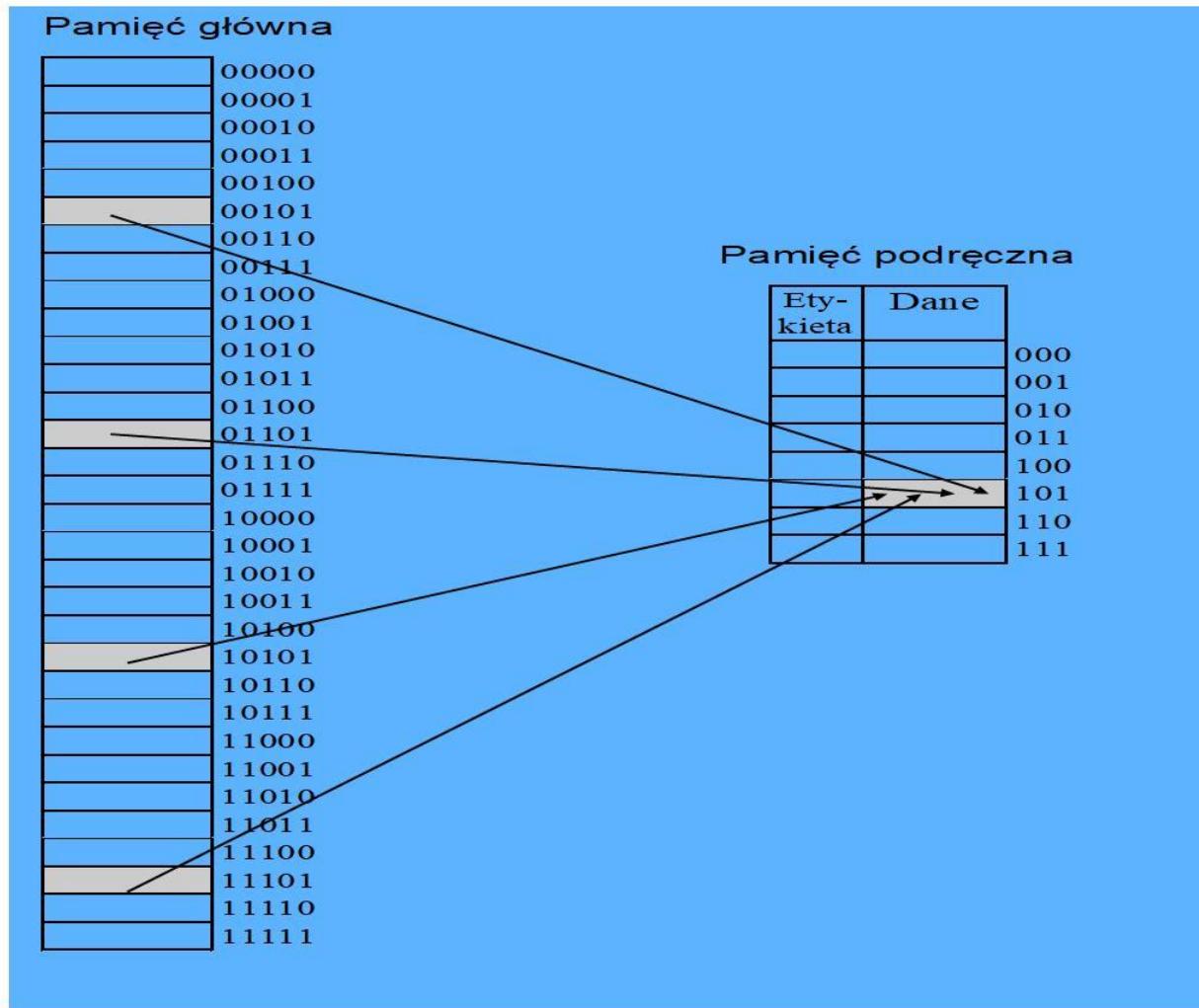
- w trakcie wykonywania instrukcji (rozkazów) procesor szuka najpierw rozkazów i danych w pamięci podręcznej:
  - trafienie (*cache hit*)
  - chybienie (*cache miss*), rzędu do 20%
- inspiracją użycia jest zasada lokalności odwołań
- w trakcie wykonywania operacji zapisu procesor najpierw sprawdza czy kopia potrzebnej lokacji znajduje się w pamięci podręcznej — jeśli tak (*write hit*), to procesor zapisuje wartość do pamięci podręcznej; jeśli potrzebnej lokacji nie ma w pamięci podręcznej, to wynik zapisywany jest tylko do pamięci głównej
- w nowszych typach procesorów, kopiowany jest wymagany blok pamięci głównej do pamięci podręcznej (*cache line fill*), po czym następuje zapis do pamięci podręcznej



- zapewnienie spójności zawartości pamięci operacyjnej i pamięci podręcznej:
  - zapis przez (*write-through*) wykonuje zapis do pamięci głównej po każdej operacji zapisu w pamięci podręcznej
  - zapis z opóźnieniem (*write-back*) - zamiast natychmiastowego zapisu bloku do pamięci głównej, zmienia się tylko bit stanu

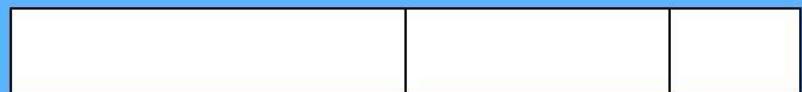
- każdy blok (4 - 16 bajtów) pamięci podręcznej zawiera pole etykiety, nazywanej też numerem bloku; (np. adresy 32-bitowe i bloki 16-bajtowe - pole etykiety - 28 bitów);





# Odwzorowanie bezpośrednie

Adres 32-bitowy generowany przez procesor



etykieta (16 bitów)

nr linii  
(12 bitów)

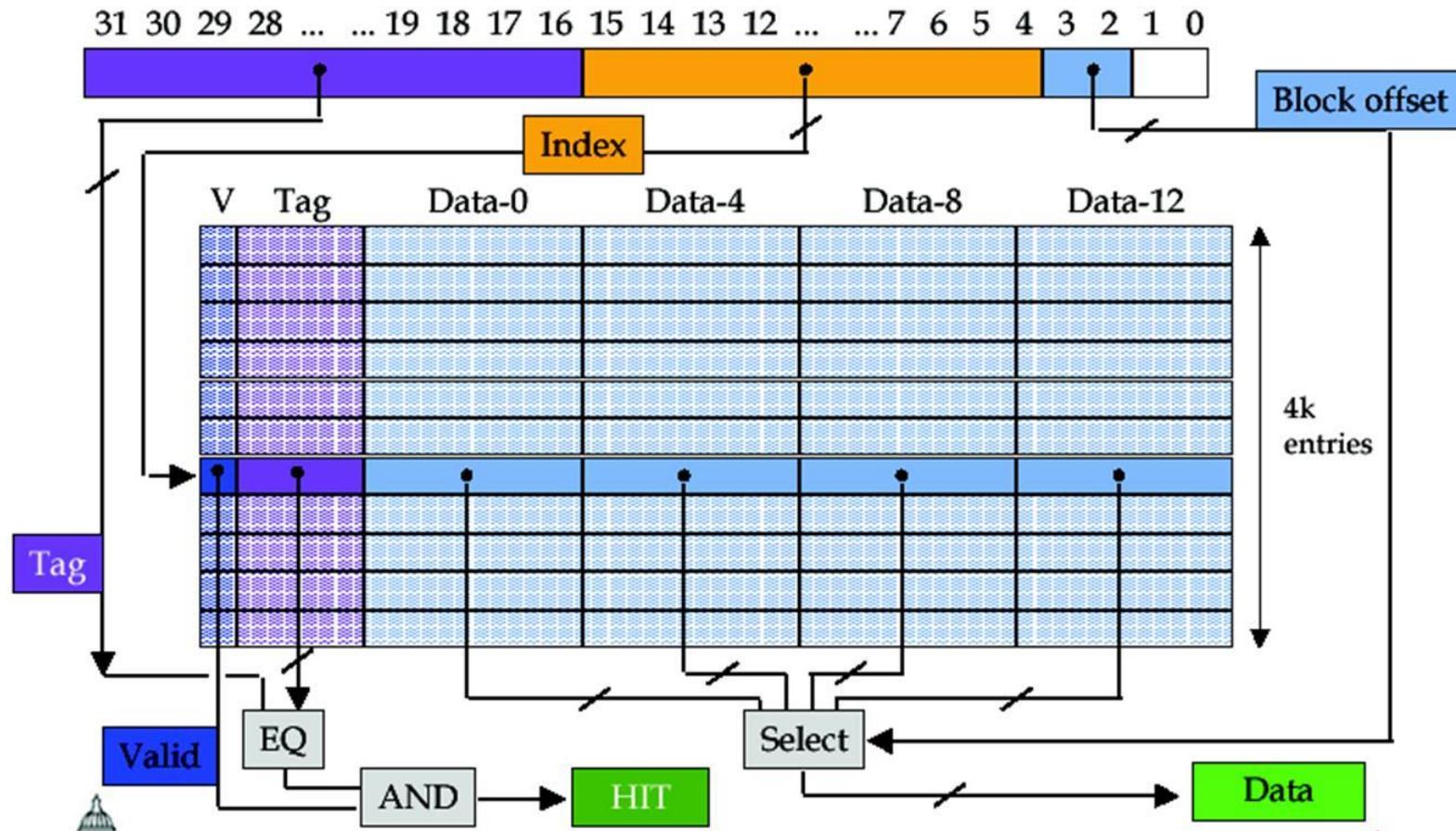
adres  
wewn.  
bloku



	etykieta	blok
000		
001		
002		
FFF		

Pamięć podręczna

# Odwzorowanie bezpośrednie





Index	V	Tag	Data
000	N		
001	N		
010	N		
011	N		
100	N		
101	N		
110	N		
111	N		

a. The initial state of the cache after power-on

Index	V	Tag	Data
000	N		
001	N		
010	Y	11 <sub>two</sub>	Memory (11010 <sub>two</sub> )
011	N		
100	N		
101	N		
110	Y	10 <sub>two</sub>	Memory (10110 <sub>two</sub> )
111	N		

c. After handling a miss of address (11010<sub>two</sub>)

Index	V	Tag	Data
000	Y	10 <sub>two</sub>	Memory (10000 <sub>two</sub> )
001	N		
010	Y	11 <sub>two</sub>	Memory (11010 <sub>two</sub> )
011	Y	00 <sub>two</sub>	Memory (00011 <sub>two</sub> )
100	N		
101	N		
110	Y	10 <sub>two</sub>	Memory (10110 <sub>two</sub> )
111	N		

e. After handling a miss of address (00011<sub>two</sub>)

Index	V	Tag	Data
000	N		
001	N		
010	N		
011	N		
100	N		
101	N		
110	Y	10 <sub>two</sub>	Memory(10110 <sub>two</sub> )
111	N		

b. After handling a miss of address (10110<sub>two</sub>)

Index	V	Tag	Data
000	Y	10 <sub>two</sub>	Memory (10000 <sub>two</sub> )
001	N		
010	Y	11 <sub>two</sub>	Memory (11010 <sub>two</sub> )
011	N		
100	N		
101	N		
110	Y	10 <sub>two</sub>	Memory (10110 <sub>two</sub> )
111	N		

d. After handling a miss of address (10000<sub>two</sub>)

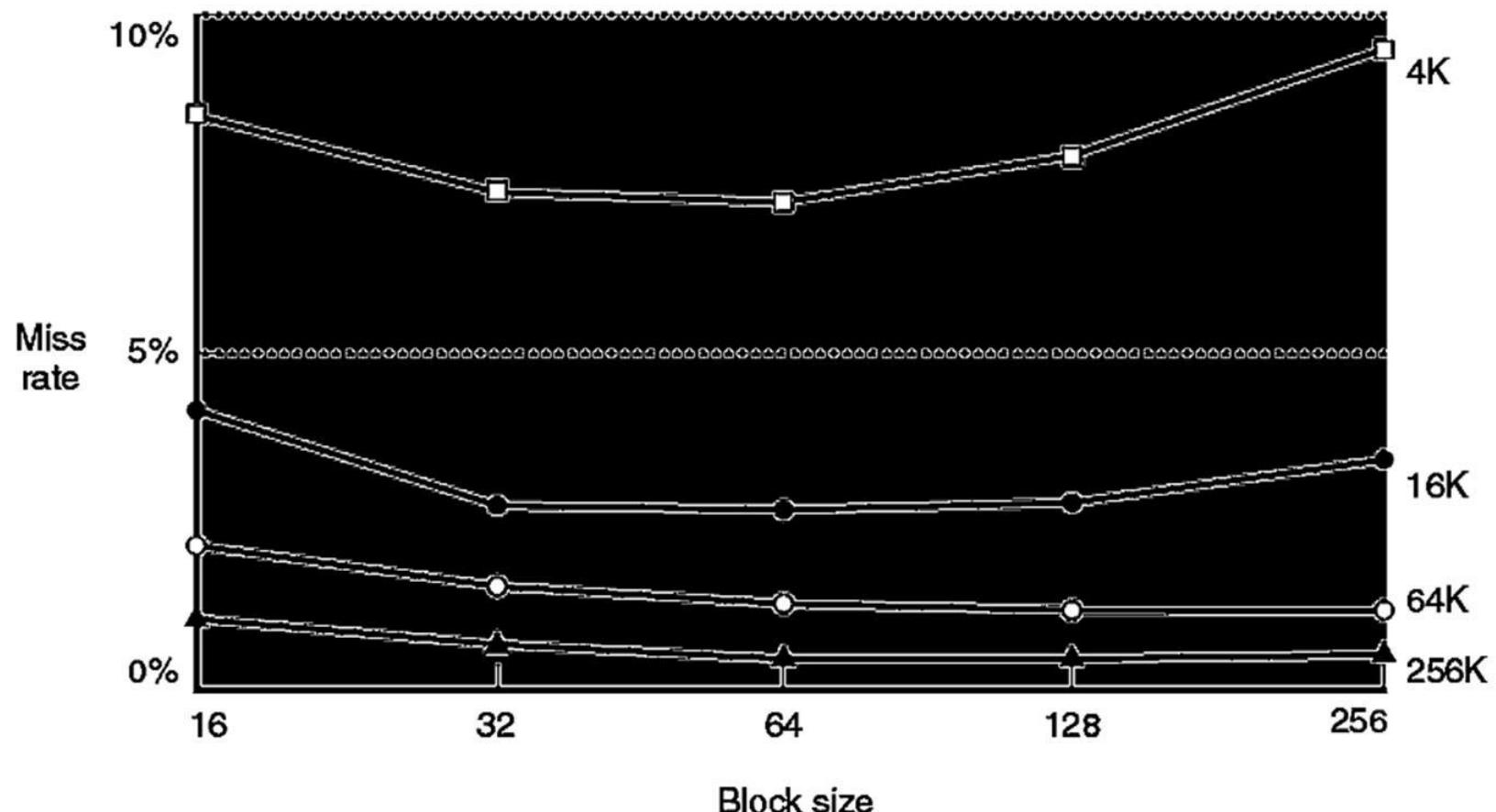
Index	V	Tag	Data
000	Y	10 <sub>two</sub>	Memory (10000 <sub>two</sub> )
001	N		
010	Y	10 <sub>two</sub>	Memory (10010 <sub>two</sub> )
011	Y	00 <sub>two</sub>	Memory (00011 <sub>two</sub> )
100	N		
101	N		
110	Y	10 <sub>two</sub>	Memory (10110 <sub>two</sub> )
111	N		

f. After handling a miss of address (10010<sub>two</sub>)



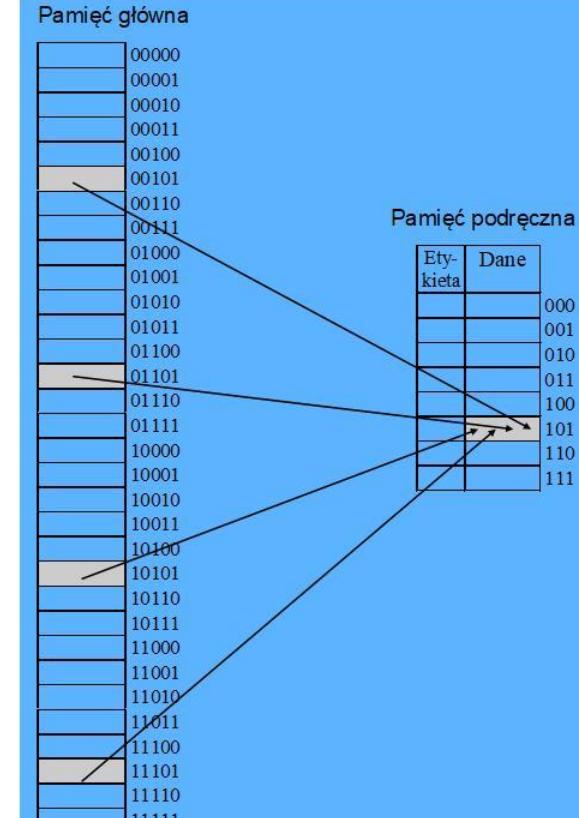
- Jaki jest całkowity rozmiar pamięci podręcznej, która pomieści 8 KB danych w liniach (blokach) o wielkości 8 bajtów i zakładając 32 bitowy adres?

# Współczynnik chybienia



# Ulepszenia pamięci cache

- w przypadku stosowania pamięci podręcznych z odwzorowaniem bezpośrednim, jeśli w programie używanych jest kilka lokacji pamięci o identycznych numerach linii (12 środkowych bitów), to tylko jedna z tych lokacji pamięci może być skopiowana do pamięci podręcznej
- tego rodzaju pamięć podręczna nazywana jest *dwukanałową* (*two way set associative*)



Etykieta	Dane	Etykieta	Dane
			00
			01
			10
			11



## Algorytmy wyboru bloku do wymiany

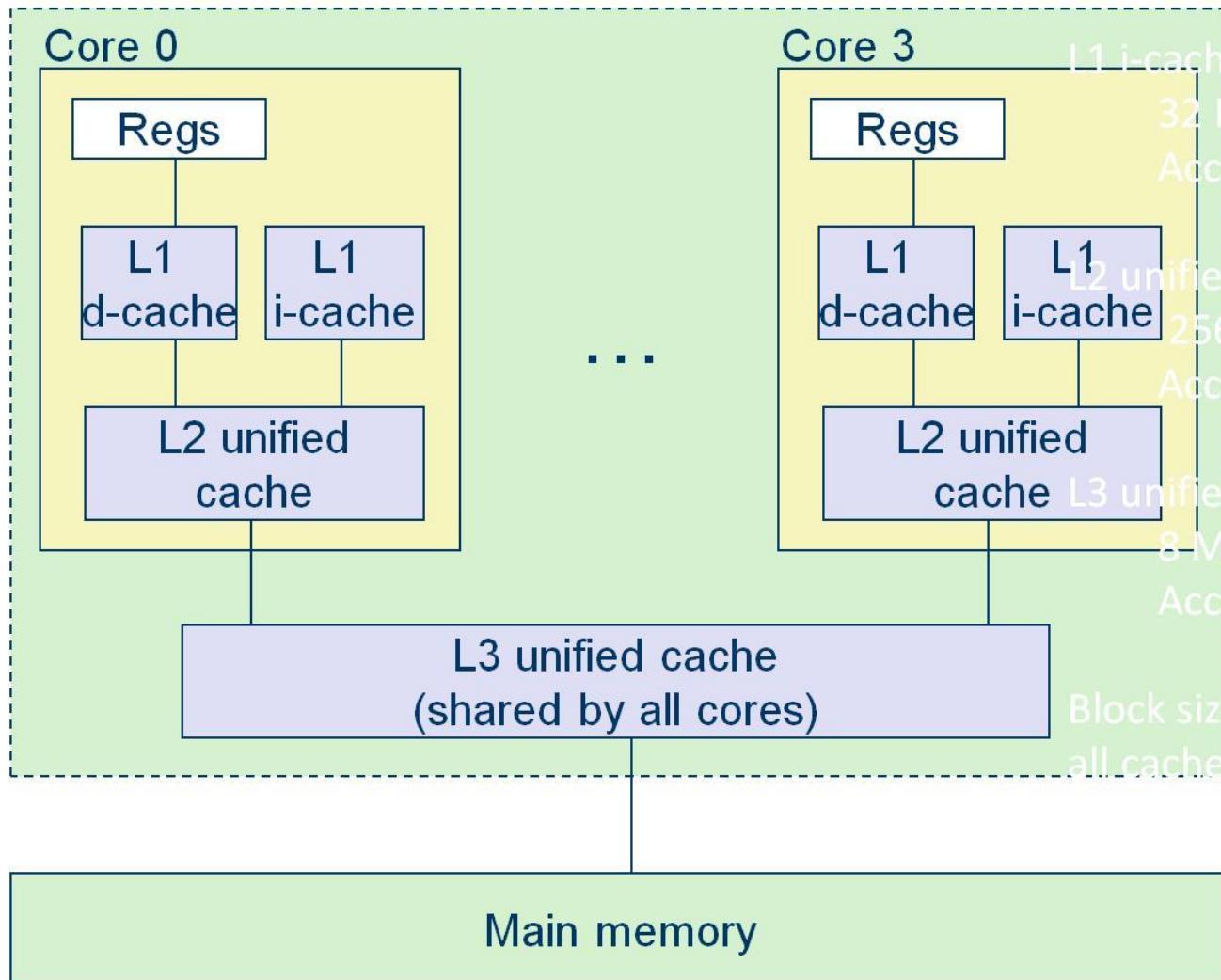
- Losowy (przesunięcia rejestru)
- LRU (Least recently used)
- FIFO



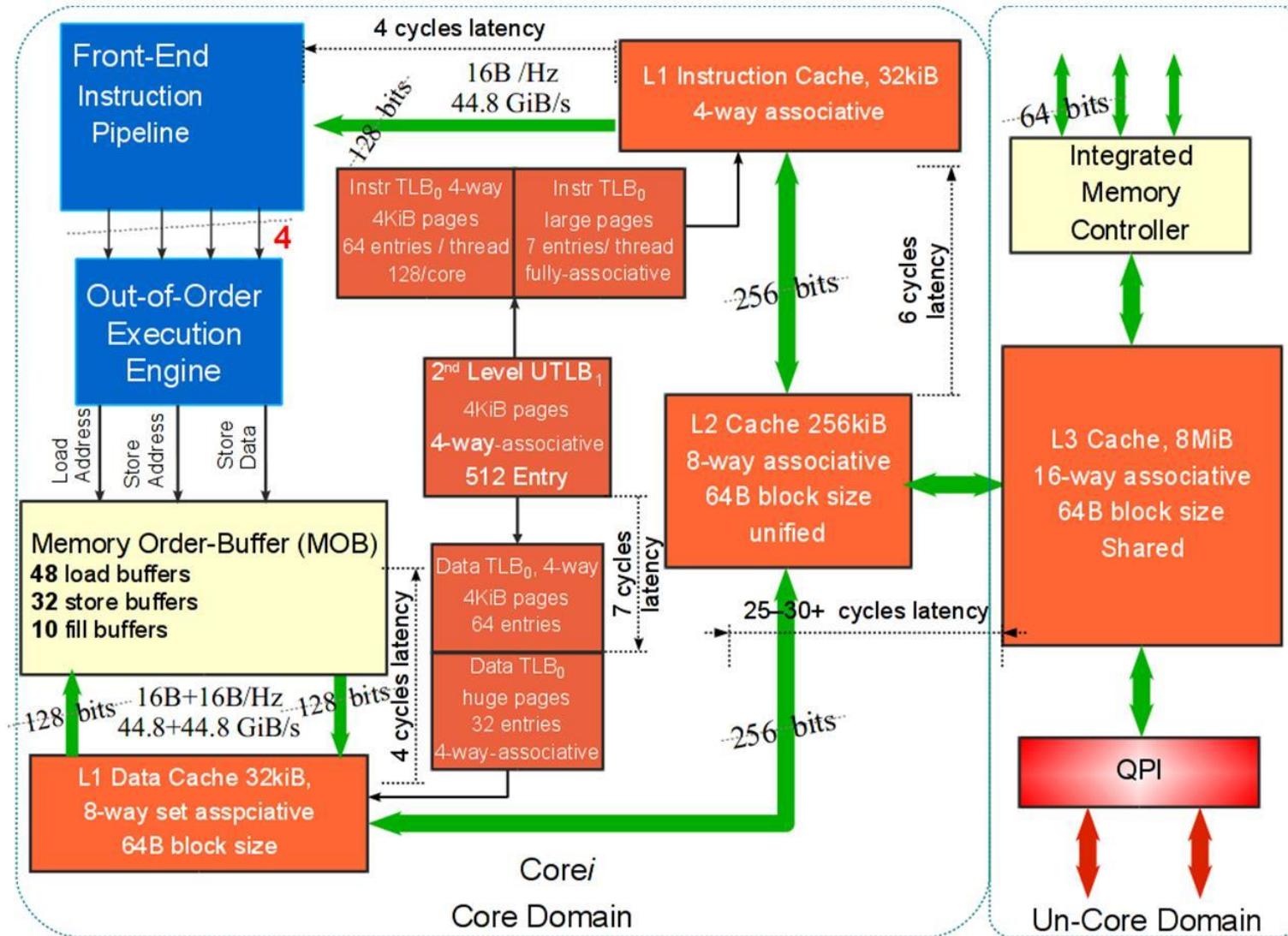
Parametry pamięci podręcznej używanej w procesorach Pentium 4:

- pamięć poziomu L1 dla danych: 8 KB, dostęp 4-kanałowy, linie 64-bajtowe;
- pamięć poziomu L1 dla rozkazów: nie implementowana (stosowna jest natomiast *trace cache* współpracująca z dekoderem rozkazów — w pamięci tej przechowywane są zdekodowane rozkazy w postaci ciągu mikrooperacji)
- pamięć poziomu L2 dla rozkazów i danych: 1 MB, dostęp 8-kanałowy, 64 bajty w linii
- pamięć poziomu L3 stosowana jest tylko w procesorach Intel Xeon, Itanium2 L1 - 32 KB (instrukcje i dane) **L2 - 256 KB L3 - 3 MB, 4 MB, 6 MB i 9 MB**

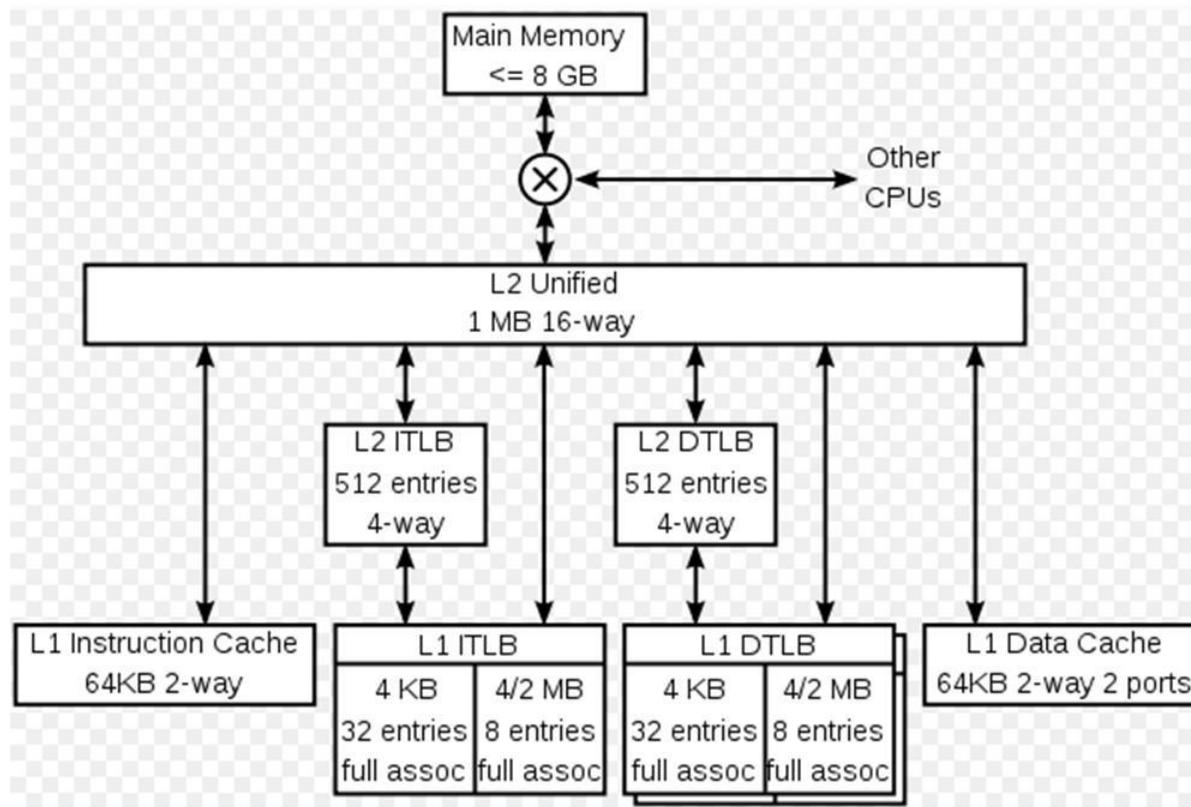
## Processor package



# Pamięć podręczna w Core i7



# Pamięć podręczna w Athlon64





MPU	AMD Opteron	Intrinsity FastMATH	Intel Pentium 4	Intel PXA250	Sun UltraSPARC IV
Instruction set architecture	IA-32, AMD64	MIPS32	IA-32	ARM	SPARC v9
Intended application	server	embedded	desktop	low-power embedded	server
Die size (mm <sup>2</sup> ) (2004)	193	122	217		356
Instructions issued/clock	3	2	3 RISC ops	1	4 × 2
Clock rate (2004)	2.0 GHz	2.0 GHz	3.2 GHz	0.4 GHz	1.2 GHz
Instruction cache	64 KB, 2-way set associative	16 KB, direct mapped	12000 RISC op trace cache (~96 KB)	32 KB, 32-way set associative	32 KB, 4-way set associative
Latency (clocks)	3?	4	4	1	2
Data cache	64 KB, 2-way set associative	16 KB, 1-way set associative	8 KB, 4-way set associative	32 KB, 32-way set associative	64 KB, 4-way set associative
Latency (clocks)	3	3	2	1	2
TLB entries (I/D/L2 TLB)	40/40/512/ 512	16	128/128	32/32	128/512
Minimum page size	4 KB	4 KB	4 KB	1 KB	8 KB
On-chip L2 cache	1024 KB, 16-way set associative	1024 KB, 4-way set associative	512 KB, 8-way set associative	—	—
Off-chip L2 cache	—	—	—	—	16 MB, 2-way set associative
Block size (L1/L2, bytes)	64	64	64/128	32	32



- Pamięć ROM jest nieulotna
- Znajduje się w niej m.in. program inicjujący pracę komputera
- Adres ROM-BIOS F000:0000 do FFFF:FFFF
- Po włączeniu komputera pierwszy rozkaz pobierany jest z adresu F000:FFF0 (5 bajtów kodu FAR JMP do rozpoczęcia POST)
- Po adresem 0000:7C00 umieszczany jest rekord ładowający

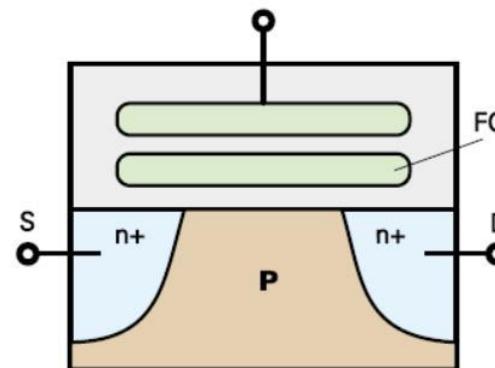
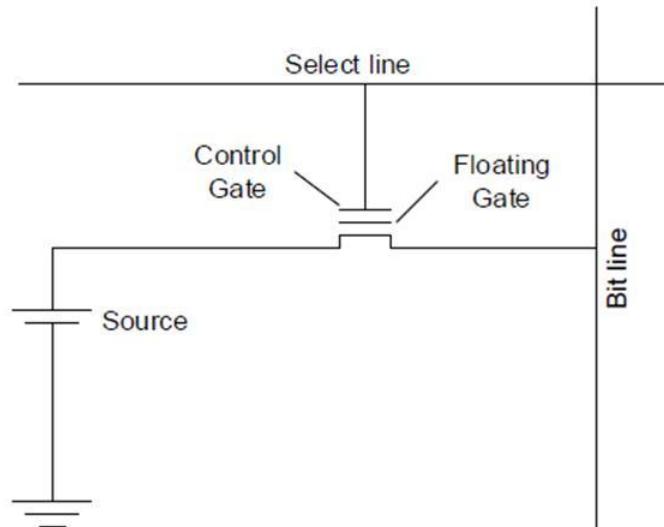


- MROM (lub ROM) – zawartość pamięci jest ustalana w trakcie produkcji i nie może być później zmieniona; czasy dostępu wynoszą 100-450 ns, pojemności 4Kbitów - 16 Mbitów
- PROM – (*programmable ROM*) pamięć jednokrotnie programowalna przez użytkownika
- EPROM – (*erasable PROM*) pamięć wielokrotnie programowalna (kasowanie promieniami ultrafioletowymi)
- EEPROM – (*electrically erasable PROM*) pamięć wielokrotnie programowana (kasowana na drodze elektrycznej) - BIOS
- FLASH – (pamięci błyskowe) możliwe jest programowanie całych bloków, czas programowania pojedynczego bajtu wynosi 10 $\mu$ s; czas dostępu wynosi 45 - 120 ns
- NVRAM – (*non volatile RAM*) połączenie pamięci SRAM z pamięcią EEPROM



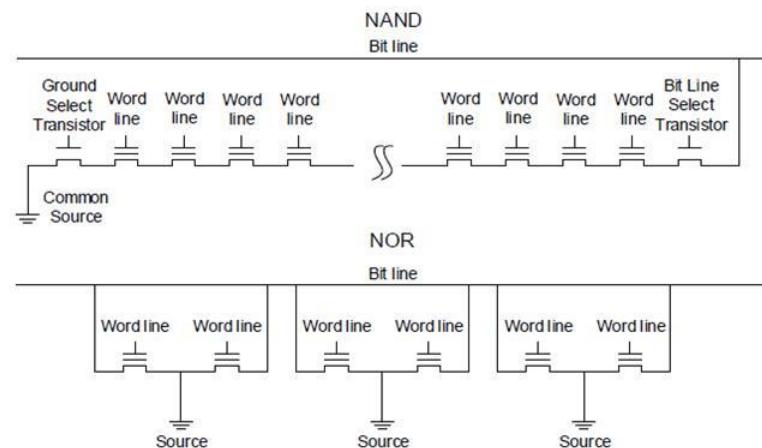
- programowanie pamięci EPROM wykonuje się w urządzeniu nazywanym programatorem (w sposób impulsowy)
- kasowanie pamięci EPROM wymaga użycia źródła promieniowania nadfioletowego UV
- liczba cykli programowania kasowania nie powinna przekraczać 100 - 1000 (dane te nie są podawane przez producentów)
- pamięci EEPROM są kasowalne elektrycznie (wyższym napięciem)
- zapis danych do wybranych komórek pamięci musi być poprzedzony kasowaniem ich zawartości,
- możliwe jest kasowanie zawartości pojedynczych bajtów
- pamięci EEPROM mogą być zazwyczaj nagrywane nawet 100000 razy
- pamięci EEPROM wytwarzane są jako:  
    równoległe  
    szeregowe

# Budowa komórka pamięci flash



	Kasowanie	Programowanie	Odczyt
G	GND	+12V	+5V
S	+12V	GND	GND
D	nie podłączony	+5...+7V	odczyt stanu

- Pływająca bramka (FG) stanowi kondensator o bardzo małej pojemności i upływności – pułapka potencjału
- Bramki NAND łączy się szeregowo, bramki NOR równolegle.
- Pozwala to na dostęp do pojedynczego bitu w NOR i blokowy w NAND
- NAND w porównaniu do pamięci NOR mają krótsze czasy zapisu, odczytu, większą gęstość zapisu i co za tym idzie niższy koszt na MB pojemności. Wytrzymałość liczby cykli zapisu jest do 10 razy większa
- Czas zapisu i odczytu 70-120 ns
- Kasowania sektora 0,7 s (64kB)

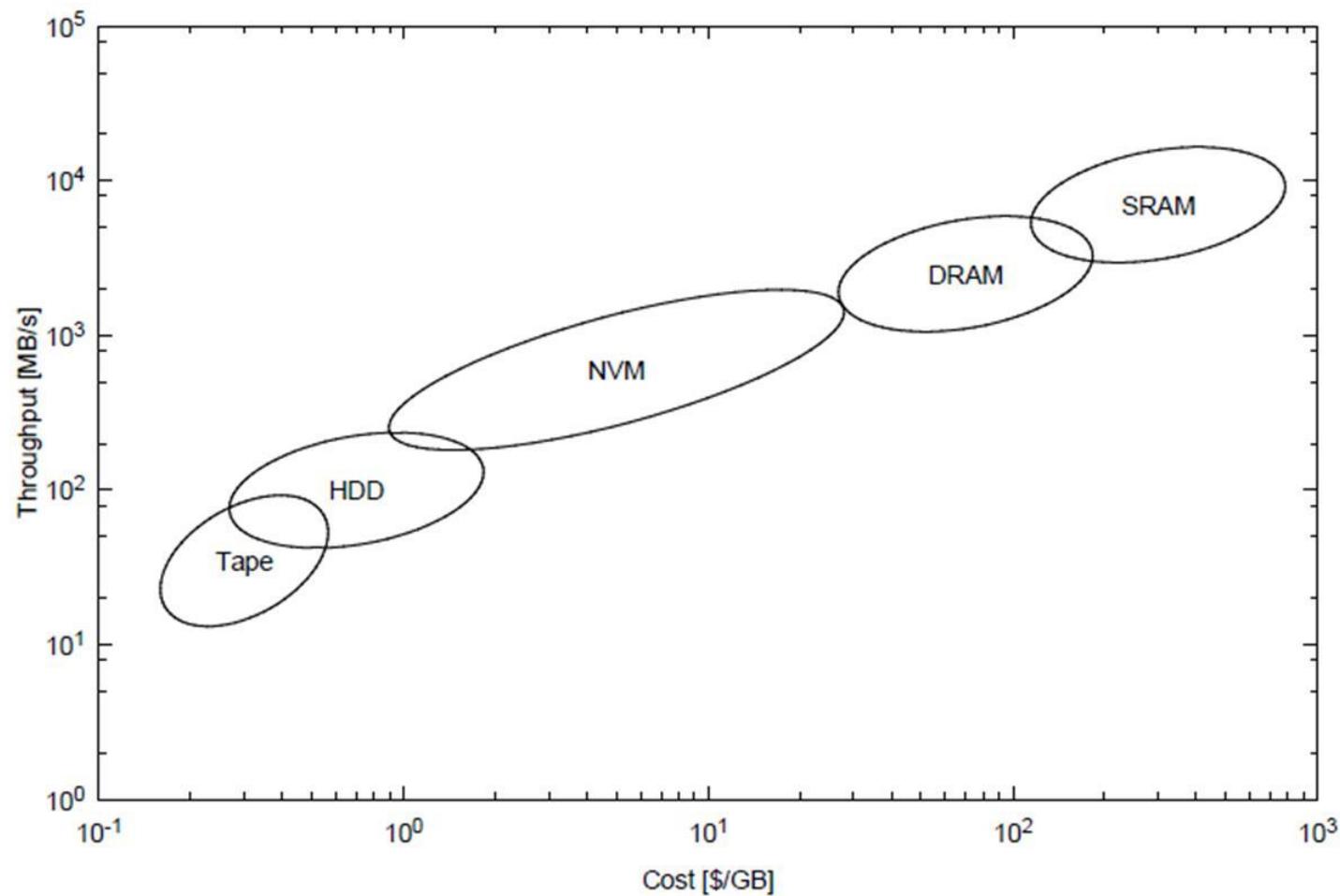




Flash memory type	NAND	NOR
Capacity	~32 Gbit	~1 Gbit
Access method	Sequential	Random
Interface	I/O interface	Full memory interface
Performance	Fast read (serial access cycle) Fast write Fast erase (approx. 2 ms/block)	Fast read (random access) Slow write Slow erase (approx. 1 s/block)
Life span	100,000–1,000,000	10,000–100,000
Price	Low	High

Technology	Read latency	Write latency	Density
DRAM	6-10 ns	6-10 ns	8 Gb/chip
FRAM	8-75 ns	8-75 ns	128 Mb/chip
MRAM	1-10 ns	1-10 ns	32 Mb/chip
STT-RAM	1-10 ns	1-10 ns	2 Mb/chip
NRAM	<10 ns	<10 ns	NA
RRAM	10 ns	20 ns	64 Kb/chip
CBRAM	<50 ns	<50 ns	2 Mb/chip
PRAM	10-100 ns	100-1000 ns	512 Mb/chip
NAND flash	25,000 ns	200,000 ns	64 Gb/chip

# Przepustowość vs koszt





---

HISTORIA MĄDROŚCIĄ  
PRZYSZŁOŚĆ WYZWANIEM