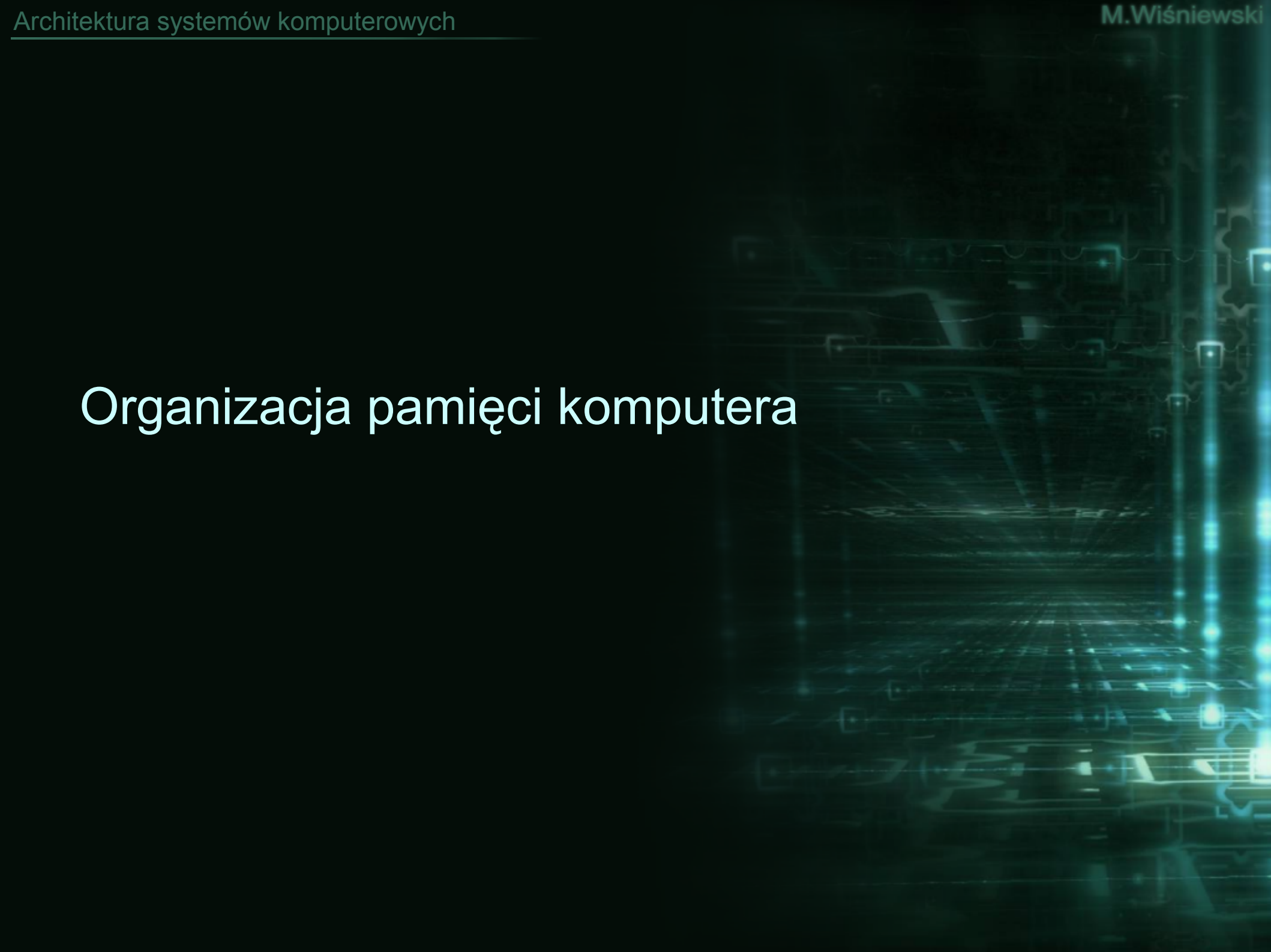
Architektura systemów komputerowych

Mariusz Wiśniewski

Politechnika Świętokrzyska w Kielcach
Katedra Informatyki



Organizacja pamięci komputera

Plan wykładu

1. Podział logiczny pamięci operacyjnej.
2. Pamięć wirtualna.
3. Pamięć podręczna.
4. Pamięć cache procesora.

Cele

Znajomość budowy współczesnego komputera. Rozległa wiedza na temat zastosowania pamięci, struktur pamięciowych, różnych pojęć związanych z pamięcią komputera.

Podział logiczny pamięci operacyjnej

Podział logiczny pamięci...

- wstęp
- segmentacja
- stronicowanie

Konstrukcyjnie pamięć operacyjna jest tablicą słów, które można adresować liniowo, jednak w praktyce ten model pamięci nie jest doskonały. Dlatego dostęp do pamięć jest zazwyczaj wielopoziomowy.

Pojęcia:
- pamięć w komputerze
- mechanizmy dostępu CPU do pamięci

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Konstrukcyjnie pamięć operacyjna jest tablicą słów, które można adresować liniowo, jednak w praktyce ten model pamięci nie jest doskonały. Dlatego dostęp do pamięć jest zazwyczaj wielopoziomowy.

Pojęcia:
- pamięć w komputerze
- mechanizmy dostępu CPU do pamięci

Architektura pamięć komputera zawiera moduły:
- matryce pamiętające – układy pamięci,
- kontroler pamięci,
- magistrale – równoległe, punkt-do-punktu.

Z punktu widzenia CPU pamięć to liniowa struktura danych, w której każdą komórkę można adresować, zapisać lub odczytać. Pamięć RAM jest wykorzystywana jako pamięć operacyjna, w której znajdują się dane i programu. Ze względu na specyfikę pracy komputera, pamięć RAM ulega z czasem fragmentacji, co uniemożliwia sprawną pracę systemu operacyjnego.

Współczesne systemy komputerowe posiadają dwupoziomową organizację pamięci:
- warstwa fizyczna – moduły pamięci,
- poziom logiczny – programowo-sprzętowy, realizujący logiczny model adresowania, zadania ochrony zasobów i zapobiegania fragmentacji pamięci.

M.Wiśniewski

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Konstrukcyjnie pamięć operacyjna jest tablicą słów, które można adresować liniowo, jednak w praktyce ten model pamięci nie jest doskonały. Dlatego dostęp do pamięć jest zazwyczaj wielopoziomowy.

Pojęcia:
- pamięć w komputerze
- mechanizmy dostępu CPU do pamięci

Mikroprocesor może zapewniać dostęp do pamięci na obu poziomach architektonicznych:
- poziom fizyczny – odpowiada rzeczywistym wyjściom procesora, za pomocą których CPU łączy się z komponentami komputera, w tym z pamięcią,
- poziom logiczny – dostępny dla programów, wymaga obecności w CPU modułów realizujących translację adresów logicznych na fizyczne, w ogólnością są to:
  - moduł segmentacji,
  - moduły związane ze stronicowaniem.

  Translacja adresów logicznych na fizyczne jest transparentna dla programów.

Podział logiczny pamięci umożliwia zapobieganiu jej fragmentacji oraz pozwala na łatwą implementację ochrony obszarów pamięci przydzielonych do zadań.

Architektura pamięć komputera zawiera moduły:
- matryce pamiętające – układy pamięci,
- kontroler pamięci,
- magistrale – równoległe, punkt-do-punktu.

Z punktu widzenia CPU pamięć to liniowa struktura danych, w której każdą komórkę można adresować, zapisać lub odczytać. Pamięć RAM jest wykorzystywana jako pamięć operacyjna, w której znajdują się dane i programu. Ze względu na specyfikę pracy komputera, pamięć RAM ulega z czasem fragmentacji, co uniemożliwia sprawną pracę systemu operacyjnego.

Współczesne systemy komputerowe posiadają dwupoziomową organizację pamięci:
- warstwa fizyczna – moduły pamięci,
- poziom logiczny – programowo-sprzętowy, realizujący logiczny model adresowania, zadania ochrony zasobów i zapobiegania fragmentacji pamięci.

Podział logiczny pamięci...

- wstęp
- segmentacja
- stronicowanie

Segmentacja jest jedną z metod logicznego podziału pamięci operacyjnej i w ogólności polega podziale pamięci na segmenty (sekcje).

Definicje:
- segment pamięci
- adresowanie pamięci
- wielozadaniowość

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Segmentacja jest jedną z metod logicznego podziału pamięci operacyjnej i w ogólności polega podziale pamięci na segmenty (sekcje).

Definicje:
- segment pamięci
- adresowanie pamięci
- wielozadaniowość

Segment pamięci zawiera następujące dane:
- adres początku segmentu w pamięci RAM,
- wielkość segmentu (liczba słów),
- atrybuty – odczyt / zapis / wykonanie, itp.

Segment pamięci definiuje system operacyjny, zgodnie z żądaniem aplikacji. Aplikacja otrzymuje jedynie identyfikator segmentu i posługuje się nim w trakcie adresowania pamięci.

Definicje segmentów przechowuje się w specjalnych strukturach danych – w x86 deskryptorach segmentów, tworzących tablice GDT – dla SO oraz LDT dla zadań (programów).

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Segmentacja jest jedną z metod logicznego podziału pamięci operacyjnej i w ogólności polega podziale pamięci na segmenty (sekcje).

Definicje:
- segment pamięci
- adresowanie pamięci
- wielozadaniowość

Segment pamięci zawiera następujące dane:
- adres początku segmentu w pamięci RAM,
- wielkość segmentu (liczba słów),
- atrybuty – odczyt / zapis / wykonanie, itp.

Segment pamięci definiuje system operacyjny, zgodnie z żądaniem aplikacji. Aplikacja otrzymuje jedynie identyfikator segmentu i posługuje się nim w trakcie adresowania pamięci.

Definicje segmentów przechowuje się w specjalnych strukturach danych – w x86 deskryptorach segmentów, tworzących tablice GDT – dla SO oraz LDT dla zadań (programów).

Adresowanie pamięci z segmentacją w x86:

W trakcie operacji w segmentach może pojawić się wyjątek związany z naruszeniem ochrony pamięci.

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

– wstęp

– segmentacja

– stronicowanie

Segmentacja jest jedną z metod logicznego podziału pamięci operacyjnej i w ogólności polega podziale pamięci na segmenty (sekcje).

Definicje:
- segment pamięci
- adresowanie pamięci
- wielozadaniowość

Zazwyczaj segmentacja posiada wbudowane mechanizmy ochrony pamięci. Definicje segmentów ograniczają ich zastosowania poprzez zabronienie odczytu, zapisu lub wykonania programu. Ponadto próba dostępu do obszaru RAM poza segmentem generuje wyjątek.

Powyższe mechanizmy gwarantują ochronę obszarów pamięci należących do zadań, wspierając tym samym wielozadaniowość.

Segment pamięci zawiera następujące dane:
- adres początku segmentu w pamięci RAM,
- wielkość segmentu (liczba słów),
- atrybuty – odczyt / zapis / wykonanie, itp.

Segment pamięci definiuje system operacyjny, zgodnie z żądaniem aplikacji. Aplikacja otrzymuje jedynie identyfikator segmentu i posługuje się nim w trakcie adresowania pamięci.

Definicje segmentów przechowuje się w specjalnych strukturach danych – w x86 deskryptorach segmentów, tworzących tablice GDT – dla SO oraz LDT dla zadań (programów).

Adresowanie pamięci z segmentacją w x86:

W trakcie operacji w segmentach może pojawić się wyjątek związany z naruszeniem ochrony pamięci.

Podział logiczny pamięci...

- wstęp
- segmentacja
- stronicowanie

Mechanizm stronicowania pozwala na odwzorowanie nieciągłych obszarów pamięci fizycznej w ciągły obszar pamięci logicznej.

Pojęcia:
- model stronicowania
- stronicowanie a segmentacja
- mechanizm stronicowania

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Mechanizm stronicowania pozwala na odwzorowanie nieciągłych obszarów pamięci fizycznej w ciągły obszar pamięci logicznej.

Pojęcia:
- model stronicowania
- stronicowanie a segmentacja
- mechanizm stronicowania

W ogólności stronicowanie wymaga dwupoziomowej realizacji dostępu do pamięci:
- poziom logiczny – realizowany przez warstwę programową (system operacyjny),
- poziom fizyczny – odnoszący się do CPU i pamięci fizycznej (RAM).

W praktyce pamięć fizyczną dzieli się na obszary o stałej długości, nazywane stronami. Rolą systemu operacyjnego jest stworzenie tzw. katalogu stron i zarządzanie stronami.

Podział logiczny pamięci...

architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Mechanizm stronicowania pozwala na odwzorowanie nieciągłych obszarów pamięci fizycznej w ciągły obszar pamięci logicznej.

Pojęcia:
- model stronicowania
- stronicowanie a segmentacja
- mechanizm stronicowania

W ogólności stronicowanie wymaga dwupoziomowej realizacji dostępu do pamięci:
- poziom logiczny – realizowany przez warstwę programową (system operacyjny),
- poziom fizyczny – odnoszący się do CPU i pamięci fizycznej (RAM).

W praktyce pamięć fizyczną dzieli się na obszary o stałej długości, nazywane stronami. Rolą systemu operacyjnego jest stworzenie tzw. katalogu stron i zarządzanie stronami.

Podczas adresowania pamięci aplikacja posługuje się adresami logicznymi komórek w RAM. W przypadku procesorów wspierających zarówno stronicowanie i segmentację adres logiczny podlega translacji na adres fizyczny w jednostce MMU, która:
- wpierw wykorzystuje mechanizmy związane z segmentacją,
- następnie stosuje stronicowanie.

Powyższe czynności są transparentne dla aplikacji i po części dla SO.

Podział logiczny pamięci...

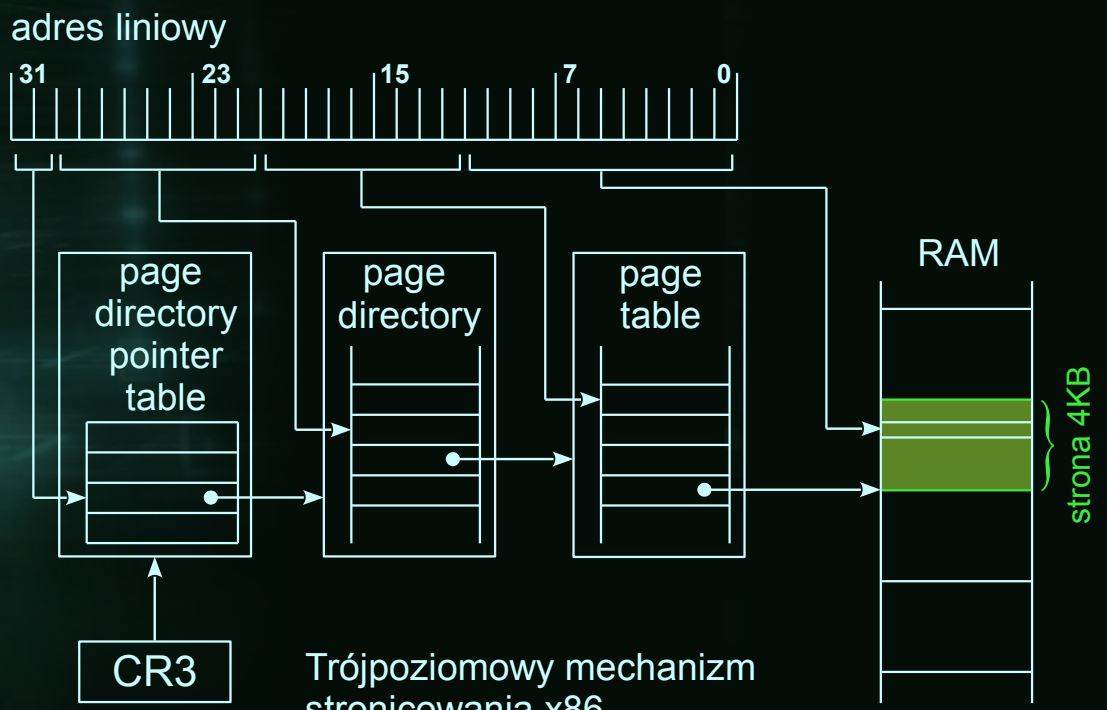
architektura pamięci komputera klasy PC

- wstęp
- segmentacja
- stronicowanie

Mechanizm stronicowania pozwala na odwzorowanie nieciągłych obszarów pamięci fizycznej w ciągły obszar pamięci logicznej.

Pojęcia:
- model stronicowania
- stronicowanie a segmentacja
- mechanizm stronicowania

Stronicowanie wymaga zastosowania dodatkowej logiki układowej i jest najczęściej wbudowane w układ MMU procesora.



Trójpoziomowy mechanizm stronicowania x86.

W ogólności stronicowanie wymaga dwupoziomowej realizacji dostępu do pamięci:
- poziom logiczny – realizowany przez warstwę programową (system operacyjny),
- poziom fizyczny – odnoszący się do CPU i pamięci fizycznej (RAM).

W praktyce pamięć fizyczną dzieli się na obszary o stałej długości, nazywane stronami. Rolą systemu operacyjnego jest stworzenie tzw. katalogu stron i zarządzanie stronami.

Podczas adresowania pamięci aplikacja posługuje się adresami logicznymi komórek w RAM. W przypadku procesorów wspierających zarówno stronicowanie i segmentację adres logiczny podlega translacji na adres fizyczny w jednostce MMU, która:
- wpierw wykorzystuje mechanizmy związane z segmentacją,
- następnie stosuje stronicowanie.

Powyższe czynności są transparentne dla aplikacji i po części dla SO.

Pamięć wirtualna

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

- wstęp
- mechanika

Pamięć aplikacji może mieć charakter wirtualny, powodując wrażenie posiadania dostępu do ciągłego bloku pamięci operacyjnej, podczas gdy rzeczywista przestrzeń jest nieciągła i może w ogóle nie znajdować się w RAM.

Pojęcia:
- pamięć wirtualna
- mechanizmy
- przestrzeń adresowa

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

- wstęp
- mechanika

Pamięć aplikacji może mieć charakter wirtualny, powodując wrażenie posiadania dostępu do ciągłego bloku pamięci operacyjnej, podczas gdy rzeczywista przestrzeń jest nieciągła i może w ogóle nie znajdować się w RAM.

Pojęcia:

- pamięć wirtualna
- mechanizmy
- przestrzeń adresowa

Pamięć wirtualna jest bezpośrednio związana ze stronicowaniem. Można wyróżnić dwie techniki wirtualizacji pamięci operacyjnej:

- prowadzące do eliminowania nieciągłości pamięci,
- zwiększające dostępne zasoby pamięciowe.

W praktyce stosuje się obie techniki, w drugim przypadku korzystając z zasobów pamięci stałej – dysków HDD/SSD.

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

– wstęp

– mechanika

Pamięć aplikacji może mieć charakter wirtualny, powodując wrażenie posiadania dostępu do ciągłego bloku pamięci operacyjnej, podczas gdy rzeczywista przestrzeń jest nieciągła i może w ogóle nie znajdować się w RAM.

Pojęcia:
- pamięć wirtualna
- mechanizmy
- przestrzeń adresowa

Pamięć wirtualna jest bezpośrednio związana ze stronicowaniem. Można wyróżnić dwie techniki wirtualizacji pamięci operacyjnej:
- prowadzące do eliminowania nieciągłości pamięci,
- zwiększające dostępne zasoby pamięciowe.

W praktyce stosuje się obie techniki, w drugim przypadku korzystając z zasobów pamięci stałej – dysków HDD/SSD.

Obsługa pamięci wirtualnej wymaga odpowiednio przygotowanego CPU, który musi:
- obsługiwać stronicowanie,
- przechowywać dodatkowe informacje o stronach (w katalogu stron, TLB) – odpowiednio:
  - bit obecności strony (present),
  - bit użycia strony (accessed),
  - bit modyfikacji strony (dirty / modified).,
- posiadać mechanizmy pozwalające na wymianę stron.

M.Wiśniewski

Pamięć wirtualna

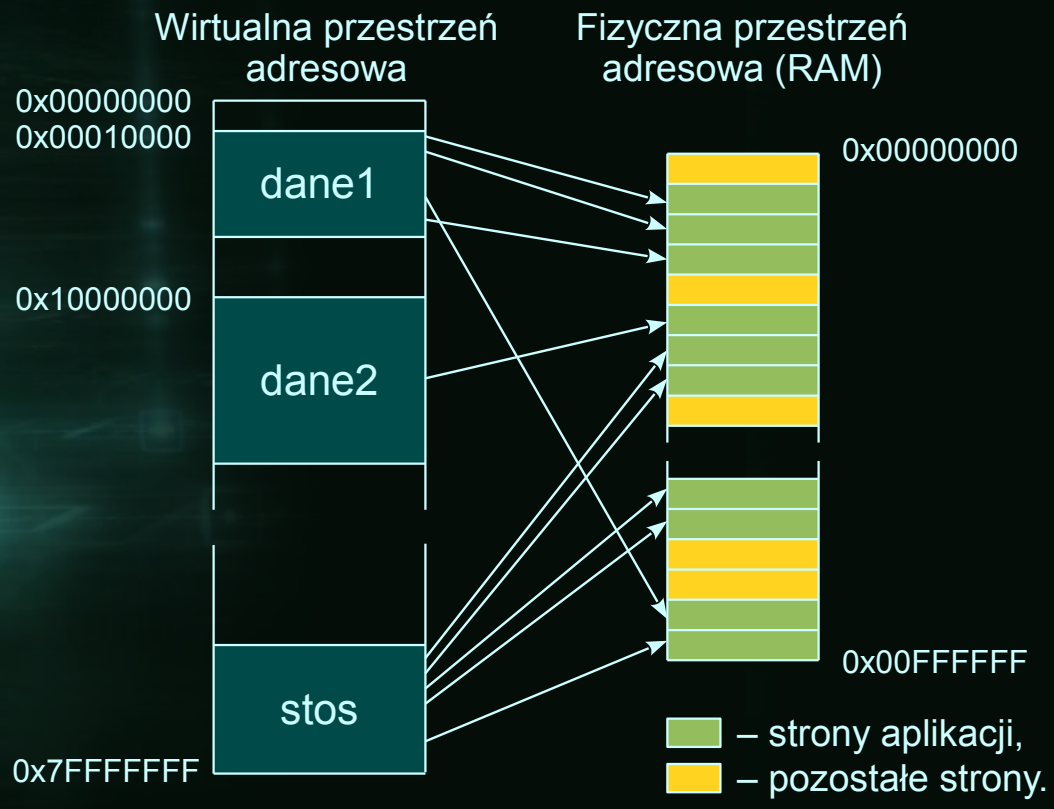
mechanizm zwiększający zasoby pamięciowe procesów

– wstęp

– mechanika

Pamięć aplikacji może mieć charakter wirtualny, powodując wrażenie posiadania dostępu do ciągłego bloku pamięci operacyjnej, podczas gdy rzeczywista przestrzeń jest nieciągła i może w ogóle nie znajdować się w RAM.

Pojęcia:
- pamięć wirtualna
- mechanizmy
- przestrzeń adresowa



Pamięć wirtualna jest bezpośrednio związana ze stronicowaniem. Można wyróżnić dwie techniki wirtualizacji pamięci operacyjnej:
- prowadzące do eliminowania nieciągłości pamięci,
- zwiększające dostępne zasoby pamięciowe.

W praktyce stosuje się obie techniki, w drugim przypadku korzystając z zasobów pamięci stałej – dysków HDD/SSD.

Obsługa pamięci wirtualnej wymaga odpowiednio przygotowanego CPU, który musi:
- obsługiwać stronicowanie,
- przechowywać dodatkowe informacje o stronach (w katalogu stron, TLB) – odpowiednio:
  - bit obecności strony (present),
  - bit użycia strony (accessed),
  - bit modyfikacji strony (dirty / modified).,
- posiadać mechanizmy pozwalające na wymianę stron.

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

- wstęp
- mechanika

Obsługa pamięci wirtualnej wymaga posiadania odpowiednio wyposażonego procesora oraz systemu operacyjnego wspierającego tą technologię.

Pojęcia:
- tablica stron
- wymagania dla SO
- proces translacji adresów

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

- wstęp
- mechanika

Obsługa pamięci wirtualnej wymaga posiadania odpowiednio wyposażonego procesora oraz systemu operacyjnego wspierającego tą technologię.

Pojęcia:
- tablica stron
- wymagania dla SO
- proces translacji adresów

W ogólności wpisy w tablicy stron wskazują miejsce w pamięci RAM, gdzie znajduje się strona lub zawierają informację, że wybrana strona jest zapisana w pamięci masowej.

Zarządzaniem tablicą stron zajmuje się system operacyjny.

M.Wiśniewski

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

– wstęp

– mechanika

Obsługa pamięci wirtualnej wymaga posiadania odpowiednio wyposażonego procesora oraz systemu operacyjnego wspierającego tą technologię.

Pojęcia:

- tablica stron
- wymagania dla SO
- proces translacji adresów

W ogólności wpisy w tablicy stron wskazują miejsce w pamięci RAM, gdzie znajduje się strona lub zawierają informację, że wybrana strona jest zapisana w pamięci masowej.

Zarządzaniem tablicą stron zajmuje się system operacyjny.

Podczas obsługi pamięci wirtualnej system operacyjny ma następujące zadania:

- tworzenie i modyfikowanie tablic stron,
- obsługa wyjątków związanych ze stronicowanie pamięci wirtualnej,
- wymiana stron pamięci wirtualnej.

Proces wymiany stron nie można wykonać dla tzw. stron krytycznych, które należą do:

- procedur obsługi przerwań opartych o wskaźniki do kodu dla poszczególnych przerwań,
- tablice stron,
- bufory danych wykorzystywane w komunikacji z urządzeniami komputera,
- fragmenty jądra systemu operacyjnego.

M.Wiśniewski

Pamięć wirtualna

mechanizm zwiększający zasoby pamięciowe procesów

– wstęp

– mechanika

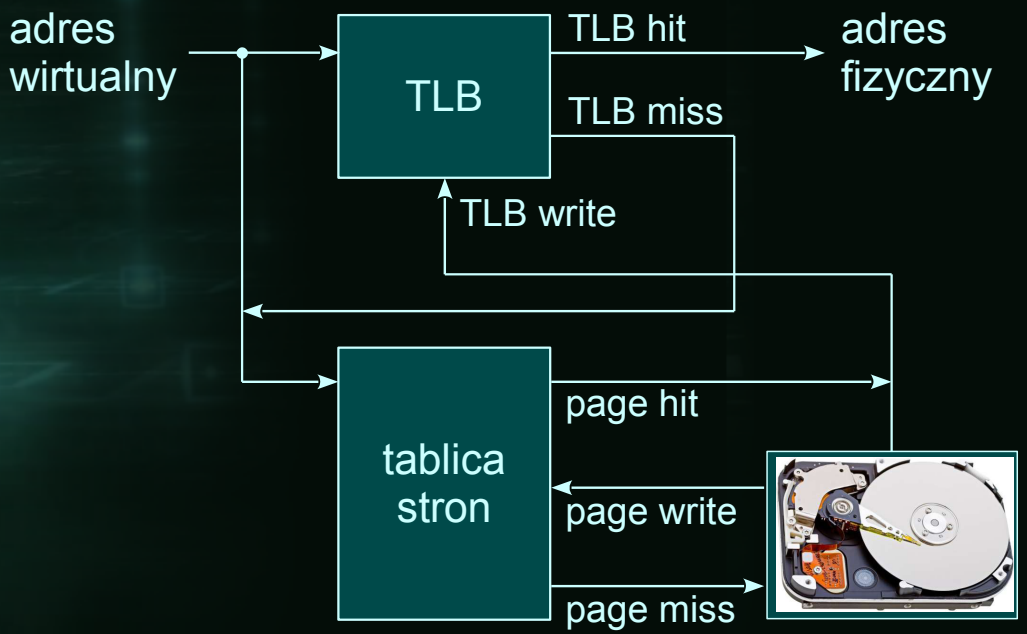
Obsługa pamięci wirtualnej wymaga posiadania odpowiednio wyposażonego procesora oraz systemu operacyjnego wspierającego tą technologię.

Pojęcia:
- tablica stron
- wymagania dla SO
- proces translacji adresów

W ogólności wpisy w tablicy stron wskazują miejsce w pamięci RAM, gdzie znajduje się strona lub zawierają informację, że wybrana strona jest zapisana w pamięci masowej.

Zarządzaniem tablicą stron zajmuje się system operacyjny.

W procesie translacji adresów wirtualnych na fizyczne bierze udział tablica TLB (translation lookaside buffer), będącą rodzajem pamięci asocjacyjnej.



Podczas obsługi pamięci wirtualnej system operacyjny ma następujące zadania:
- tworzenie i modyfikowanie tablic stron,
- obsługa wyjątków związanych ze stronicowanie pamięci wirtualnej,
- wymiana stron pamięci wirtualnej.

Proces wymiany stron nie można wykonać dla tzw. stron krytycznych, które należą do:
- procedur obsługi przerwań opartych o wskaźniki do kodu dla poszczególnych przerwań,
- tablice stron,
- bufory danych wykorzystywane w komunikacji z urządzeniami komputera,
- fragmenty jądra systemu operacyjnego.

Pamięć podręczna

Pamięć podręczna

techniki optymalizacji dostępu do danych

- wstęp
- przykłady zastosowań

Technika buforowania danych nie jest jedynie domeną techniki mikroprocesorowej. Buforowanie można spotkać w wielu systemach komputerowych.

Pojęcia:
- cechy pamięci podręcznej
- strategie i heurystyki

Pamięć podręczna

techniki optymalizacji dostępu do danych

- wstęp
- przykłady zastosowań

Technika buforowania danych nie jest jedynie domeną techniki mikroprocesorowej. Buforowanie można spotkać w wielu systemach komputerowych.

Pojęcia:
- cechy pamięci podręcznej
- strategie i heurystyki

W ogólności pamięć podręczna jest buforem o większej przepustowości i mniejszym czasie dostępu niż nośnik danych z którym jest połączona.

Podstawą funkcjonowania pamięci podręcznej jest przypuszczenie, że skoro nastąpiło odwołanie do pewnych danych, to może to nastąpić ponownie w najbliższej przyszłości.

Najprostsze systemy pamięci podręcznej bazują na powyższej zasadzie, najbardziej złożone często analizują dane, informując system nadrzędny, który może podjąć decyzje wpływające na efektywność buforowania.

Moduł pamięci podręcznej powinien:
- być możliwie automatyczny,
- zwiększać wydajność systemu.

W projektowaniu pamięci podręcznej bierze się pod uwagę:
- specyfikę buforowanego systemu,
- rodzaj buforowanych danych.

M.Wiśniewski

Pamięć podręczna

techniki optymalizacji dostępu do danych

– wstęp

– przykłady zastosowań

Technika buforowania danych nie jest jedynie domeną techniki mikroprocesorowej. Buforowanie można spotkać w wielu systemach komputerowych.

Pojęcia:
- cechy pamięci podręcznej
- strategie i heurystyki

W ogólności istotą buforowania jest zmniejszanie czasu dostępu do danych. Aby uzyskać zadowalające wyniki układy sterujące pamięcią podręczną stosują pewne strategie pracy oparte na prawdopodobieństwie wystąpienia żądania do zbuforowanych danych. Najczęściej stosuje się:
- lokalność czasową:
  - jeśli żądano dostępu obiektu, to możliwe, że żądanie wkrótce zostanie powtórzone.
- lokalność przestrzenną:
  - jeśli żądano obiektu o określonym adresie, to możliwe, że zostaną również zażądane dane znajdujące się w jego pobliżu.
- atrakcyjność obiektów:
  - zazwyczaj obiekty występujące w danym środowisku powszechniej są żądane częściej, niż obiekty posiadające cechy specjalne.

Strategie optymalizacji buforowania mogą opierać się na kombinacjach powyższych heurystyk.

W ogólności pamięć podręczna jest buforem o większej przepustowości i mniejszym czasie dostępu niż nośnik danych z którym jest połączona.

Podstawą funkcjonowania pamięci podręcznej jest przypuszczenie, że skoro nastąpiło odwołanie do pewnych danych, to może to nastąpić ponownie w najbliższej przyszłości.

Najprostsze systemy pamięci podręcznej bazują na powyższej zasadzie, najbardziej złożone często analizują dane, informując system nadrzędny, który może podjąć decyzje wpływające na efektywność buforowania.

Moduł pamięci podręcznej powinien:
- być możliwie automatyczny,
- zwiększać wydajność systemu.

W projektowaniu pamięci podręcznej bierze się pod uwagę:
- specyfikę buforowanego systemu,
- rodzaj buforowanych danych.

Pamięć podręczna

techniki optymalizacji dostępu do danych

- wstęp
- przykłady zastosowań

Pamięć podręczna jest w zasadzie elementem wszystkich systemów komputerowych, zarówno programowych, jak i czysto sprzętowych.

Pamięć podręczna:
- w procesorze
- w dysku HDD
- systemu plików
- przeglądarki www

Pamięć podręczna

techniki optymalizacji dostępu do danych

- wstęp
- przykłady zastosowań

Pamięć podręczna jest w zasadzie elementem wszystkich systemów komputerowych, zarówno programowych, jak i czysto sprzętowych.

Pamięć podręczna:

- w procesorze
- w dysku HDD
- systemu plików
- przeglądarki www

Współczesne procesory wykonują instrukcje znacznie szybciej niż pracuje pamięć RAM, a pamięć podręczna CPU przyspiesza dostęp do pamięci operacyjnej. Obecnie procesory posiadają 2 lub 3 poziomy pamięci podręcznej.

Zazwyczaj pamięć pierwszego poziomu jest zintegrowana z CPU, a pozostałe poziomy znajdują się w tej samej obudowie co rdzeń procesora.

Pamięć podręczna

techniki optymalizacji dostępu do danych

- wstęp
- przykłady zastosowań

Pamięć podręczna jest w zasadzie elementem wszystkich systemów komputerowych, zarówno programowych, jak i czysto sprzętowych.

Pamięć podręczna:

- w procesorze
- w dysku HDD
- systemu plików
- przeglądarki www

Współczesne procesory wykonują instrukcje znacznie szybciej niż pracuje pamięć RAM, a pamięć podręczna CPU przyspiesza dostęp do pamięci operacyjnej. Obecnie procesory posiadają 2 lub 3 poziomy pamięci podręcznej.

Zazwyczaj pamięć pierwszego poziomu jest zintegrowana z CPU, a pozostałe poziomy znajdują się w tej samej obudowie co rdzeń procesora.

Pamięć podręczna dysku twardego może mieć pojemność wielu megabajtów i jest zazwyczaj podzielona na dwie części:

- dla podsystemu odczytu z wyprzedzeniem – zazwyczaj zajmuje znaczną część pamięci podręcznej dysku,
- dla podsystemu zapisu z opóźnieniem.

Buforowany odczyt danych zazwyczaj stosuje strategię związaną z heurystyką lokalności przestrzennej danych, a jej efektywność znacznie zwiększa defragmentacja dysku.

M.Wiśniewski

Pamięć podręczna

techniki optymalizacji dostępu do danych

– wstęp

– przykłady zastosowań

Pamięć podręczna jest w zasadzie elementem wszystkich systemów komputerowych, zarówno programowych, jak i czysto sprzętowych.

Pamięć podręczna:

- w procesorze
- w dysku HDD
- systemu plików
- przeglądarki www

Współczesne procesory wykonują instrukcje znacznie szybciej niż pracuje pamięć RAM, a pamięć podręczna CPU przyspiesza dostęp do pamięci operacyjnej. Obecnie procesory posiadają 2 lub 3 poziomy pamięci podręcznej.

Zazwyczaj pamięć pierwszego poziomu jest zintegrowana z CPU, a pozostałe poziomy znajdują się w tej samej obudowie co rdzeń procesora.

Pamięć podręczna dysku twardego może mieć pojemność wielu megabajtów i jest zazwyczaj podzielona na dwie części:

- dla podsystemu odczytu z wyprzedzeniem – zazwyczaj zajmuje znaczną część pamięci podręcznej dysku,
- dla podsystemu zapisu z opóźnieniem.

Buforowany odczyt danych zazwyczaj stosuje strategię związaną z heurystyką lokalności przestrzennej danych, a jej efektywność znacznie zwiększa defragmentacja dysku.

Pamięć podręczna systemu plików jest częścią systemu operacyjnego – zazwyczaj jest realizowana w pamięci RAM, a jej zadaniem jest zmniejszenie liczby odwołań do stosunkowo wolnej pamięci masowej. W pamięci podręcznej systemu plików buforuje się:

- metadane systemu plików,
- struktury katalogów i plików,
- tablice alokacji plików, np. FAT.

M.Wiśniewski

Pamięć podręczna

techniki optymalizacji dostępu do danych

– wstęp

– przykłady zastosowań

Pamięć podręczna jest w zasadzie elementem wszystkich systemów komputerowych, zarówno programowych, jak i czysto sprzętowych.

Pamięć podręczna:

- w procesorze
- w dysku HDD
- systemu plików
- przeglądarki www

Współczesne procesory wykonują instrukcje znacznie szybciej niż pracuje pamięć RAM, a pamięć podręczna CPU przyspiesza dostęp do pamięci operacyjnej. Obecnie procesory posiadają 2 lub 3 poziomy pamięci podręcznej.

Zazwyczaj pamięć pierwszego poziomu jest zintegrowana z CPU, a pozostałe poziomy znajdują się w tej samej obudowie co rdzeń procesora.

Przeglądarki internetowe przechowują odwiedzane strony witryn wykorzystując jako pamięć podręczną lokalny dysk HDD/SSD. W tym przypadku dostęp do pamięci masowej jest szybszy niż pobieranie tej samej treści z sieci Internet. Przeglądarka aktualizuje swoją pamięć podręczną wysyłając do właściwego serwisu krótkie zapytania o aktualność podstron witryn.

Pamięć podręczna dysku twardego może mieć pojemność wielu megabajtów i jest zazwyczaj podzielona na dwie części:

- dla podsystemu odczytu z wyprzedzeniem – zazwyczaj zajmuje znaczną część pamięci podręcznej dysku,
- dla podsystemu zapisu z opóźnieniem.

Buforowany odczyt danych zazwyczaj stosuje strategię związaną z heurystyką lokalności przestrzennej danych, a jej efektywność znacznie zwiększa defragmentacja dysku.

Pamięć podręczna systemu plików jest częścią systemu operacyjnego – zazwyczaj jest realizowana w pamięci RAM, a jej zadaniem jest zmniejszenie liczby odwołań do stosunkowo wolnej pamięci masowej. W pamięci podręcznej systemu plików buforuje się:

- metadane systemu plików,
- struktury katalogów i plików,
- tablice alokacji plików, np. FAT.

Pamięć cache procesora

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:

- struktura cache
- poziomy
- wydajność

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

gdzie:
- tag jest częścią adresu komórek RAM,
- blok danych zawiera dane pobrane z komórki pamięci operacyjnej,
- flagi zawierają bity stanu wiersza cache.

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

gdzie:
- tag odpowiada polu tag z pamięci cache,
- indeks jest adresem wiersza w cache,
- offset bloku odpowiada adresowi słowa wewnątrz wiersza cache.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Pojemność cache jest znacznie mniejsza niż pamięci RAM. Z tego powodu podczas pracy z pamięcią cache mogą wystąpić dwa zdarzenia:
- cache hit – dane znajduje się w cache,
- cache miss – spowodowane brakiem danych w cache.

Zwiększenie wydajności przez cache:
- średni czas dostępu:

$$T = T_{RAM} - (T_{RAM} - T_C) * h$$

gdzie:
- T_C – czas dostępu do cache (bardzo mała wartość),
- T_{RAM} – czas dostępu do RAM (duża wartość),
- h – współczynnik trafień (cache hit).

- współczynnik chybień:

$$M_{RATE} = 1 - h$$

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

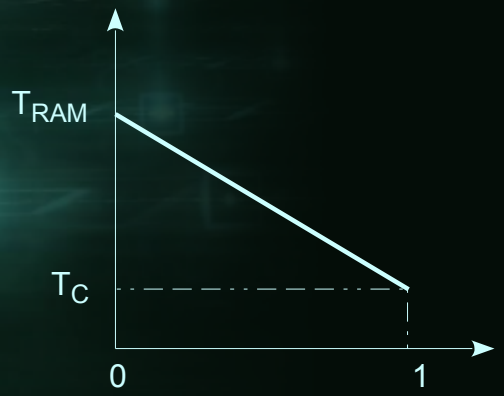
- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

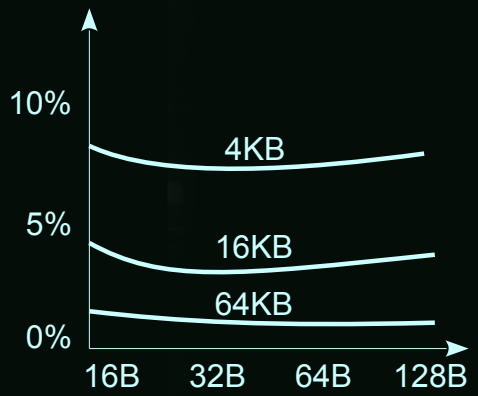
Pojęcia:
- struktura cache
- poziomy
- wydajność

Pojemność cache jest znacznie mniejsza niż pamięci RAM. Z tego powodu podczas pracy z pamięcią cache mogą wystąpić dwa zdarzenia:
- cache hit – dane znajduje się w cache,
- cache miss – spowodowane brakiem danych w cache.

Zwiększenie wydajności przez cache:



średni czas dostępu



współczynnik trafień względem rozmiaru wiersza cache

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Pojemność cache jest znacznie mniejsza niż pamięci RAM. Z tego powodu podczas pracy z pamięcią cache mogą wystąpić dwa zdarzenia:
- cache hit – dane znajduje się w cache,
- cache miss – spowodowane brakiem danych w cache.

Zwiększenie wydajności przez cache dla :
- procesora bez cache:
  - współczynnik CPI = 1,
  - dostęp do RAM wymaga 10 cykli,
  - 30% instrukcji wymaga dostępu do pamięci.
  - dla 100 instrukcji szybkość pracy = $130 * 10 = 1300$
- procesor z cache:
  - cache hit instrukcji = 0.95, cache hit danych = 0.90,
  - koszt cache miss = 15 cykli,
  - dla 100 instrukcji szybkość pracy wynosi:

$$100*(0.95*1+0.05*15)+30*(0.9*1+0.1*15) = 242$$

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Wydajność CPU zwiększa się wraz ze wzrostem liczby poziomów cache. Z reguły kolejne poziomy cache są mniej specjalizowane, wolniejsze i o większej pojemności.

Przykład:
- częstotliwość pracy CPU = 5GHz, cykl = 0.2ns, CPI = 1,
- czas dostępu do pamięci równy 100ns = 500 cykli,
- uwzględniając czas dostępu do pamięci:
  CPI = 1 + |czas oczekiwania na dane z pamięci| = 501

Dla CPU z pamięcią L1(współczynnik trafień = 0.95):
- CPI = 1 + |cykle oczekiwania|
  = 1 + 0.95 ∗ 0 + 0.05 ∗ 500 = 26
- zwiększenie szybkości: 501/26 = 19.3

Dla CPU z L1 i L2(wsp. trafień = 0.90, cykl = 25ns):
- CPI = 1 + |cykle oczekiwania|
  = 1 + 0.95 ∗ 0 + 0.05 ∗ (0.9 ∗ 25 + 0.1 ∗ 525) = 4.75
- zwiększenie szybkości: 501/4.75 = 105.5

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć operacyjna komputera pracuje znacznie wolniej niż procesor. Z tego powodu stosuje się specjalne mechanizmy niwelujące tą niedogodność.

Pojęcia:
- struktura cache
- poziomy
- wydajność

Niezależnie od liczby poziomów pamięci cache, wzrost wydajności wynikający z jej zastosowania jest możliwy jedynie w sytuacji, gdy zostaną zachowane współczynniki trafień. Aby było to możliwe pamięć cache posiada własny kontroler, realizujący następujące zadania:
- pobieranie danych z pamięci RAM,
- zapisywanie danych do RAM,
- pobieranie danych z wyprzedzeniem – realizacja strategii wymiany danych w cache.

Istnieje wiele algorytmów wymiany danych w cache, oparte na częstości wykorzystania danych, asocjacjach, podziale na kanały itp.

Dobrze dobrany algorytm cache powoduje, że wpływ czasu odpowiedzi pamięci na żądanie CPU nie wpływa na jego współczynnik CPI.

Moduł pamięci cache składa się z:
- kontrolera i magistral,
- wierszy pamięci:

tag	blok danych (cache line)	flagi

Przy dostępie do cache adres efektywny RAM dzieli się na części:

tag	indeks	offset bloku

Pamięć cache organizuje się w bloki logiczne, nazywane poziomami. Szybkość działania jak i wielkość cache na każdym poziomie jest różna, odpowiednio:
- poziom 1 – niewielki rozmiar, pracująca z szybkością CPU, zazwyczaj podzielona na dane i rozkazy,
- poziom 2 – pamięć wolniejsza od L1,
- poziom 3 – pamięć wolniejsza od L2, wspólna dla wszystkich rdzeni CPU, zazwyczaj o pojemności kilkukrotnie większej niż L2.

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć cache przechowuje dane w specjalnie przystosowanych strukturach, których konstrukcja wynika ze specyfiki jej pracy.

Pojęcia:
- moduł cache
- mapowania cache

Pamięć cache procesora

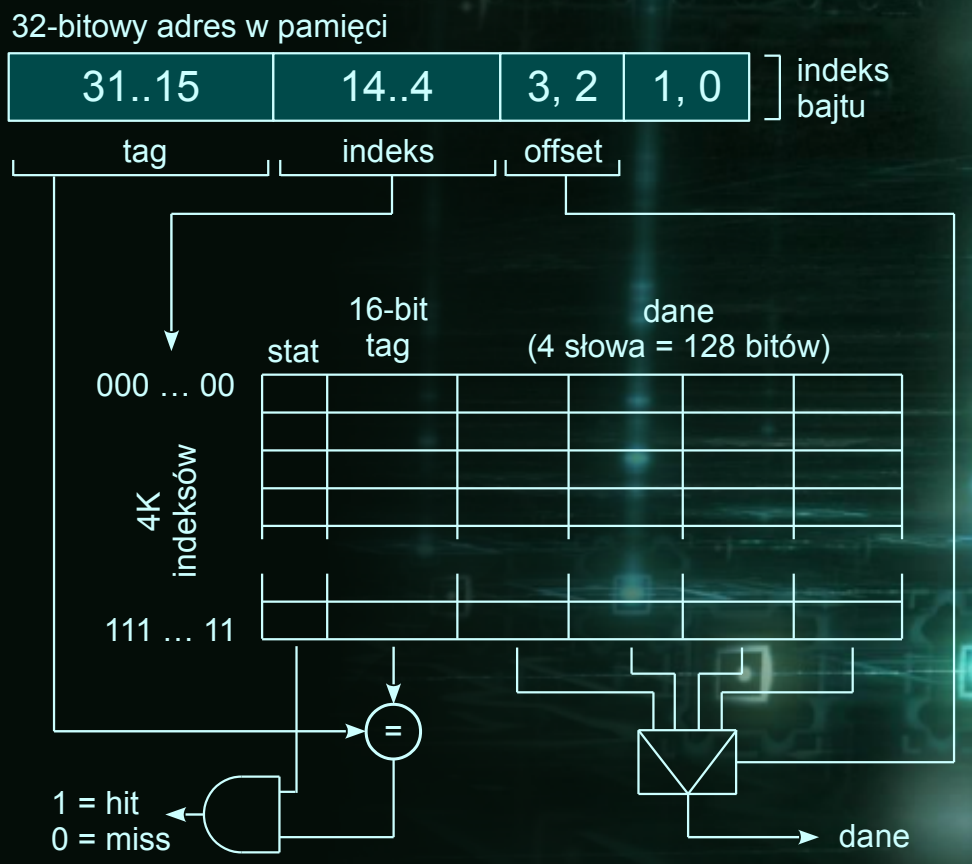
mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Pamięć cache przechowuje dane w specjalnie przystosowanych strukturach, których konstrukcja wynika ze specyfiki jej pracy.

Pojęcia:
- moduł cache
- mapowania cache

Konstrukcja modułu cache:



- cache 16K słów,
- rozmiar bloku = 4 słowa

M.Wiśniewski

Pamięć cache procesora

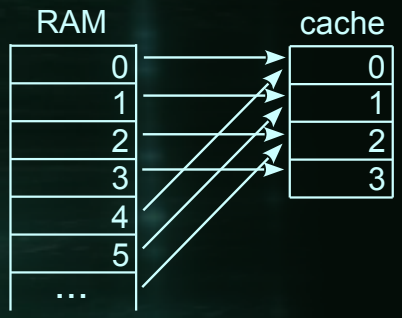
mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

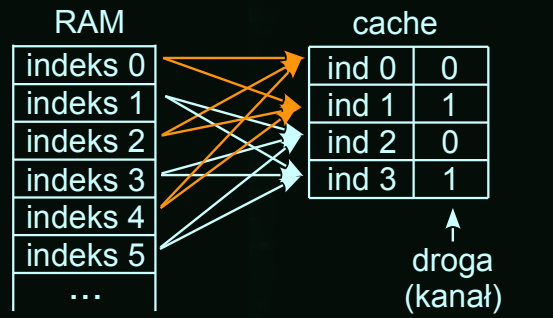
Pamięć cache przechowuje dane w specjalnie przystosowanych strukturach, których konstrukcja wynika ze specyfiki jej pracy.

Pojęcia:
- moduł cache
- mapowania cache

Kontroler cache kopiuje komórki pamięci operacyjnej do pamięci podręcznej stosując pewne strategie:

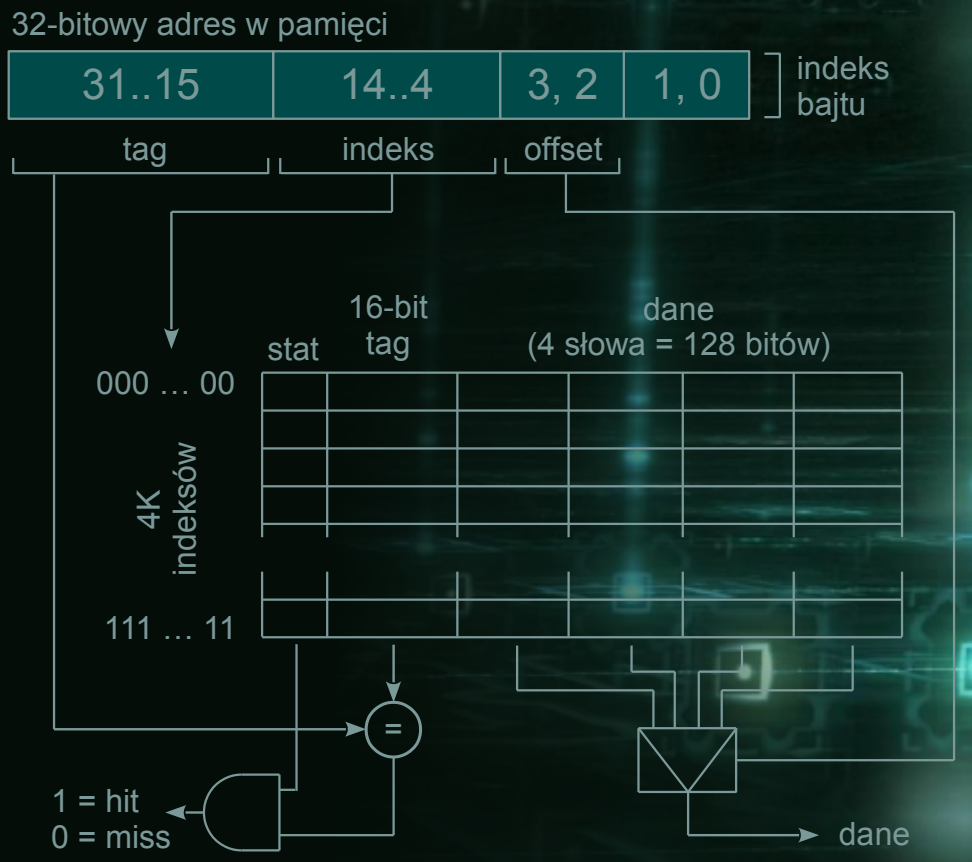


mapowanie bezpośrednie



mapowanie dwudrożne (każda komórka RAM może znajdować się w dwóch lokacjach w cache)

Konstrukcja modułu cache:



- cache 16K słów,
- rozmiar bloku = 4 słowa

M.Wiśniewski

Pamięć cache procesora

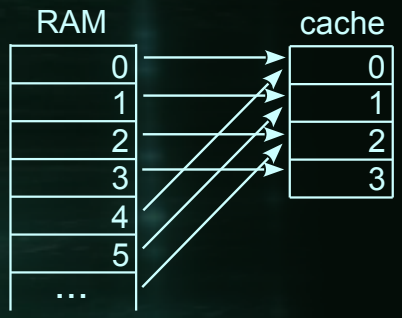
mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

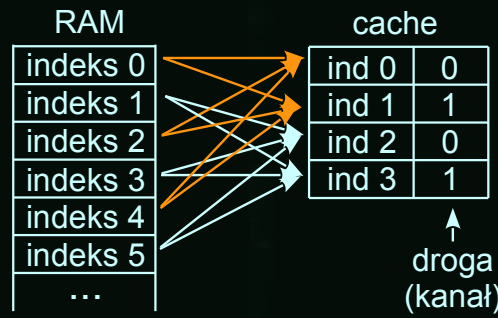
Pamięć cache przechowuje dane w specjalnie przystosowanych strukturach, których konstrukcja wynika ze specyfiki jej pracy.

Pojęcia:
- moduł cache
- mapowania cache

Kontroler cache kopiuje komórki pamięci operacyjnej do pamięci podręcznej stosując pewne strategie:



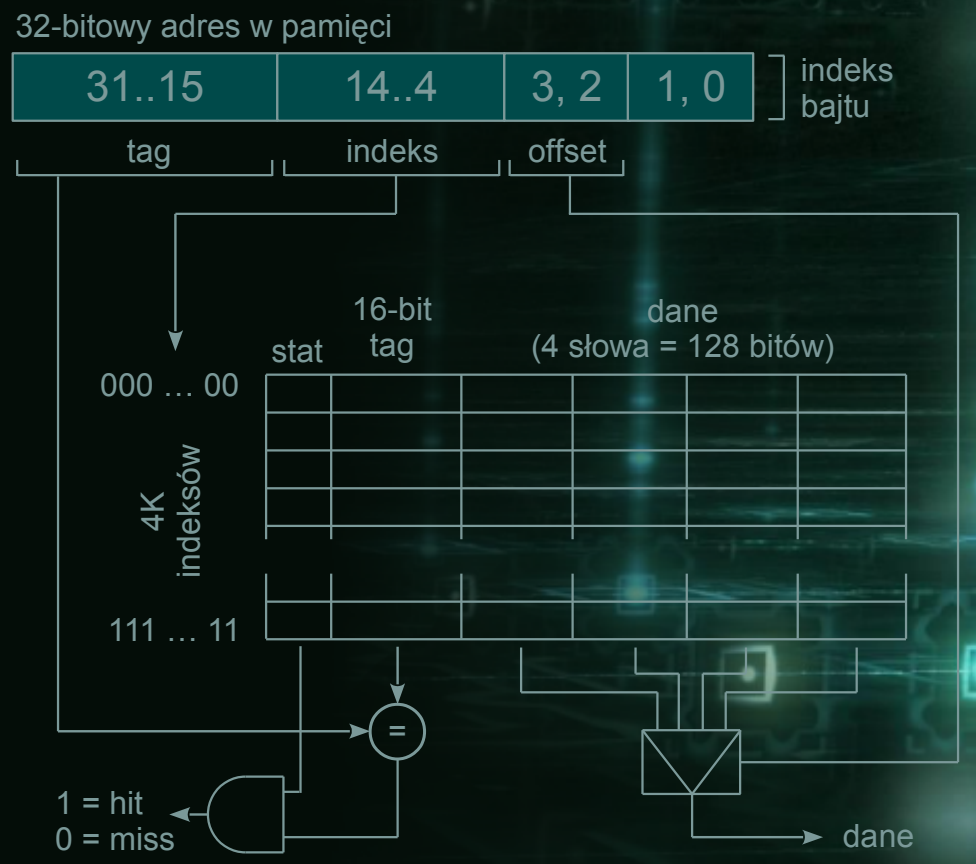
mapowanie bezpośrednie



mapowanie dwudrożne (każda komórka RAM może znajdować się w dwóch lokacjach w cache)

Rodzaje cache:
- z mapowaniem bezpośrednim (szybka, prosta),
- dwudrożna / skośnie dwudrożna asocjacyjna,
- czterodrożna asocjacyjna,
- w pełni asocjacyjna (najmniejsza liczba chybień).

Konstrukcja modułu cache:



- cache 16K słów,
- rozmiar bloku = 4 słowa

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Nowoczesne mikroprocesory współpracują z pamięcią operacyjną poprzez pamięć cache, co powoduje konieczność obsługi dodatkowych zdarzeń.

Pojęcia:
- spójność danych
- obsługa chybień
- przeplot RAM

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Nowoczesne mikroprocesory współpracują z pamięcią operacyjną poprzez pamięć cache, co powoduje konieczność obsługi dodatkowych zdarzeń.

Pojęcia:
- spójność danych
- obsługa chybień
- przeplot RAM

Pamięć operacyjna i cache stają się niespójne, gdy dane zostaną zapisane do cache, ale nie do RAM. Istnieją dwie drogi rozwiązania tego problemu:
- zapis jednoczesny (write-through):
  - zapis do RAM następuje jednocześnie z zapisem do cache,
  - zapis do RAM jest ok. 100 razy wolniejszy.
- zapis opóźniony (write-back/behind):
  - zapis wykonuje się do pamięci cache, ustawiając bit statusu (dirty/modified),
  - zapis do RAM następuje gdy to możliwe, lub gdy CPU zażąda ponownie dostępu do tej komórki RAM.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

- wstęp
- konstrukcja
- współpraca z procesorem

Nowoczesne mikroprocesory współpracują z pamięcią operacyjną poprzez pamięć cache, co powoduje konieczność obsługi dodatkowych zdarzeń.

Pojęcia:
- spójność danych
- obsługa chybień
- przeplot RAM

Pamięć operacyjna i cache stają się niespójne, gdy dane zostaną zapisane do cache, ale nie do RAM ...

Chybienie cache to sytuacja, gdy CPU zażąda danych nie znajdujących się w cache. W takim wypadku kontroler CPU wykona:
- wstrzymanie potoku (jeśli występuje),
- zablokowanie zawartości rejestrów,
- włączenie kontrolera cache, który:
  1. gdy cache jest pełny:
     - wybierze najdawniej wykorzystywany wiersz cache i oznaczy go do zapisu,
     - jeśli wybrany blok nie jest spójny, to wykona jego zapis do RAM,
  2. skopiuje do cache blok danych zawierający żądane dane
- wznowienie instrukcji.

Chybienie może nastąpić w trakcie pobierania rozkazu, wtedy czynności są podobne do powyższych.

M.Wiśniewski

Pamięć cache procesora

mechanizm optymalizacji dostępu do pamięci operacyjnej

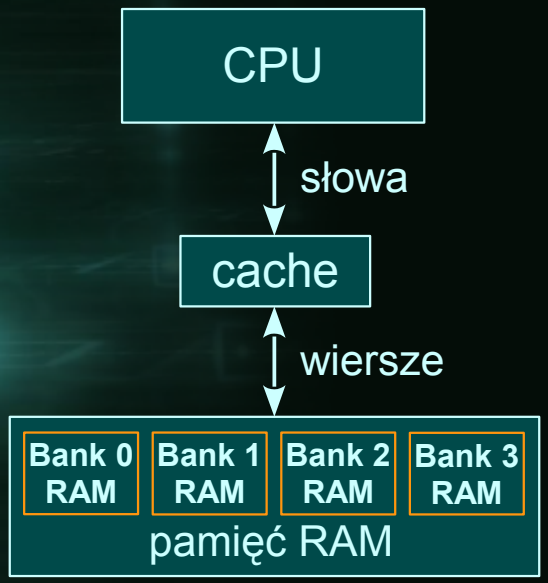
- wstęp
- konstrukcja
- współpraca z procesorem

Nowoczesne mikroprocesory współpracują z pamięcią operacyjną poprzez pamięć cache, co powoduje konieczność obsługi dodatkowych zdarzeń.

Pojęcia:
- spójność danych
- obsługa chybień
- przeplot RAM

Pamięć operacyjna i cache stają się niespójne, gdy dane zostaną zapisane do cache, ale nie do RAM ...

W systemach komputerowych z pamięcią podręczną stosuje się podział pamięci RAM na banki, uzyskując:
- zmniejszenie opóźnienia z powodu chybienia,
- odczyt słów wiersza za pomocą jednej instrukcji RAM.



Przykład:
- rozmiar wiersza cache równy 4 słowa,
- organizacja RAM w 4 banki,
- odczyt RAM ~ 15 cykli,
- opóźnienie chybienia:
  - 1 cykl na adres RAM,
  - 15 cykli na odczyt,
  - 4 cykle na zapis cache.

  sumarycznie = 20 cykli,
- opóźnienie chybienia bez przeplotu = 65 cykli.

Chybienie cache to sytuacja, gdy CPU zażąda danych nie znajdujących się w cache. W takim wypadku kontroler CPU wykona:
- wstrzymanie potoku (jeśli występuje),
- zablokowanie zawartości rejestrów,
- włączenie kontrolera cache, który:
  1. gdy cache jest pełny:
     - wybierze najdawniej wykorzystywany wiersz cache i oznaczy go do zapisu,
     - jeśli wybrany blok nie jest spójny, to wykona jego zapis do RAM,
  2. skopiuje do cache blok danych zawierający żądane dane
- wznowienie instrukcji.

Chybienie może nastąpić w trakcie pobierania rozkazu, wtedy czynności są podobne do powyższych.

Koniec wykładu