Przykłady praktycznych rozwiązań architektur systemów obliczeniowych AMD, Intel, NUMA, SMP

Wykład przetwarzanie równoległe cz.3

NUMA versus SMP systemy wieloprocesorowe

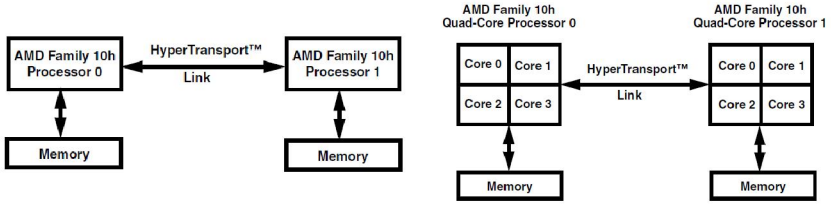
NUMA- każdy procesor jest bliżej pewnych części pamięci niż innych (nie dotyczy pamięci podręcznej). W systemach NUMA system operacyjny próbuje szeregować wątki na te procesory, które są bliżej pamięci, którą wykorzystują .

SMP – procesory i rdzenie podłączone są do pojedynczej współdzielonej pamięci głównej, przepustowość magistrali łączącej procesory z pamięcią stanowi silne ograniczenie dla efektywności systemu przy zwiększającej się liczbie procesorów. W SMP wątek może być przydzielony do dowolnego procesora. Program szeregujący dysponuje zbiorem procesorów, które są wykorzystywane do współbieżnego przetwarzania wątków. Szeregowanie jest uzależnione od priorytetów wątków. Dodatkowo optymalizacja wykorzystania pamięci podręcznej w procesach wielowątkowych może być realizowana za pomocą funkcji ustalających: **Thread Affinity – powinowactwo wątków i Thread Ideal Processor – najlepszy procesor dla wątku.**

Systemy NUMA

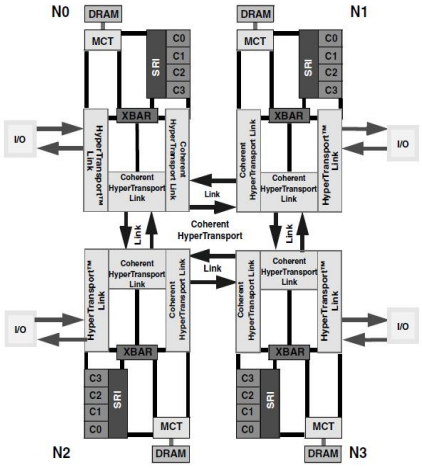
- Systemy NUMA pozwalają na wzrost prędkości przetwarzania bez konieczności wzrostu obciążenia magistrali komputera. Procesory uzyskują szybciej dostęp do bliskiej im pamięci, a do dalszej czas dostępu może być dłuższy. Jednostki przetwarzające tworzą podsystemy zwane węzłami. Każdy węzeł posiada własne procesory i pamięć, węzły są połączone z innymi węzłami za pomocą magistral zapewniających spójność pamięci podręcznych.
- System operacyjny próbuje zwiększyć wydajność przez szeregowanie wątków do procesorów w węzłach, gdzie znajduje się wykorzystywana przez nie pamięć, próbuje realizować żądania przydziału pamięci w ramach tego samego węzła, lecz również innych jeśli będzie to konieczne. Środowisko uruchomieniowe dostarcza funkcji pozwalających na określenie topologii systemu dostępnego dla aplikacji. Efektywność aplikacji może zostać podwyższona za pomocą funkcji dla NUMA pozwalających na optymalizację szeregowania i wykorzystania pamięci.

Systemy NUMA przykład I



System dwuprocesorowy z procesorami rodziny AMD 10h

SYSTEMY NUMA PRZYKŁAD I



Cztery rdzenie komunikują się przez: **system request interface** (SRQ) połączony z **non-blocking crossbar** (XBar). **XBar** jest połączony ze sterownikiem pamięci (**MCT**) i różnymi łączami typu HyperTransport . Sterownik pamięci MCT jest połączony z pamięcią lokalnego węzła. MCT, SRQ i XBar każdego węzła posiadają wewnętrzne bufory wykorzystywane do kolejkowania pakietów transakcji HyperTransport . SRQ, XBar i MCT tworzą Northbridge węzła.

System 4 procesorów 4 rdzeniowych rodziny 10h AMD

SYSTEMY DLA LABORATORIUM

- Komputery znajdujące się w ***Laboratorium Systemów Równoległych sala 2.7.6*** posiadają po jednym procesorze AMD typu PHENOM II X4 945
- System składa się z 4 procesorów logicznych – 4 rdzeni w ramach jednego procesora. System SMP.

PROCESOR PHENOM

Zgodność 32 bitowa X86 IA

wspomaganie SSE, SSE2, SSE3, SSE4a, ABM, MMX™, 3DNow!™

Technologia AMD64

rozszerzenia AMD64 technology instruction-set

Adresowanie 48-bitowe

16 rejestrów 64-bit dla integer

16 rejestrów 128-bit SSE/SSE2/SSE3/SSE4a

Architektura wielordzeniowa

opcje: Triple-core, **quad-core** lub six-core

AMD Balanced Smart Cache

oddzielne pp L1 i L2 dla każdego rdzenia

współdzielona L3

Struktura procesora

superskalarny 3 drożny (dekodowanie, wykonanie integer i FP, generacja adresu)

Struktura pp

64-Kbyte 2 drożna dzielona asocjacyjna pp danych L1

dwa dostępy 64-bit na cykl, 3 cyklowe opóźnienie

64-Kbyte 2 drożna dzielona asocjacyjna pp kodu L1

32 bajtowe pobrania

512-Kbyte 16 drożna dzielona asocjacyjna pp L2

Zarządzanie pamięcią na zasadzie wyłączności przechowywania danych L1 i L2

6-Mbyte Maximum, maksymalnie 64 drożna dzielona asocjacyjna pp L3 współdzielona

Technologia 45 nm

Złącze HyperTransport™

Procesor zintegrowany ze sterownikiem pamięci

PROCESOR PHENOM – PP KODU L1

- Układ dynamicznego wykonania instrukcji posiada 64KB pp kodu L1
- Dane w przypadku braku trafienia są pobierane do pp kodu L1 z L2, z L3 lub z pamięci systemowej w ilości 64 bajtów (pobranie) oraz kolejne 64 bajty (wstępne pobranie), po pobraniu realizowane jest wstępne dekodowanie instrukcji dla określenia granic między instrukcjami (zmiennej długości), usuwanie linii z pp jest realizowane zgodnie z algorytmem LRU (ang. least recently used**)**

PROCESOR PHENOM – PP DANYCH L1

- 64 kB dwu-sekcyjna, dwa porty 128 bitowe
- Strategia zapisu: Write-allocate cache – zapis realizowany do pp (przeciwna strategia do nowrite alllocation)
- Writeback cache – zapis poza pp realizowany w przypadku braku miejsca lub na skutek zlecenia zapisu stanu w pamięci głównej
- Algorytm LRU dla usuwania danych i protokół zapewnienia spójności MOESI

PROCESOR PHENOM – PP L2 I L3

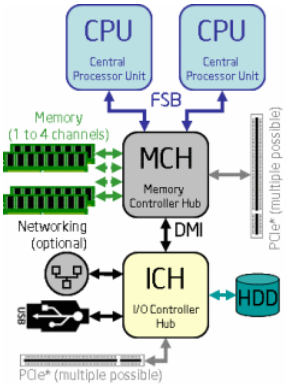
- PP L2 - victim i copy-back cache – zapisuje dane usunięte z pp L1, dane w pp są w trybie wyłącznym w L1 lub w L2
- PP L3 – victim i copy-back cache dla pp L2, głównie non-inclusive cache w przypadku, gdy dane żądane są przez jeden z rdzeni i jest mało prawdopodobne, że będą potrzebne innym, lecz możliwe powielenie.

Przykład 2
Struktura systemu z Intel Xeon SMP

System klastra składa się z węzłów 2 procesorowych.
Każdy węzeł to:

- system 2 procesorowy ze współdzieloną pamięcią,
- 2 magistrale FSB (front side bus) łączące procesory z pamięcią,
- koncentratory (Intel 5000P) sterujące pamięcią i I/O
- Każdy CPU to 4 rdzenie i 2 pp L2

chipset implementuje protokół zapewnienia spójności bazujący na podglądaniu - sprawdzaniu czy dostęp do informacji jednego procesora dotyczy informacji zapisanych w pamięci podręcznej innego procesora, ilość informacji przekazywanych z jednej do drugiej magistrali FSB jest ograniczona poprzez filtr (snoop filter) zimplementowany w MCH, informacje o lokalnych dostępach do linii, które są w pp drugiego procesora i nie są w stanie INV są przesyłane do drugiej FSB.



Procesor Intel

- Intel Xeon E5345 2.33 GHz – Quad-Core Intel Xeon Processor (**Clovertown)**
  - architektura procesora: IA-32, Intel 64 - 64 bitowy
  - architektura sprzętu: Intel Core
  - technologia quad-core - 4 rdzenie
  - prędkość FSB 1333MHz
  - technologia 65 nm
- Architektura procesorów Intela 3 poziomy opisu:
  - Architektura procesora: definicja zbioru instrukcji i zakresu zgodności np. EPIC, IA-32, IXA[2]
  - Architektura sprzętu (microarchitecture) np. P5, P6, Intel NetBurst, Intel Core, Mobile, Next Generation Intel Microarchitecture (Nehalem) w ramach IA-32
  - Procesory – implementacje mikroarchitektury np. Intel Xeon Procesor serii 5300

cat /proc/cpuinfo

processor : 0
vendor_id : GenuineIntel
cpu family : 6
model : 15
model name : Intel(R) Xeon(R) CPU E5345 @ 2.33GHz
stepping : 11
cpu MHz : 2333.423
cache size : 4096 KB
physical id : 0
siblings : 4
core id : 0
cpu cores : 4

apicid : 0
initial apicid : 0
fpu : yes
fpu_exception : yes
cpuid level : 10
wp : yes
clflush size : 64
cache_alignment : 64
address sizes : 38 bits physical, 48 bits virtual

Struktura pamięci systemu

PP L1

- 8 drożna dzielona asocjacyjna
- dla każdego rdzenia w procesorze:
  32kB instrukcji i 32kB danych

PP L2 16 drożna dzielona asocjacyjna

- 8 MB I+D w procesorze,
- 4MB współdzielone dla 2 rdzeni

Linia pp – 64 bajty dla powyższych pamięci, dane w pp wyrównane do 64 bajtowej granicy

Cechy Intel Core microarchitecture
od 2006

- technologia 65 nm
- Intel Wide Dynamic Execution
- Intelligent Power Capability
- Advanced Smart Cache
- Smart Memory Access
- Advanced Digital Media Boost

Cechy:
Intel Wide Dynamic Execution

1. analiza przepływu danych
2. spekulatywne wykonanie rozkazów
3. dynamiczne wykonywanie rozkazów
4. superskalarność

Cechy:
Intel Wide Dynamic Execution

- 14 etapowy potok przetwarzania
- 3 jednostki arytmetyczno-logiczne
- 4 dekodery (do 4 instrukcji na cykl)
- Techniki: Micro i macro fusion dla wzrostu przepustowości przetwarzania
  Micro-fusion – łaczenie operacji wewnętrznych procesora w jedną (micro-op – to co przetwarza potok)
  Macro-fusion – łączenie typowych instrukcji w jedną operację wewnętrzną procesora (np. porównanie i warunkowy skok)
- Incjowanie do 6 mikrooperacji, kończenie-zatwierdzanie do 4 mikrooperacji na cykl – SUPERSKALARNOŚĆ
- Zaawansowana predykcja rozgałęzień
- Układ śledzący wskaźnik stosu dla poprawy efektywności uruchamiania i kończenia procedur i funkcji
- Większy bufor instrukcji „w trakcie przetwarzania" dla przetwarzania dynamicznego

Cechy: Intel Advanced Smart Cache

- Pamięć podręczna L2 do 4 MB z 16 drożnym odwzorowaniem sekcyjno-skojarzeniowego (zbiorowo-asocjacyjne)
  - dane z konkretnego adresu mogą być umieszczone wyłącznie w 16 różnych lokacjach PP - 16 bloków pamięci bezpośrednio-adresowalnej.
- Pamięć L2 współdzielona między dwa rdzenie – np. Quad-Core Intel Xeon 5300 każde dwa rdzenie współdzielą 4 MB L2
- 256 bitowa ścieżka danych między L2 a L1

Cechy: Intel Smart Memory Access

Cel techniki: ukrywanie opóźnienia dostępu do pamięci

Realizacja :

1) Wcześniejsze pobranie możliwe przez wykorzystanie faktycznego braku zależności danych (ang. memory disambiguation – usunięcie niejednoznaczności istnienia konfliktu)
2) Zaawansowane układy wstępnego pobrania danych:
   - dla zapewnienia dostępności w PP danych wymaganych przez poszczególne rdzenie procesora
   - wykrywają wiele strumieni dostępu oraz wzorce dostępów z przesunięciem lokacji
   - liczba układów: 2(rdzenie) *(2xL1DC+L1IC)+2xL2C

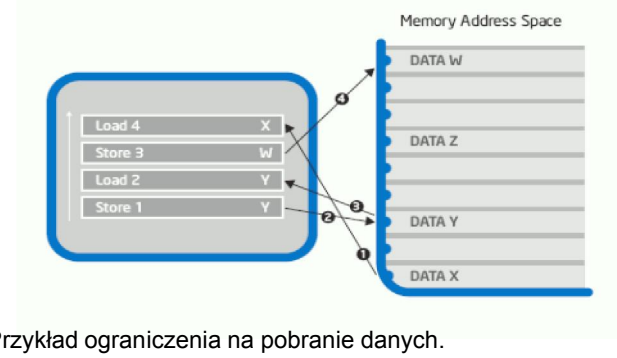
Zaawansowane układy wyprzedzającego pobrania kodu i danych

- Układ wyprzedzającego pobrania dla pamięci podręcznej danych L1 jest oparty o licznik instrukcji
  - tablica historii pobrań (adresy instrukcji i danych)
  - monitorowanie ilości pobrań i wykorzystania zasobów
  - możliwość określenia parametrów sterujących działaniem
- Działanie: wykrywanie kolejnych odwołań realizowanych instrukcją spod danego adresu do danych na adresach odległych od siebie o określony odstęp - przewidywanie adresu wymaganych danych
- Pamięć L1- lokalna rdzenia:
  - dwa układy wp dla PP danych
  - układ wp dla PP instrukcji
- Pamięć L2 – współdzielona
  - dla każdego rdzenia jeden układ wp

Memory disambiguation

Podejście do problemu niejednoznaczności istnienia konfliktu w kolejności dostępu do pamięci

- Ignorowanie warunku blokowania odczytu ze względu na nieznany adres wcześniejszego zapisu.
- Stosuje się algorytm analizy kodu dla określenia: czy pobranie danych może być zrealizowane z wyprzedzeniem w stosunku do zapisu danych?
- Jeśli pobranie jest prawdopodobnie możliwe dane są pobierane spekulatywnie dla instrukcji, które są bliskie wykonania. Predykcja opłacalności pobrania bazuje na liczniku zliczającym pomyślnie zakończone rozkazy spod tego adresu wykonane z wyprzedzającym pobraniem.



Przykład ograniczenia na pobranie danych.

Numery w kółkach określają chronologiczną kolejność realizacji instrukcji a strzałka po lewej stronie określa kolejność programu. Load 2 nie może być przesunięty w czasie przed Store 1 aby wartość Y była poprawna. Jednostka analizująca konflikty dostępu do pamięci może wykryć, że Load 4 jest niezależny od innych instrukcji i może być zrealizowany przed zapisami Store 1 i Store 3. Zrealizowanie Load 4 kilka cykli wcześniej powoduje, że procesor posiada Data X i może realizować wymagające ich instrukcje. Pozwala to na zmniejszenie opóźnienia dostępu do pamięci i uzyskanie wyższego poziomu równoległości na poziomie instrukcji.

Rysunek z: Inside Intel® Core™ Microarchitecture and Smart Memory Access

Intel Advanced Digital Media Boost

- wzrost efektywności przetwarzania instrukcji SSE (Streaming SIMD Extension)
  - instrukcja SIMD na 128 bit w czasie 1 cyklu
  - do 8 operacji (w zależności od rozmiaru zmiennej) zmiennoprzecinkowych na cykl

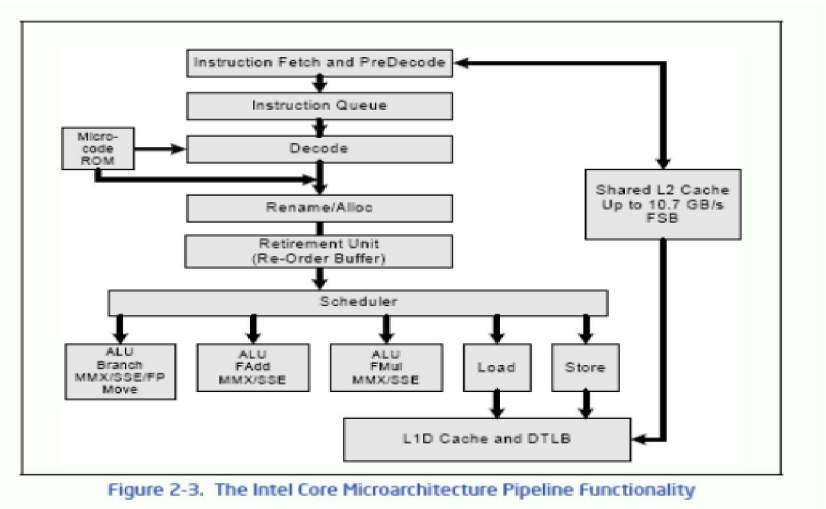


Figure 2-3. The Intel Core Microarchitecture Pipeline Functionality

Rysunek z: Intel® 64 and IA-32 Architectures Software Developer's Manual

Przykład 3: INTEL NUMA –
Nehalem Intel Microarchitecture

- Procesor Xeon serii 3500 i 5500 (technologia 45 nm)
- Intel® Hyper-Threading Technology – współdzielenie przez wiele wątków zasobów jednego rdzenia procesora – efekt: jednoczesna realizacja 2 wątków na rdzeń
- Intel® QuickPath Technology – nowa skalowalna, architektura współdzielonej pamięci, kontroler pamięci zintegrowany z procesorem, połączenie procesorów i innych komponentów systemem łączy nowego typu z prędkością do 32 GB/s na łącze dwukierunkowe (4 łacza)
- Możliwość przetwarzania do 4 instrukcji na cykl,
  - Wzrost okna instrukcji szeregowania dynamicznego – efektem większy potencjalny poziom równoległości
  - Szybsze mechanizmy synchronizacji wątków
  - Szybsza obsługa złej predykcji rozgałęzień kodu
- Dynamiczne zarządzanie rdzeniami, wątkami, pp, łączami i mocą
  - Dynamiczne zarządzanie rdzeniami – możliwość zwiększenia częstotliwości zegara wszystkich rdzeni w przypadku potrzeby/ dodatkowy wzrost f dla grupy aktywnych rdzeni, możliwe różne zegary rdzeni (nie jak w Core 2)

INTEL NUMA –
Nehalem Intel Microarchitecture

- Współdzielona (przez wszystkie rdzenie procesora jak pp L2 w Intel Core) pamięć podręczna 3 poziomu do 8 MB (powiela dane zapisane w innych poziomach pp) – efekt: mniejszy ruch pamięć – rdzeń (pp L1 L2) – powielenie powoduje spadek pojemności systemu pamięci, lecz umożliwia zmniejszenie ilości dostępów do L1 i L2 (brak konieczności przeszukiwania L1 i L2 przy cache miss do L3, zmniejszenie narzutu przeszukiwania (w którym rdzeniu dane powielone?) przy cache hit – dzięki dodatkowym flagom w L3)
- Liczba rdzeni i wątków w strukturze procesora: 1 do 8 rdzeni, 1 do 16 wątków
- pp L1 (32KB IC + 32KB DC)
- pp L2 256 KB na rdzeń
- 2 poziomy struktury TLB (Translation Lookaside Buffer) – translacja adresów pamięci wirtualnych na fizyczne (192 + 512 wejść)