Architektura komputerów

Wykład 7

Jan Kazimirski

Pamięć podręczna

Pamięć komputera - charakterystyka

- Położenie
  - Procesor – rejestry, pamięć podręczna
  - Pamięć wewnętrzna – pamięć podręczna, główna
  - Pamięć zewnętrzna – urządzenia zewnętrzne (np. dysk)
- Pojemność
  - Długość słowa (liczba bitów)
  - Liczba słów

Charakterystyka pamięci c.d.

- Jednostka transferu
  - Słowo
  - Blok
- Sposób dostępu
  - Sekwencyjny
  - Swobodny
  - Skojarzeniowy

Charakterystyka pamięci c.d.

- Wydajność
  - Czas dostępu
  - Czas cyklu
  - Szybkość transferu
- Technologia
  - Półprzewodnikowa
  - Magnetyczna
  - Optyczna

Charakterystyka pamięci c.d.

- Własności fizyczne
  - Ulotna
  - Nieulotna
  - Zapisywalna
  - Tylko do odczytu
- Organizacja (wewnętrzna budowa, ułożenie bitów itp.)

Hierarchia pamięci

- Projektowanie pamięci systemu komputerowego. Podstawowe pytania:

Jak dużo pamięci?

Jak szybka?

Jak droga?

Hierarchia pamięci c.d.

- Rejestry CPU, < 1KB, < 1 ns
- Pamięć podręczna L1, < 128 KB, ~ 1 ns
- Pamięć podręczna L2, 128 KB – 4 MB, 1-2 ns
- Pamięć podręczna L3, ~ 2-30 MB, 2-5 ns
- Pamięć operacyjna, 1-16 GB, 10-50 ns
- System plików, ~ 100-1000 GB, < 10 ms
- Sieć, nośniki wymienne

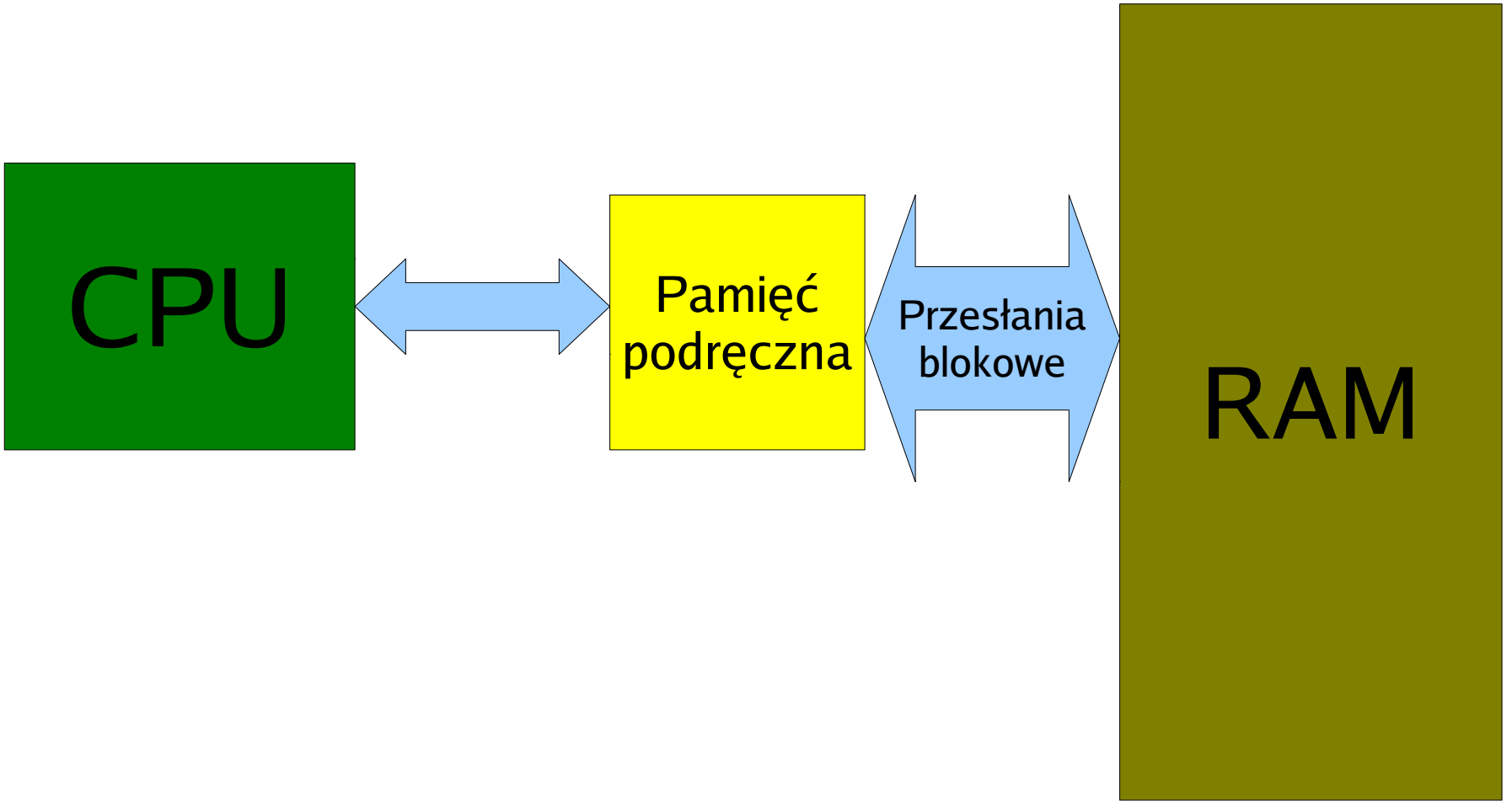
Hierarchia pamięci c.d.

- Pozwala zwiększyć wydajność przy rozsądnych kosztach.
- Wydajne użycie hierarchii pamięci zakłada tzw. Zasadę *lokalności czasowej i przestrzennej odwołań*.

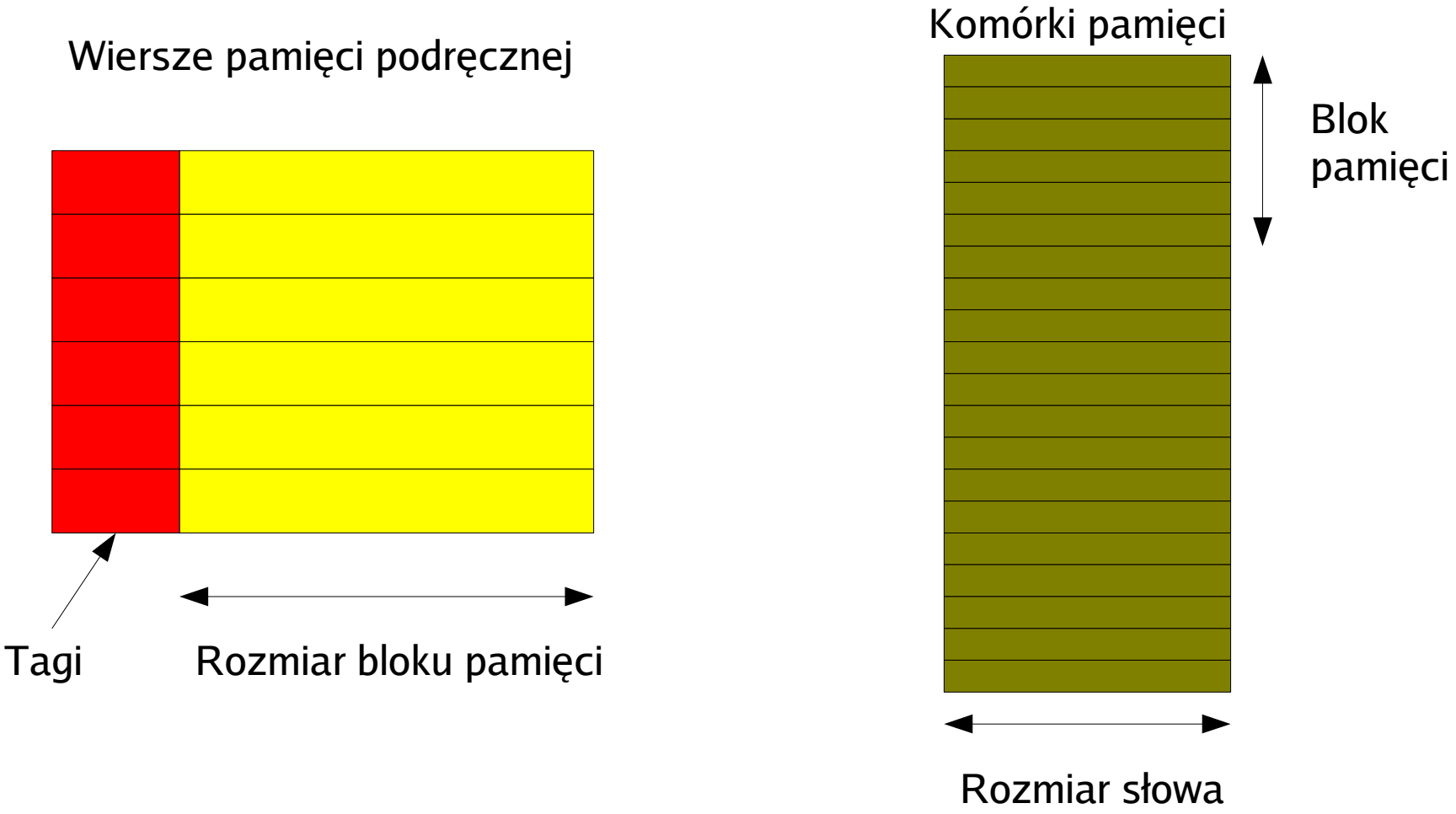
Pamięć podręczna

- Wykorzystanie pamięci podręcznej pozwala osiągnąć dwa cele:
  - Umożliwia zastosowanie najszybszej technologicznie możliwej pamięci w celu szybkiego dostępu
  - Udostępnia pamięć o dużej pojemności do przechowywania dużej liczby danych

Pamięć podręczna c.d.



Pamięć podręczna i RAM



Pamięć podręczna i RAM c.d.

- Pamięć główna – bloki adresowalnych komórek pamięci
- Pamięć podręczna – wiersze o długości odpowiadającej wielkości bloku
- Znaczniki (Tag) wiersza określają który blok jest w pamięci podręcznej.

Działanie pamięci podręcznej

- Przy każdym odwołaniu procesora do pamięci następuje sprawdzenie, czy dana jest w pamięci podręcznej.
  - „**Cache miss**” - chybienie – danej nie ma w pamięci podręcznej. Zostaje przesłana z pamięci głównej i umieszczona w pamięci podręcznej
  - „**Cache hit**” - trafienie – dana zostaje odczytana z pamięci podręcznej

Charakterystyka pamięci podręcznej

- Rozmiar
- Sposób mapowania
- Strategia wymiany
- Strategia zapisu (write through, write back)
- Wielkość bloku
- Sposób zarządzania

Cache - rozmiar

- Trudno określić „optymalny” rozmiar
  - Mała pamięć podręczna – niewielki współczynnik trafień, zmniejszona efektywność
  - Duża pamięć podręczna – czasochłonne wyszukiwanie, zmniejszona efektywność
- Typowe rozmiary pamięci podręcznej: 16-64 KB (L1), 256 KB – 4 MB (L2), 2-4 MB (L3)

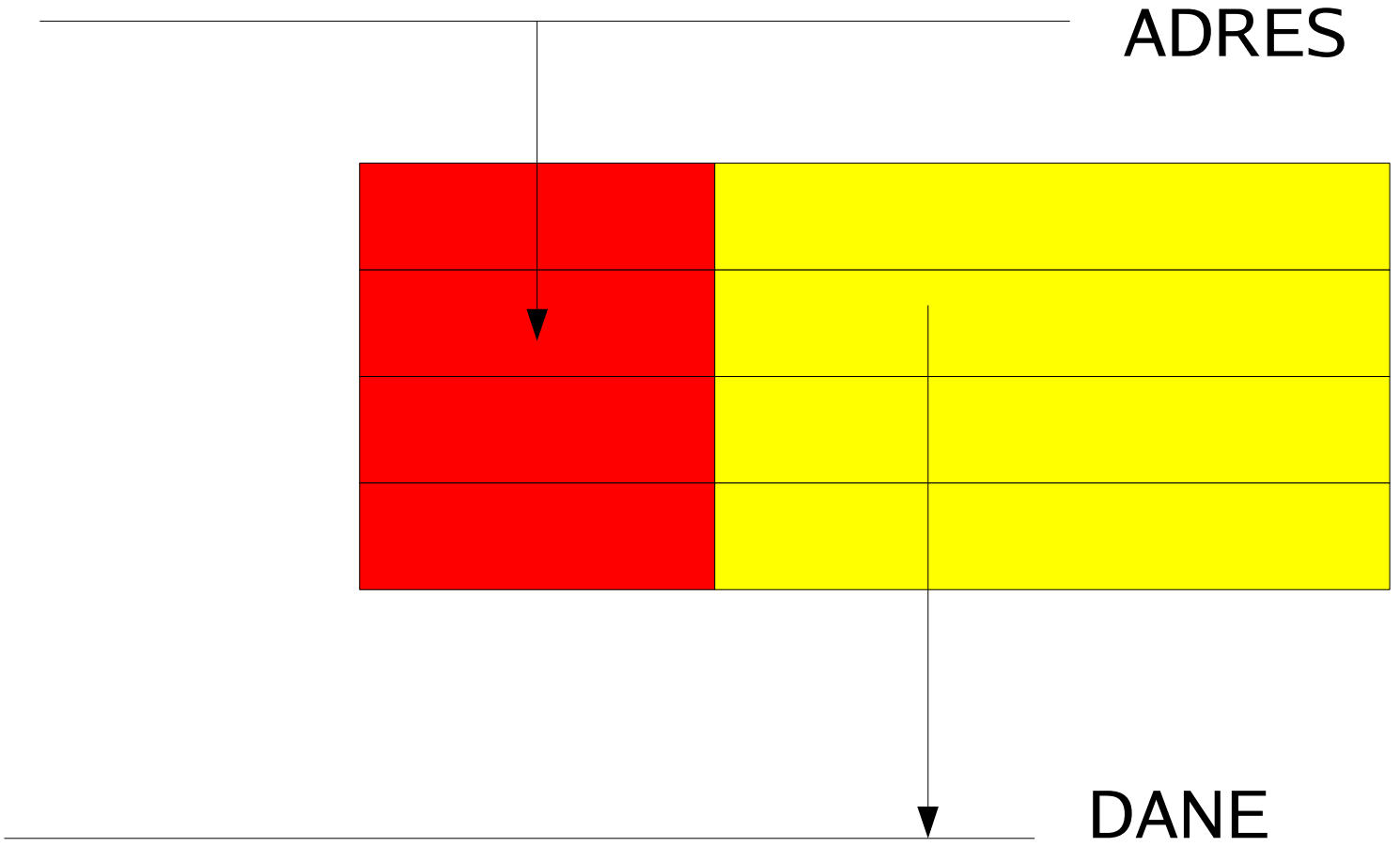
Cache – sposób mapowania

- Odwzorowanie w pełni asocjacyjne
- Odwzorowanie bezpośrednie
- Odwzorowanie zbiorowo-asocjacyjne

Pamięć asocjacyjna

- Adres w pamięci jest znacznikiem w pamięci podręcznej.
- Bloki z pamięci mogą być przechowywane w dowolnych wierszach
- Duża elastyczność przy wymianie wierszy
- Problemy z implementacją (wyszukiwanie znaczników). Niewielki rozmiar. Obecnie nie jest stosowana.

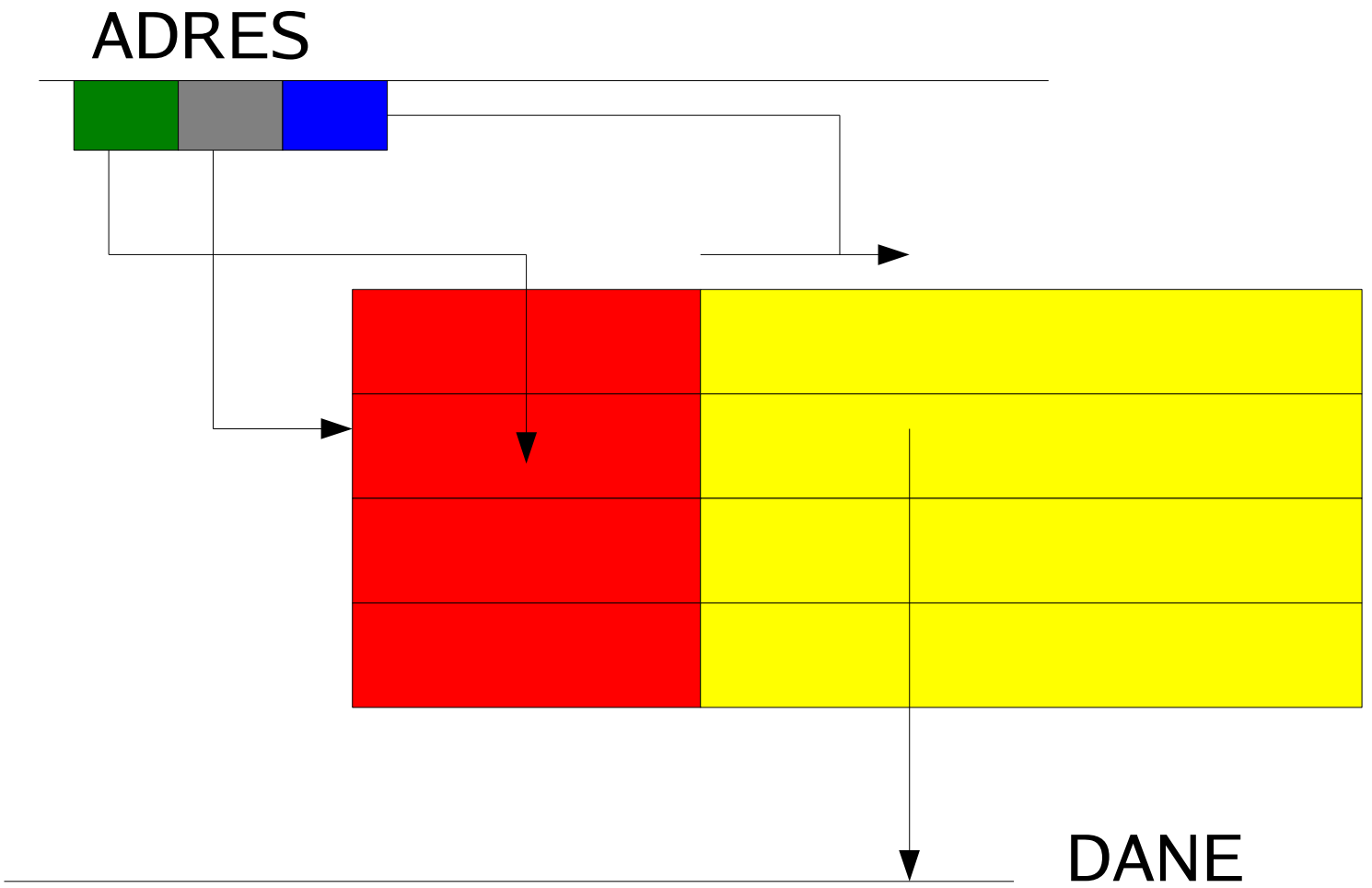
Pamieć asocjacyjna c.d.



Pamięć adresowana bezpośrednio

- Prosta w realizacji, wydajna, duża pojemność
- Sposób adresowania:
  - Najbardziej znaczące bity adresu – znacznik wiersza
  - „Środkowa” część adresu – wybór wiersza
  - Najmniej znaczące bity adresu – wybór bajtu z wiersza
- Mniejsza elastyczność – częściowo (środkowa część adresu) ustalone pozycje bloków.

Pamięć adresowana bezpośrednio



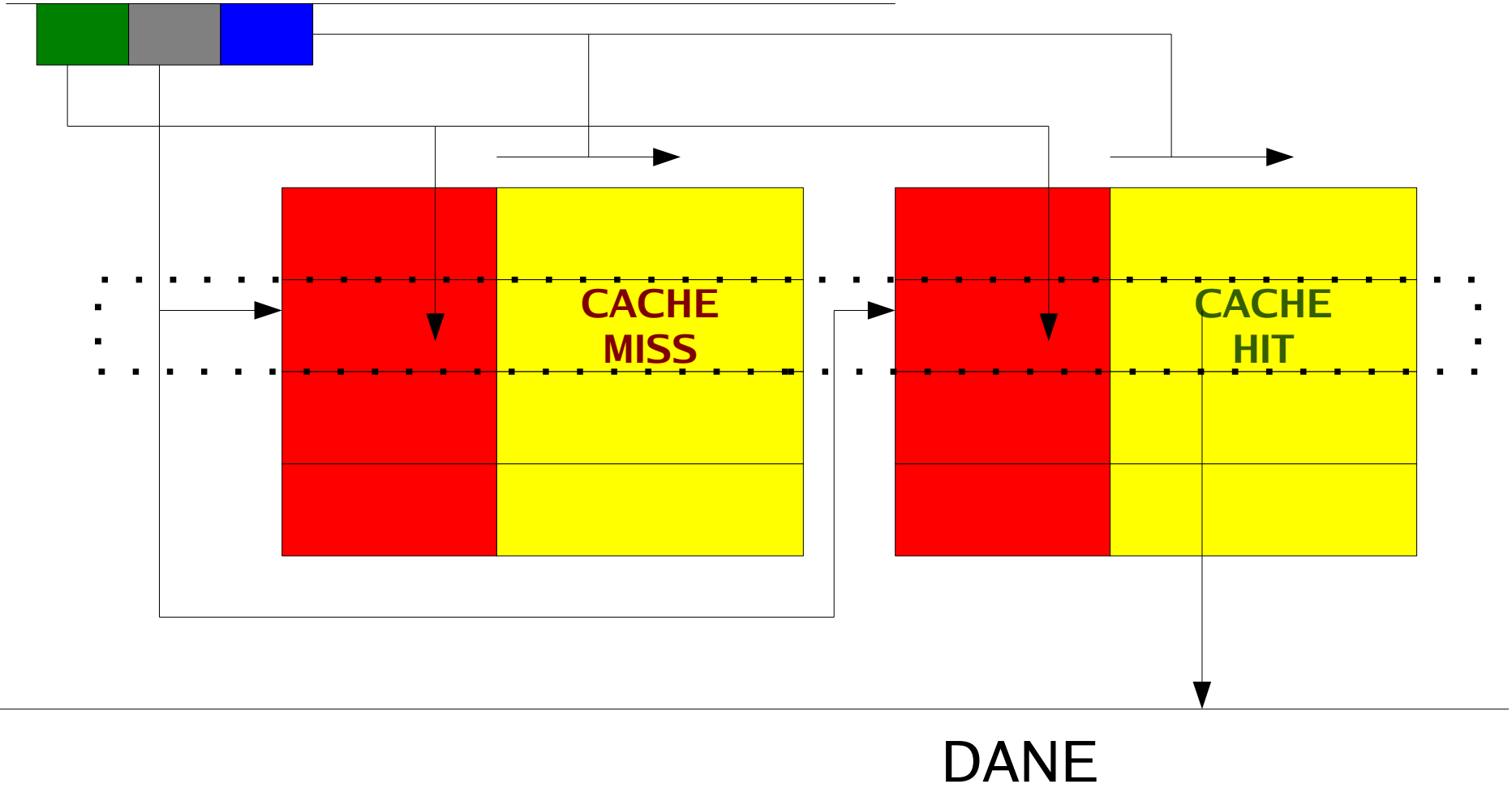
Pamięć zbiorowo asocjacyjna

- Połączenie koncepcji adresowania bezpośredniego i asocjacyjnego
- Połączenie kilku bloków pamięci podręcznej z adresowaniem bezpośrednim
- Dane z określonego adresu mogą być przechowywane w określonym wierszu ale w dowolnym bloku

Pamięć zbiorowo asocjacyjna c.d.

- W czasie odczytu przeszukiwany jest zbiór wierszy odpowiadający danemu adresowi
- Liczba bloków – stopień asocjacyjności.
- Używane również określenia „pamięć k-drożna” (k – stopień asocjacyjności)
- Najczęściej stosowany rodzaj pamięci podręcznej

Pamięć zbiorowo asocjacyjna c.d.



Strategie wymiany

- LRU – Least Recently Used – usuwany jest blok który ostatnio nie był używany.
- FIFO – First In First Out – usuwany jest blok, który jest najdłużej w pamięci podręcznej.
- LFU – Least Frequently Used – usuwany jest blok do którego było najmniej odwołań.
- Wybór losowy.

Strategia zapisu

- Problemy z synchronizacją danych w pamięci podręcznej i głównej:
  - Przed usunięciem wiersza z pamięci podręcznej trzeba uaktualnić dane w pamięci głównej
  - Urządzenia I/O mogą odwoływać się bezpośrednio do pamięci
  - Systemy SMP – problem z synchronizacją pamięci podręcznych procesorów

Write through

- Strategia write through – wszystkie operacje zapisu wykonywane są również do pamięci głównej.
- Pamięć główna jest zawsze zsynchronizowana z pamięcią podręczną
- Może powodować spadek wydajności przy dużej liczbie zapisów do pamięci.

Write back

- Wszystkie operacje zapisu tylko w pamięci podręcznej
- Zapis bloku do pamięci głównej tylko przy wymianie
- Niespójne dane w pamięci głównej i podręcznej – problemy przy obsłudze I/O

Cache i SMP

- Architektura SMP - problem spójności pamięci podręcznych procesorów
- Rozwiązania:
  - Detekcja operacji zapisu (write through)
  - Sprzętowe zapewnianie spójności pamięci podręcznych
  - Wydzielona pamięć wspólna procesorów nie wykorzystująca pamięci podręcznej.

Rozmiar wiersza

- Rozmiar wiersza (bloku danych) pamięci podręcznej ma duży wpływ na wydajność.
- Zwiększanie rozmiaru wiersza – początkowo wzrost efektywności (więcej trafień). Później bez efektu.
- Większy rozmiar bloku = mniej bloków w pamięci podręcznej
- W dużym bloku wiele danych nie będzie użyte (zasada lokalności)
- Typowe rozmiary – 8-32 bajty (64-128 dla HPC)

Pamięć specjalizowana

- Współczesne architektury komputerów często stosują osobne pamięci podręczne dla danych i instrukcji (na niskim poziomie)
- Stosowanie osobnych pamięci podręcznych dla danych i rozkazów zwiększa efektywność działania procesora

Pamięć wielopoziomowa

- Początkowo stosowano pojedyncze pamięci podręczne
- Rozwój koncepcji – pamięci 2 poziomowe
  - Pamięć podręczna w kości procesora (L1)
  - Pamięć podręczna na szynie zewnętrznej (L2)
- Dalsza rozbudowa
  - Cache L2 na osobnej szynie
  - Przesunięcie L2 do kości procesora
  - Wprowadzenie pamięci podręcznej L3

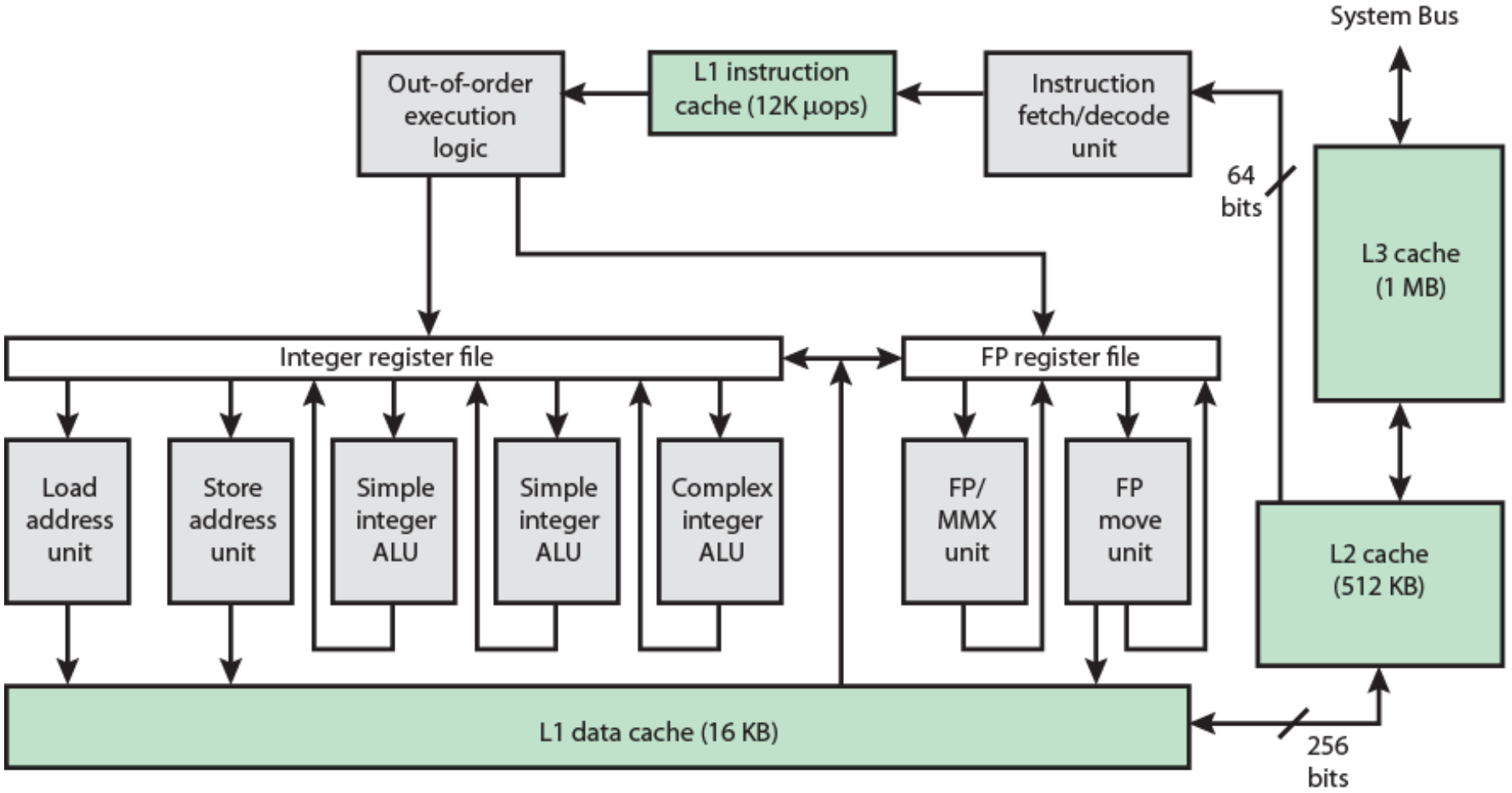
Pamięć podręczna - Intel

- 80386 – brak
- 80486 – 8k (wiersz 16 bajtów, 4-drożna)
- Pentium – L1 na kości, osobno dane i instrukcje
- Pentium III – dodana zewnętrzna pamięć podręczna (L3)
- Pentium 4 – poziomy L1, L2, L3

Pamięć podręczna – Pentium 4

- Poziom L1
  - 8k, 4-drożna, wiersz 64 bajty, osobno dane i rozkazy (mikrokod)
- Poziom L2
  - 256 lub 512k, 8-drożna, wiersz 128 bajtów
- Poziom L3
  - 1 MB, 8-drożna, wiersz 128 bajtów

Pentium 4 - architektura



(Stallings „"Computer architecture..."

Współczynnik trafień

- Określa efektywność pamięci podręcznej
- Wyliczany wzorem:

  h= l.trafień / l.odwołań

- Zależy od:
  - Pojemności pamięci podręcznej
  - Organizacji pamięci podręcznej
  - Wykonywanego programu (!)

Pamięć inkluzywna

- Rozwiązanie stosowane do 2000 r.
- Przepływ danych RAM <-> L2 <-> L1 <-> CPU
- Dane są dublowane
- Efektywna pojemność pamięci podręcznej = pojemność najpojemniejszego poziomu

Pamięć wyłączna

- Pamięć L2 służy do przechowywania danych usuniętych z L1 („victim cache”).
- Efektywna pojemność pomięci podręcznej = suma pojemności poszczególnych warstw
- Rozwiązanie stosowane obecnie (K7, K8, Pentium4, Core)

Podsumowanie

- Charakterystyka pamięci komputera
- Hierarchia pamięci
- Pamięć podręczna
  - Rodzaje
  - Kwestie wydajnościowe
  - Algorytmy zarządzania pamięcią podręczną