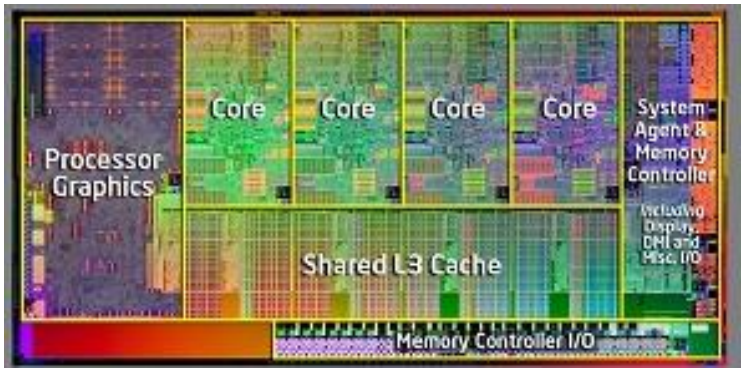
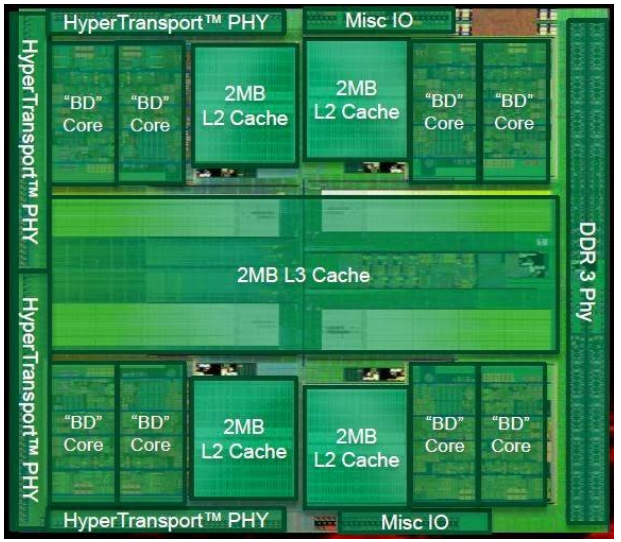




Pamięć Cache

ɱ@ʁ€Ⱪ Þud3£k0

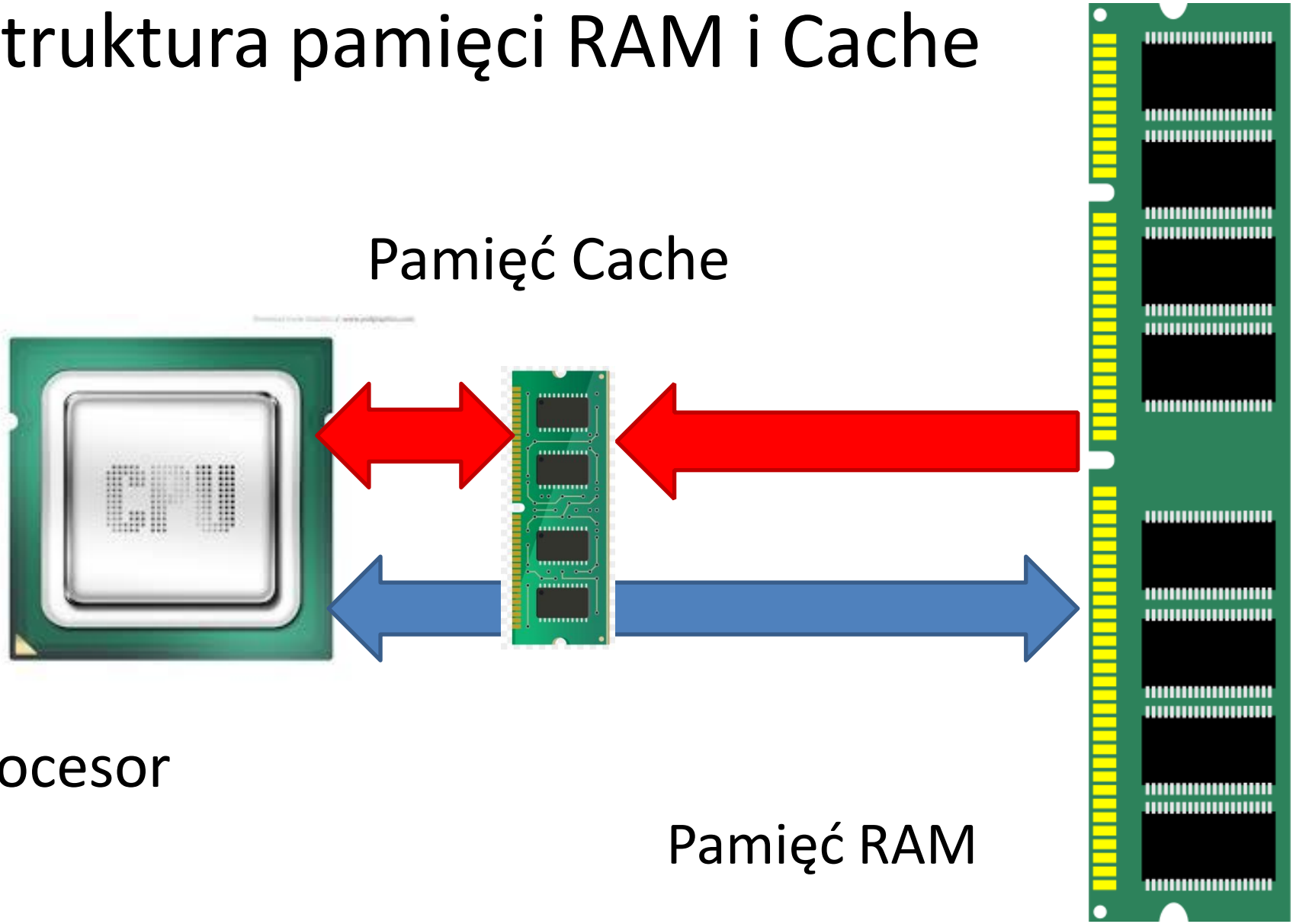
Urządzenia Techniki Komputerowej

Spis treści

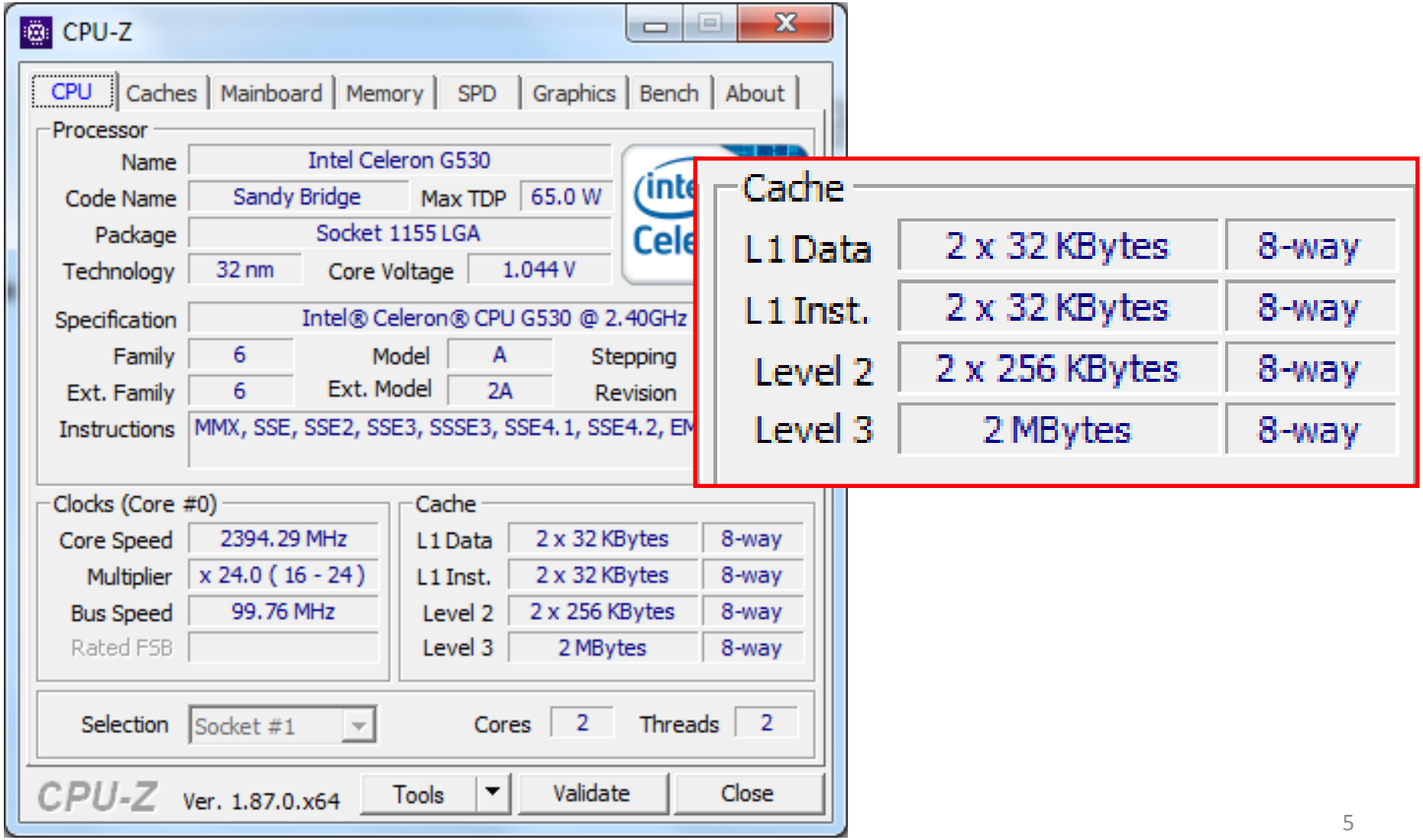
Pamięć Cache

- Szybka pamięć podręczna (ang. Cache Memory), służąca do przechowywania często używanych danych.
- Stanowi bufor pomiędzy wolną dynamiczną pamięcią operacyjną (DRAM) a procesorem.
- Typ pamięci Static RAM (SRAM)
- Wielkość jej wynosi 1 MB – 16 MB

Struktura pamięci RAM i Cache



Pamięć Cache w CPU-Z

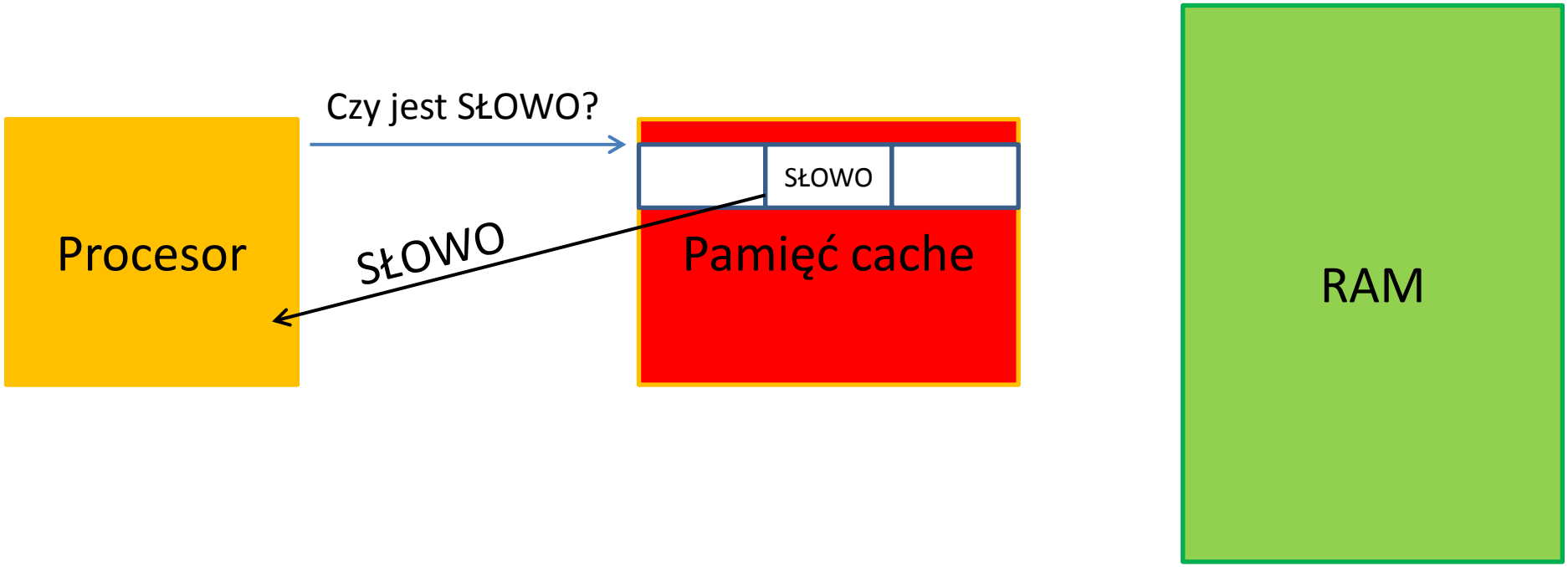


Szybkość pamięci i procesora

- Wzrost wydajności spowodowała, że procesor stał się zbyt szybki dla pamięci RAM.
- Odwoływanie się tylko do niej spowodowałoby, że większość czasu byłby niewykorzystany.
- Dodano dodatkową pamięć Cache, która jest szybsza niż RAM i stanowi magazyn często używanych danych.

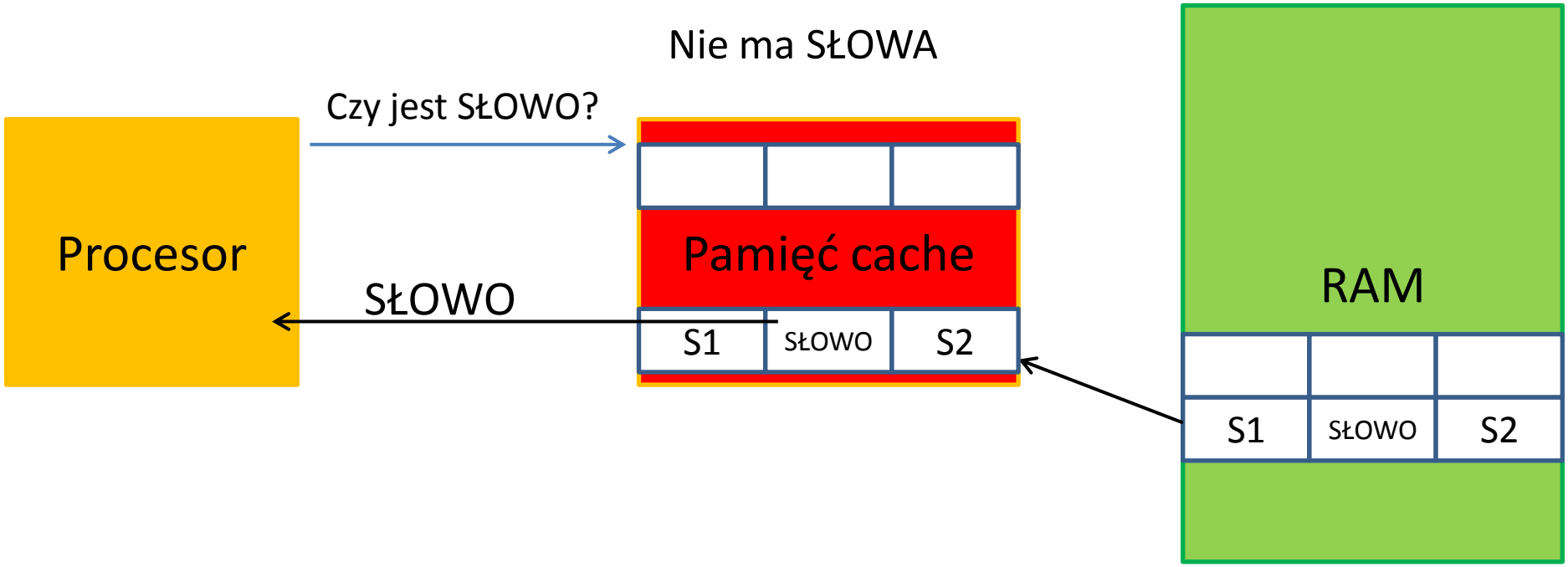
ZASADA DZIAŁANIA PAMIĘCI CACHE

Zasada działania pamięci podręcznej



- Procesor potrzebuje danego słowa.
- Szuka go najpierw w pamięci Cache.
- Gdy jest znajdzie, to ściąga je do obróbki.

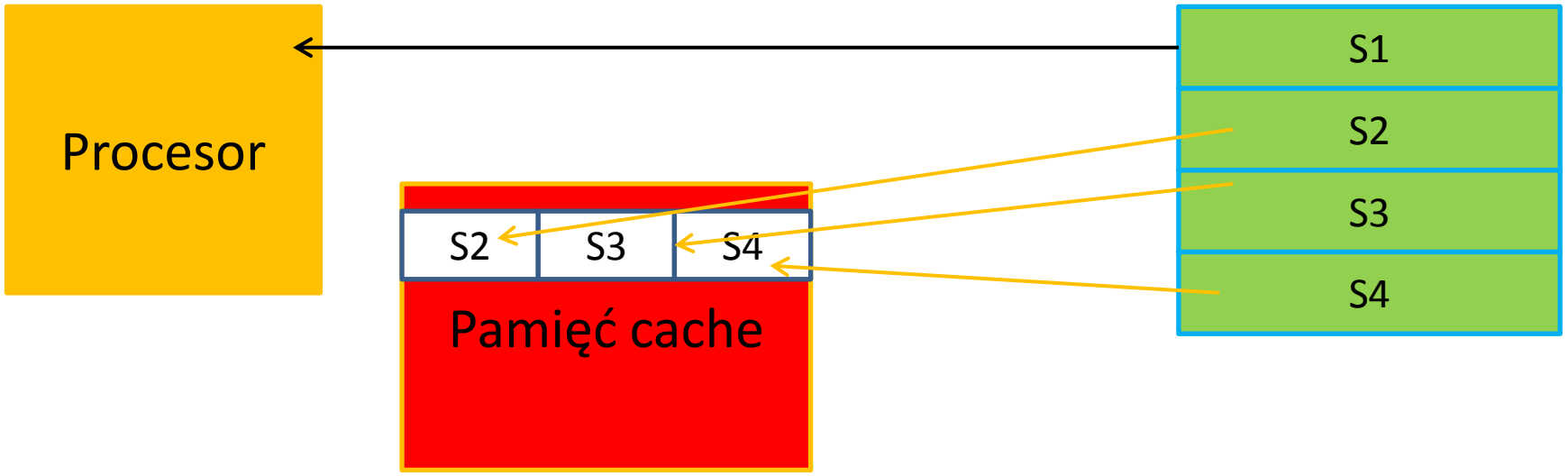
Zasada działania pamięci podręcznej



- Gdy brak danego słowa w pamięci Cache, sięga się do pamięci RAM.
- Blok pamięci RAM zawierający ustaloną liczbę słów (w tym szukane) jest wczytywany do pamięci podręcznej.
- Szukane słowo jest dostarczane do procesora.

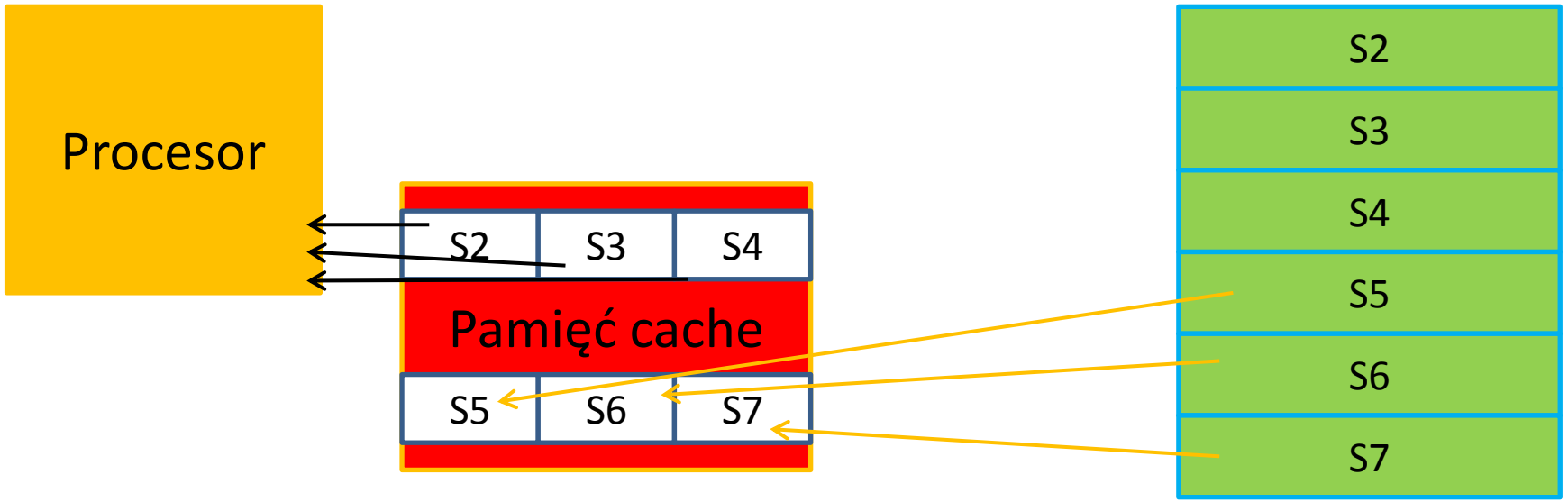
Integralność danych

- Zazwyczaj dane w pamięci RAM są połączone w logiczne grupy.
- Wykonywane są raczej sekwencyjnie.
- Pozwala to przewidzieć jakie dane i polecenia będą potrzebne za chwilę (zazwyczaj następne po aktualnie wykorzystanych) co pozwala na załadowanie ich wcześniej do pamięci cache.



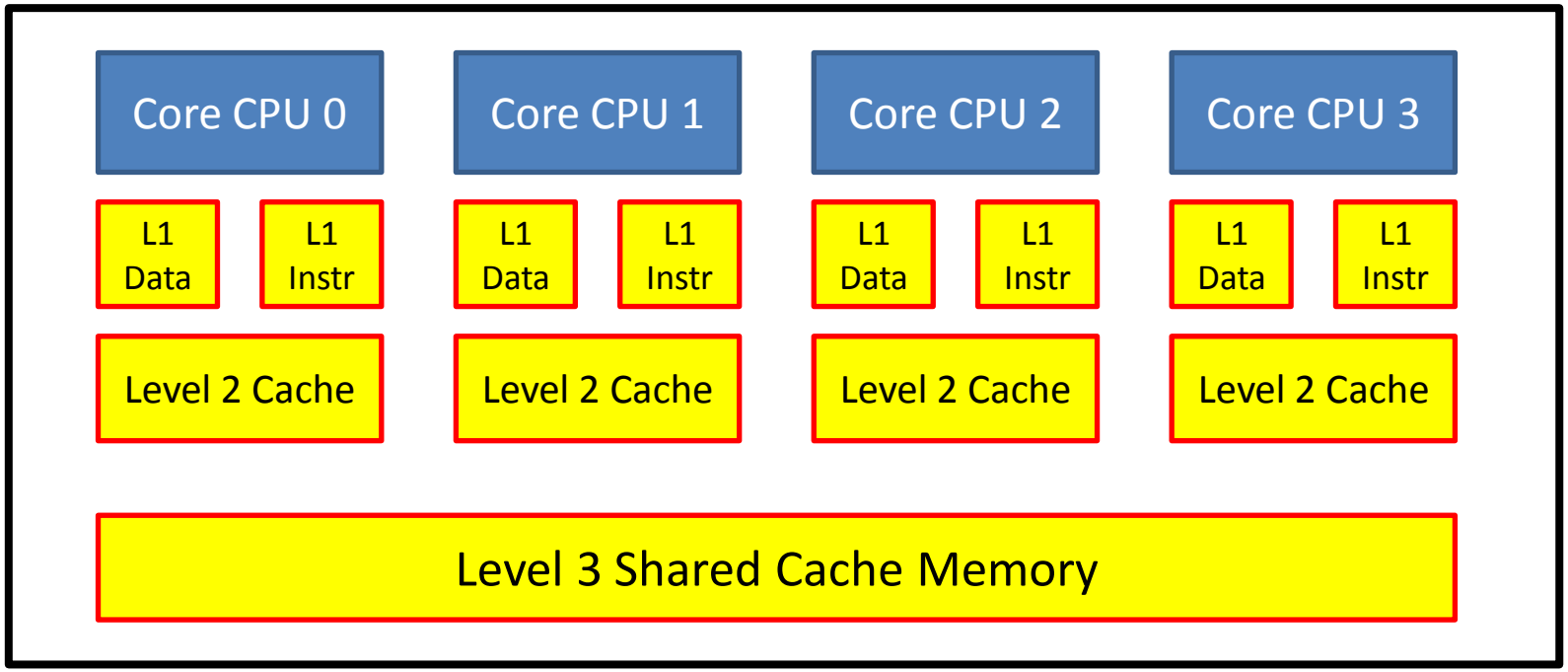
Integralność danych

- Po kolejne dane procesor sięgnie do pamięci podręcznej, co znacznie przyśpieszy obliczenia.
- Tymczasem do pamięci cache można kopiować kolejne dane z RAM.



HIERARCHIA PAMIĘCI CACHE

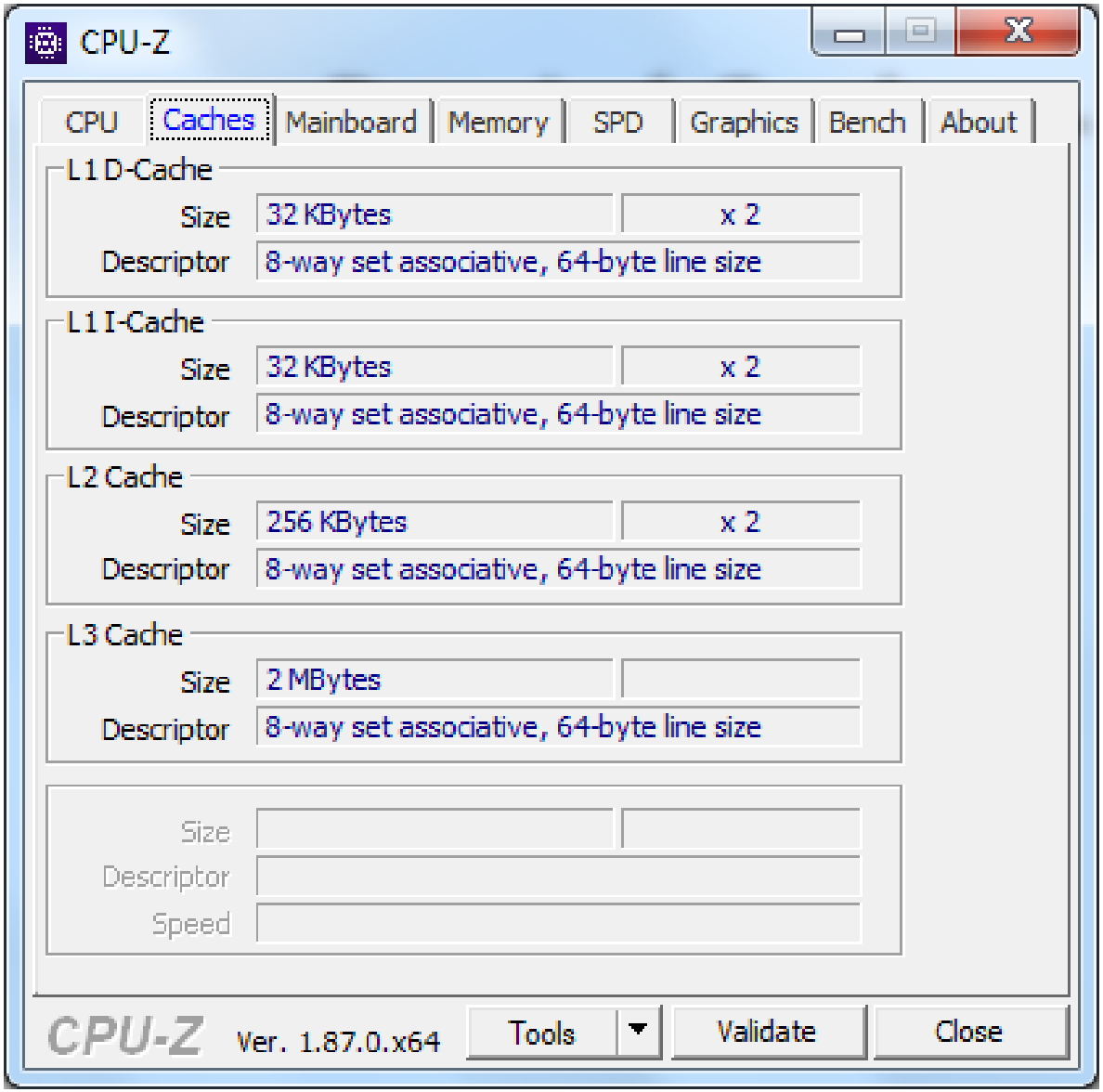
Hierarchia pamięci cache



Pamięć RAM

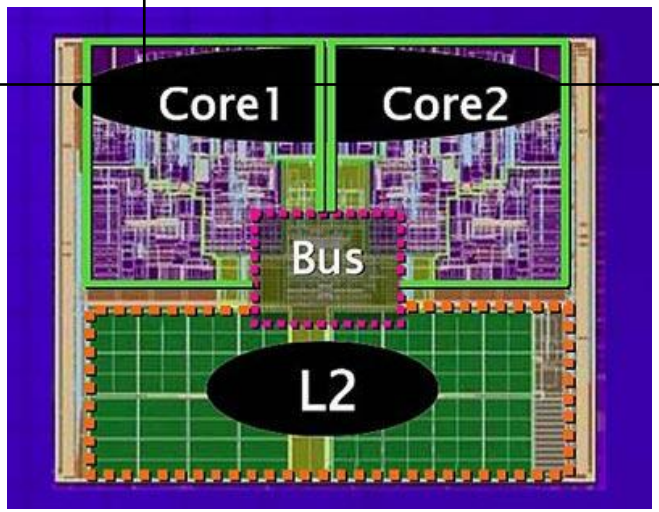
Nośniki pamięci trwałej

Pamięć Cache – CPU-Z

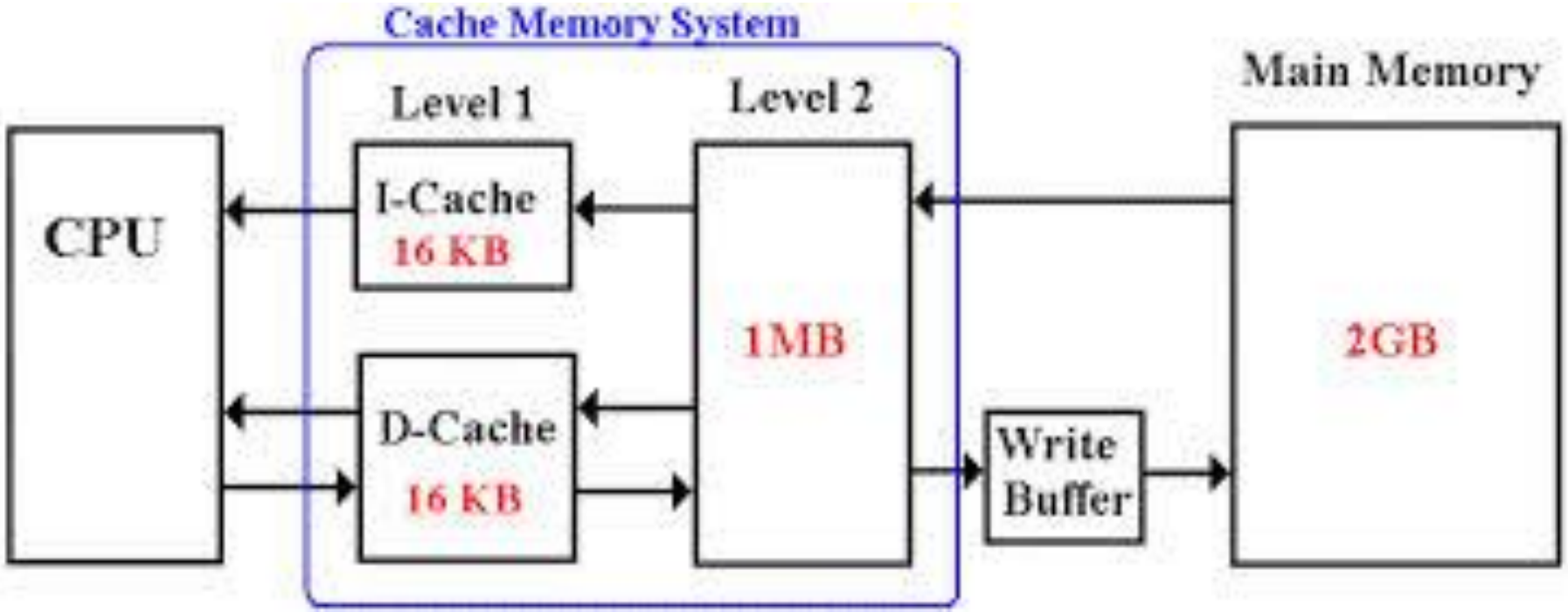


Poziomy pamięci Cache

Poziom pamięci cache	Umiejscowienie	Wielkość
L1	Zintegrowane z jądrem procesora	32kB-128kB
L2	Umieszczona na płycie głównej Może być wewnątrz wielordzeniowego procesora	1MB-8 MB
L3	Umieszczona na płycie głównej	4MB-32 MB W szybkich procesorach serwerowych jest dzielona przez wszystkie rdzenie/procesory

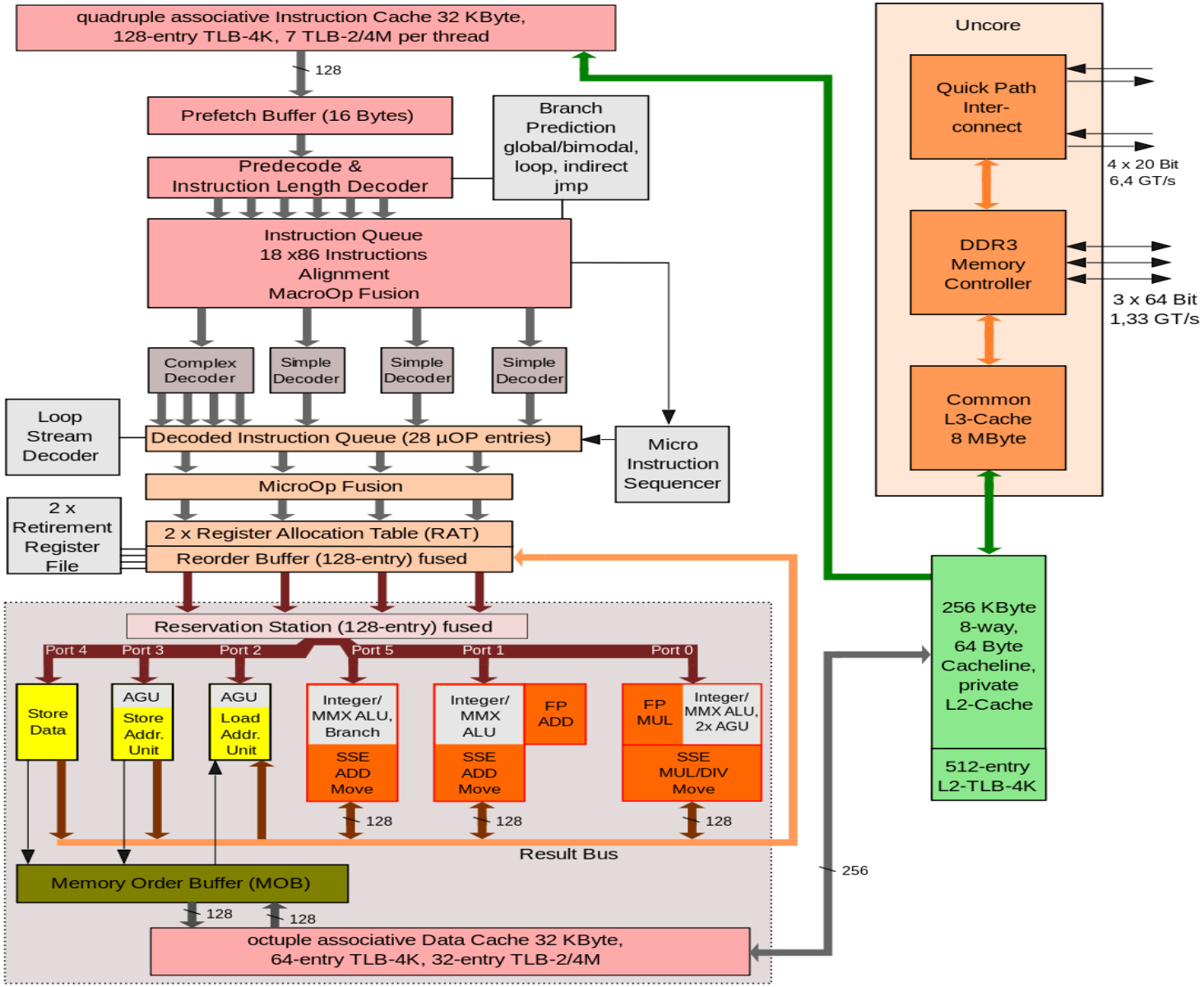


Realizacja cache



ARCHITEKTURA PAMIĘCI CACHE

Intel Nehalem microarchitecture

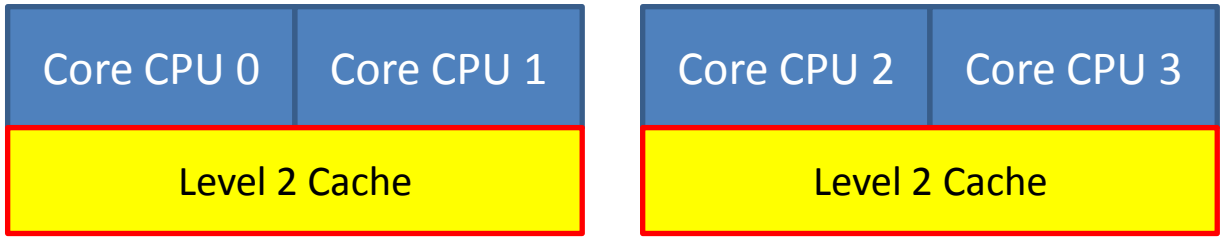


GT/s: gigatransfers per second

Architektura pamięci cache L2

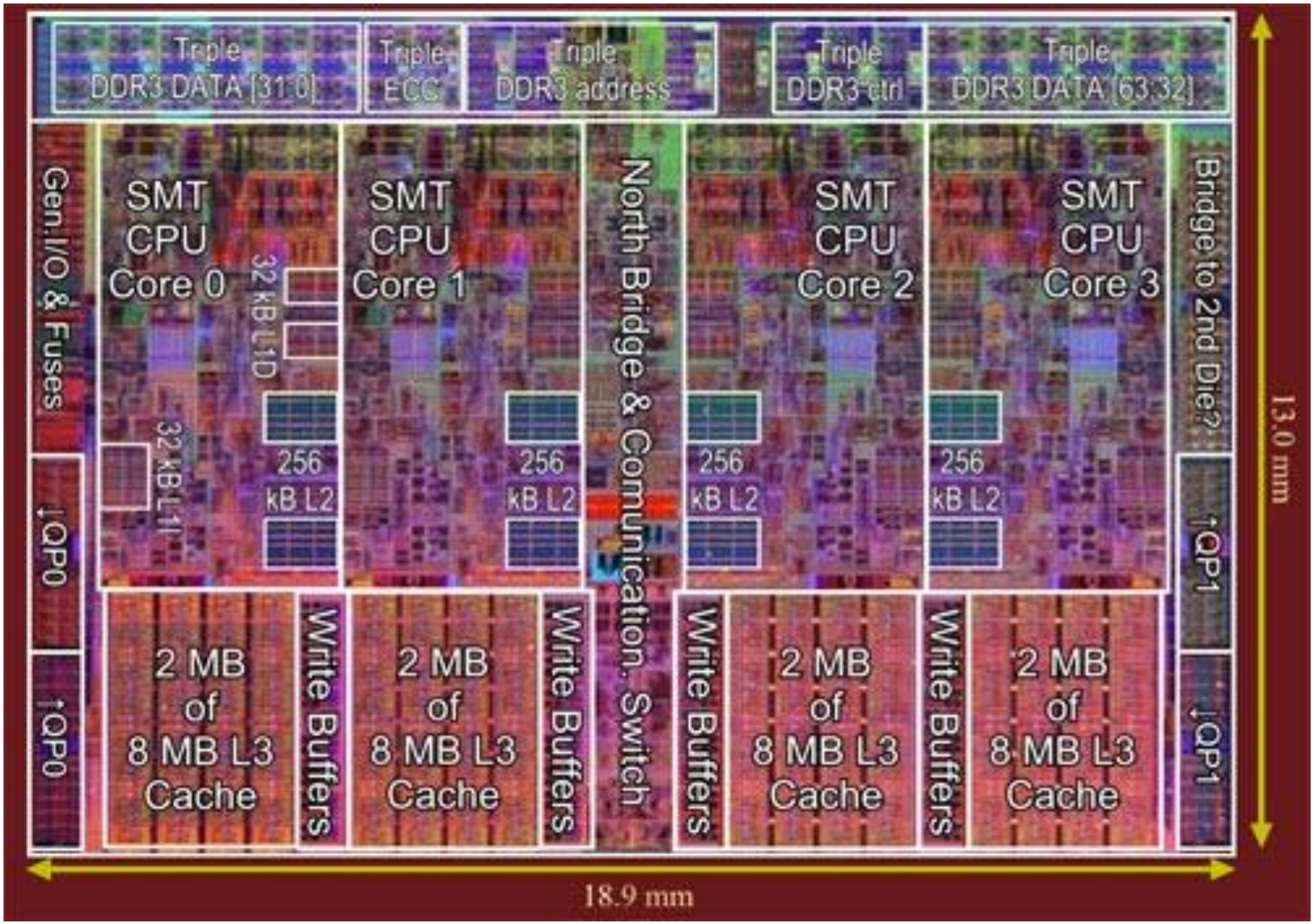
Oddzielne pamięci cache
(Separated Cache)
AMD, Pentium D

Wspólna pamięć cache
(Shared Cache)
Core Duo, Core 2 Duo

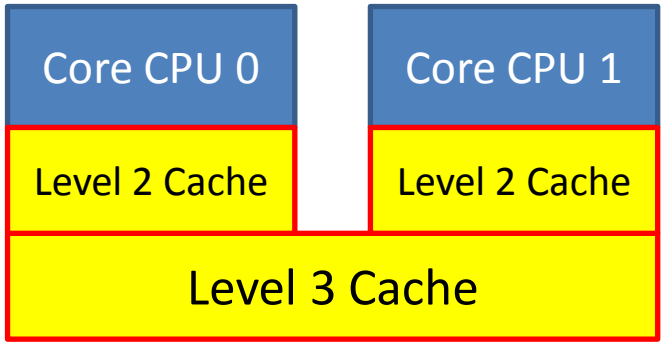


Współbieżna pamięć cache
(Current Cache)
Core 2 Quad, Core 2 Extreme QX

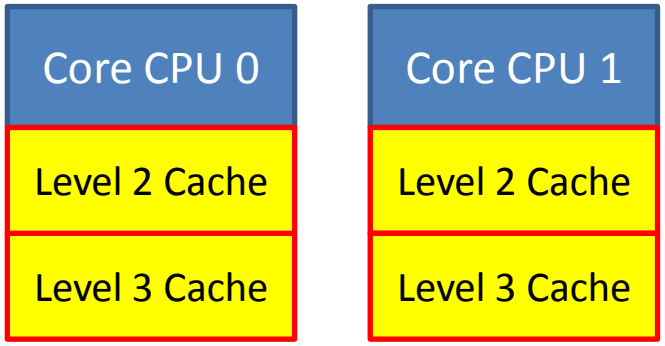
L1, L2 Cache


Triple DDR3 DATA [31:0]
Triple ECC
Triple DDR3 address
Triple DDR3 ctrl
Triple DDR3 DATA [63:32]
Gen.I/O & Fuses
SMT CPU Core 0
32 kB L1D
32 kB L1I
256 kB L2
SMT CPU Core 1
256 kB L2
North Bridge & Comunication. Switch
SMT CPU Core 2
256 kB L2
SMT CPU Core 3
256 kB L2
Bridge to 2nd Die?
↑QP0
↑QP0
2 MB of 8 MB L3 Cache
Write Buffers
2 MB of 8 MB L3 Cache
Write Buffers
Write Buffers
2 MB of 8 MB L3 Cache
Write Buffers
2 MB of 8 MB L3 Cache
↑QP1
↓QP1
13.0 mm
18.9 mm

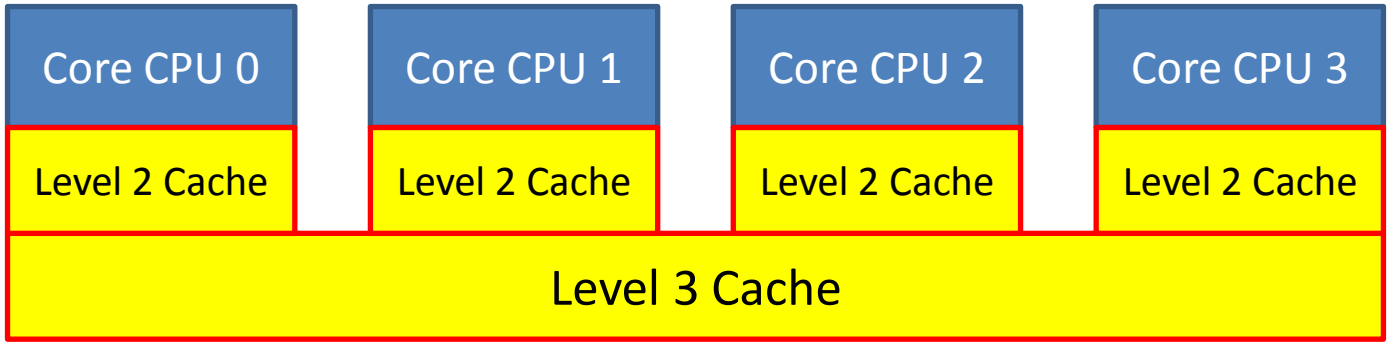
Architektura pamięci cache L3



Dual-Core

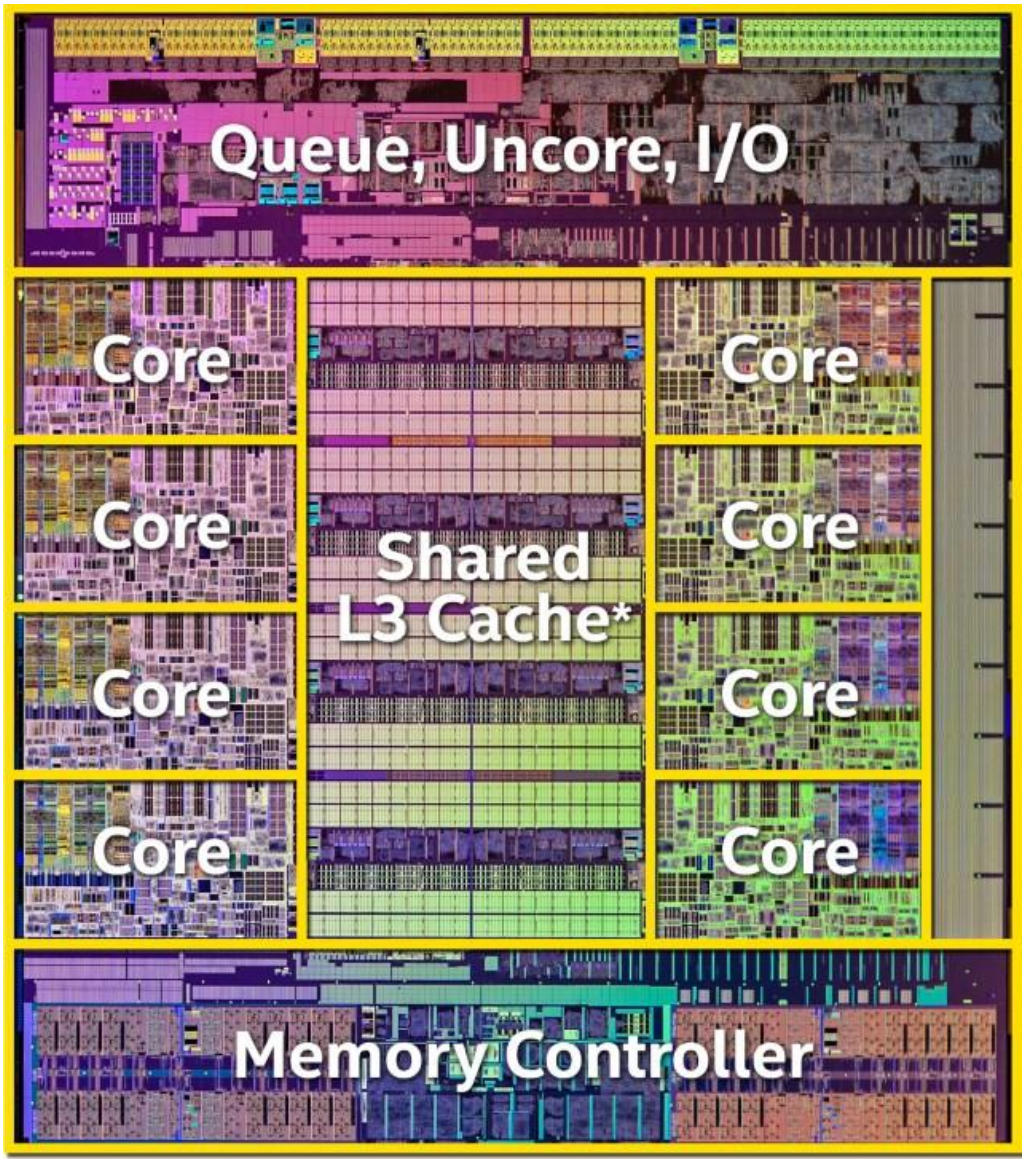


Itanium 2

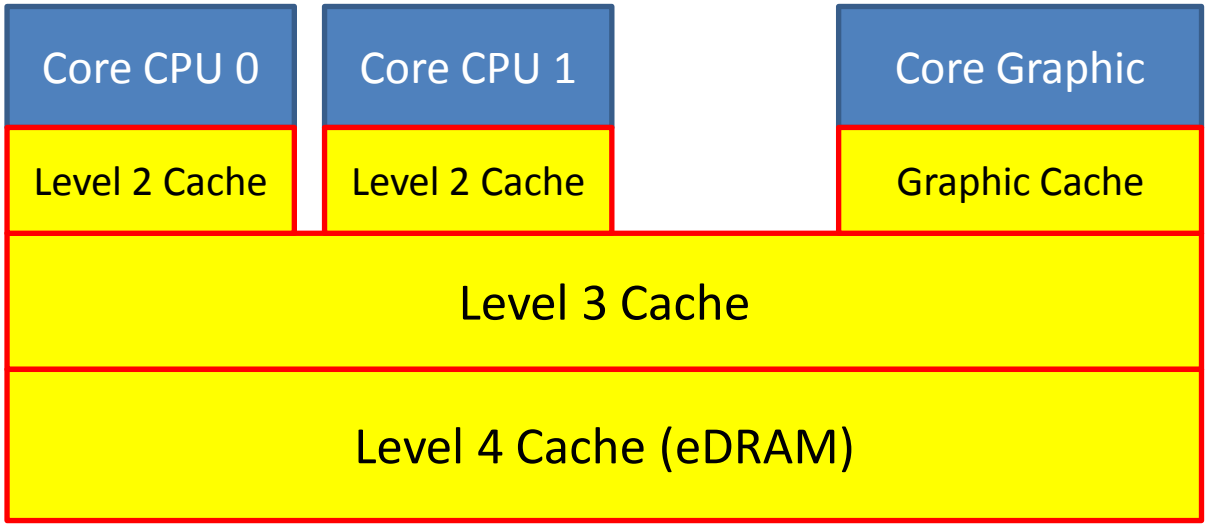


Quad-Core

L3 Cache

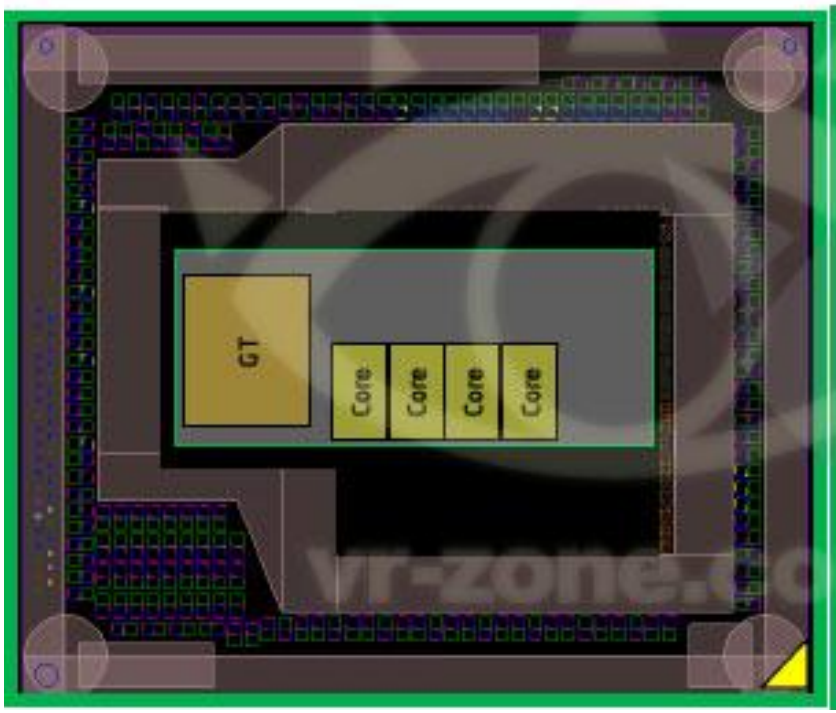

Queue, Uncore, I/O
Core
Core
Core
Core
Core
Core
Core
Core
Shared
L3 Cache*
Memory Controller

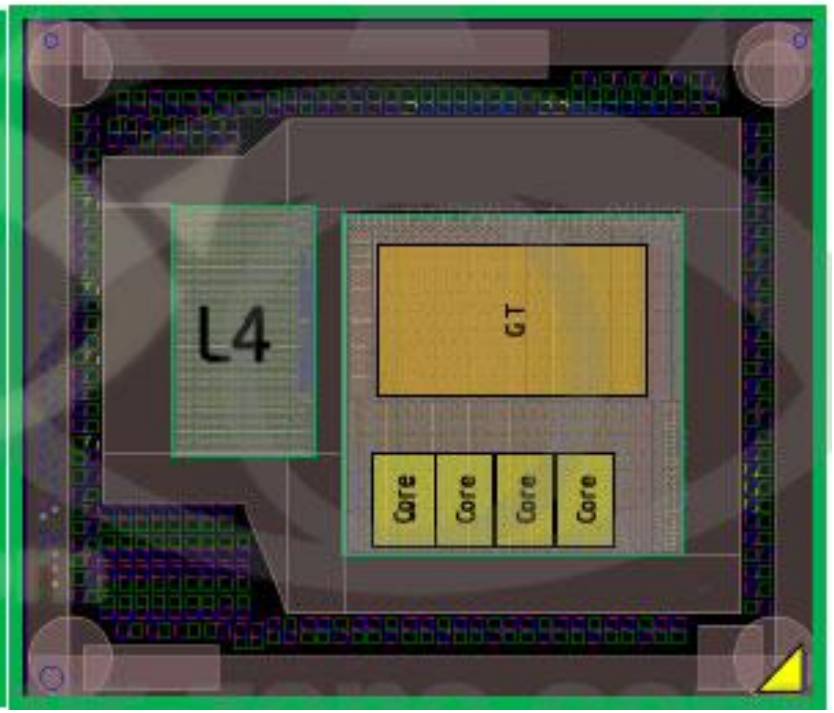
Architektura pamięci cache L4

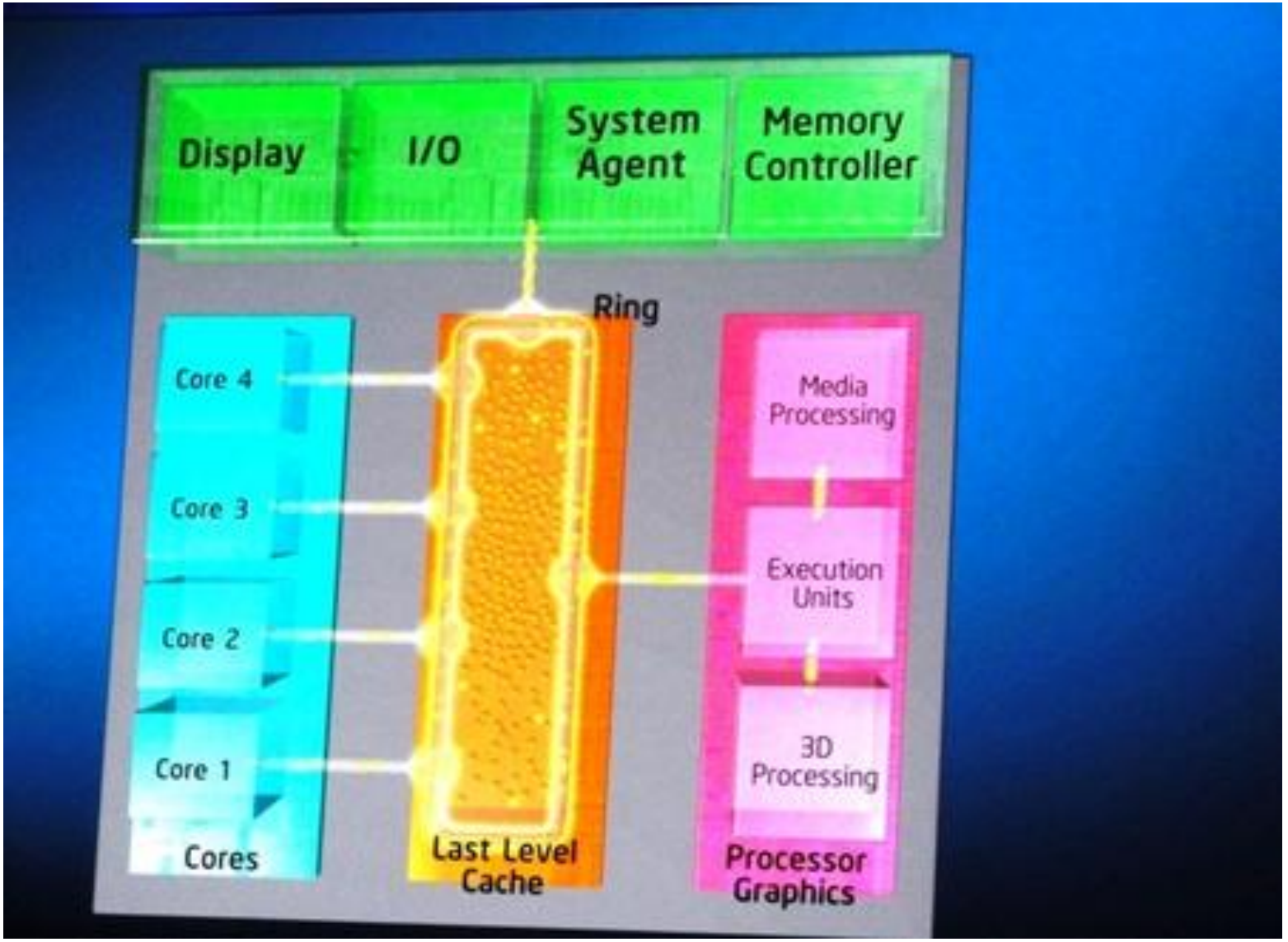


Intel Haswell, SkyLake

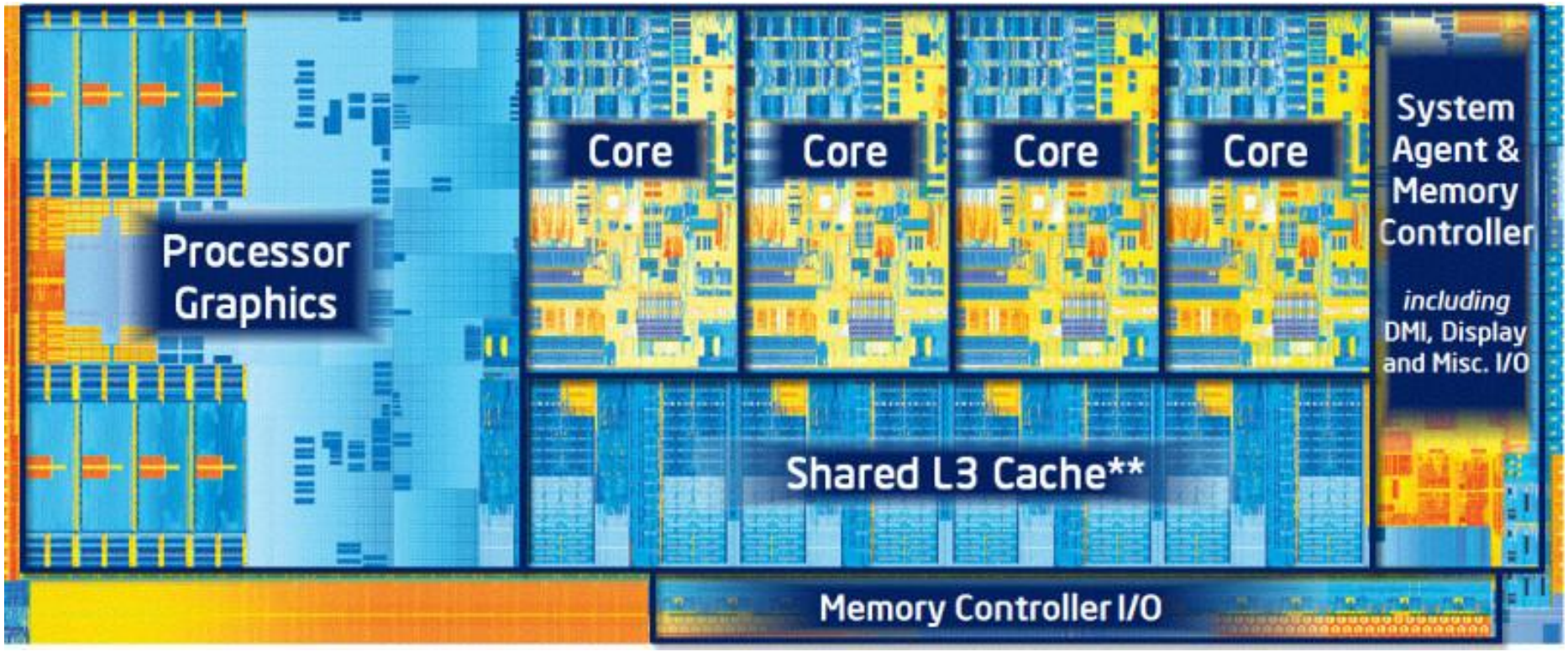
Intel Haswell



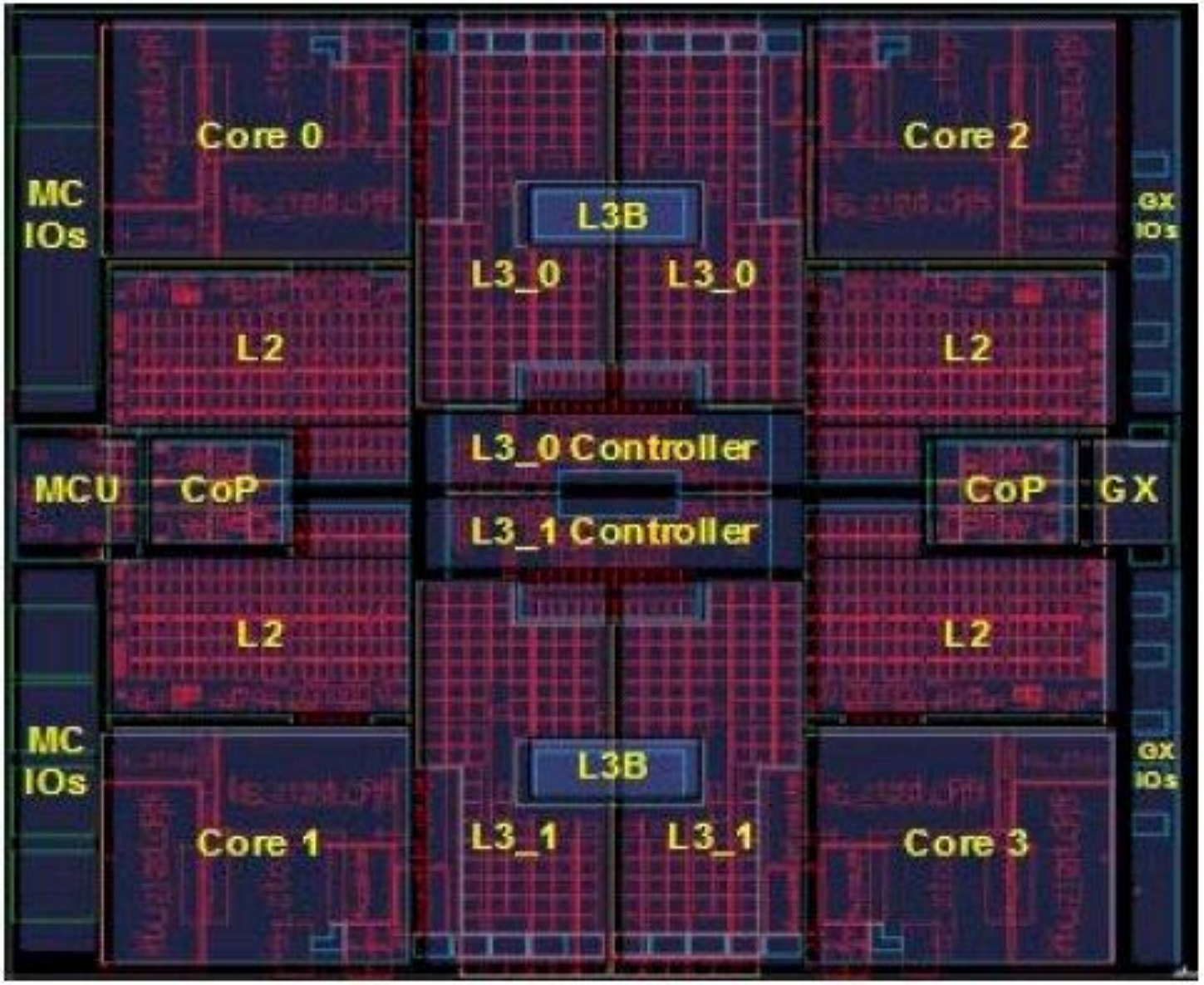



Display
I/O
System Agent
Memory Controller
Ring
Core 4
Core 3
Core 2
Core 1
Cores
Last Level Cache
Media Processing
Execution Units
3D Processing
Processor Graphics

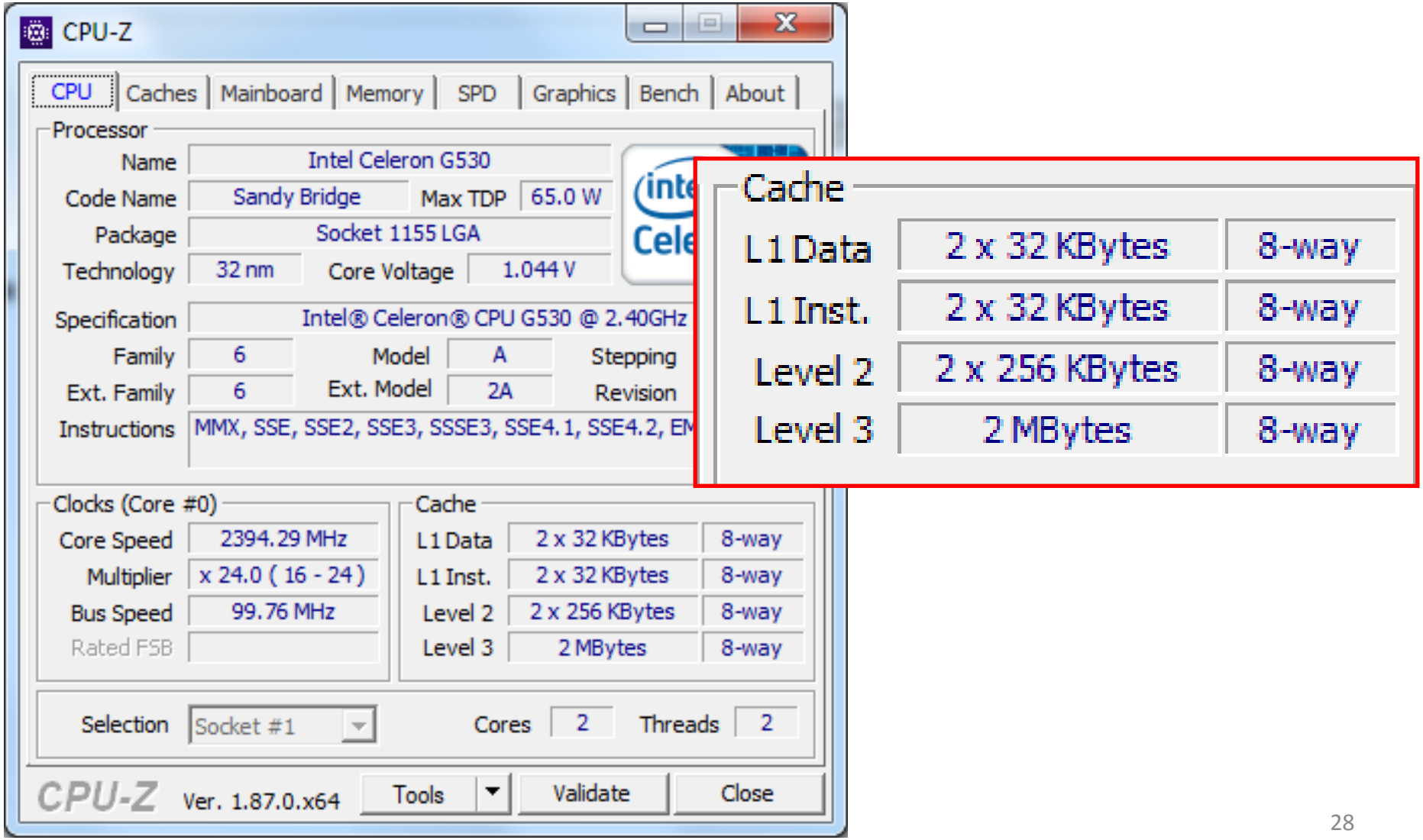
Cache w procesorze



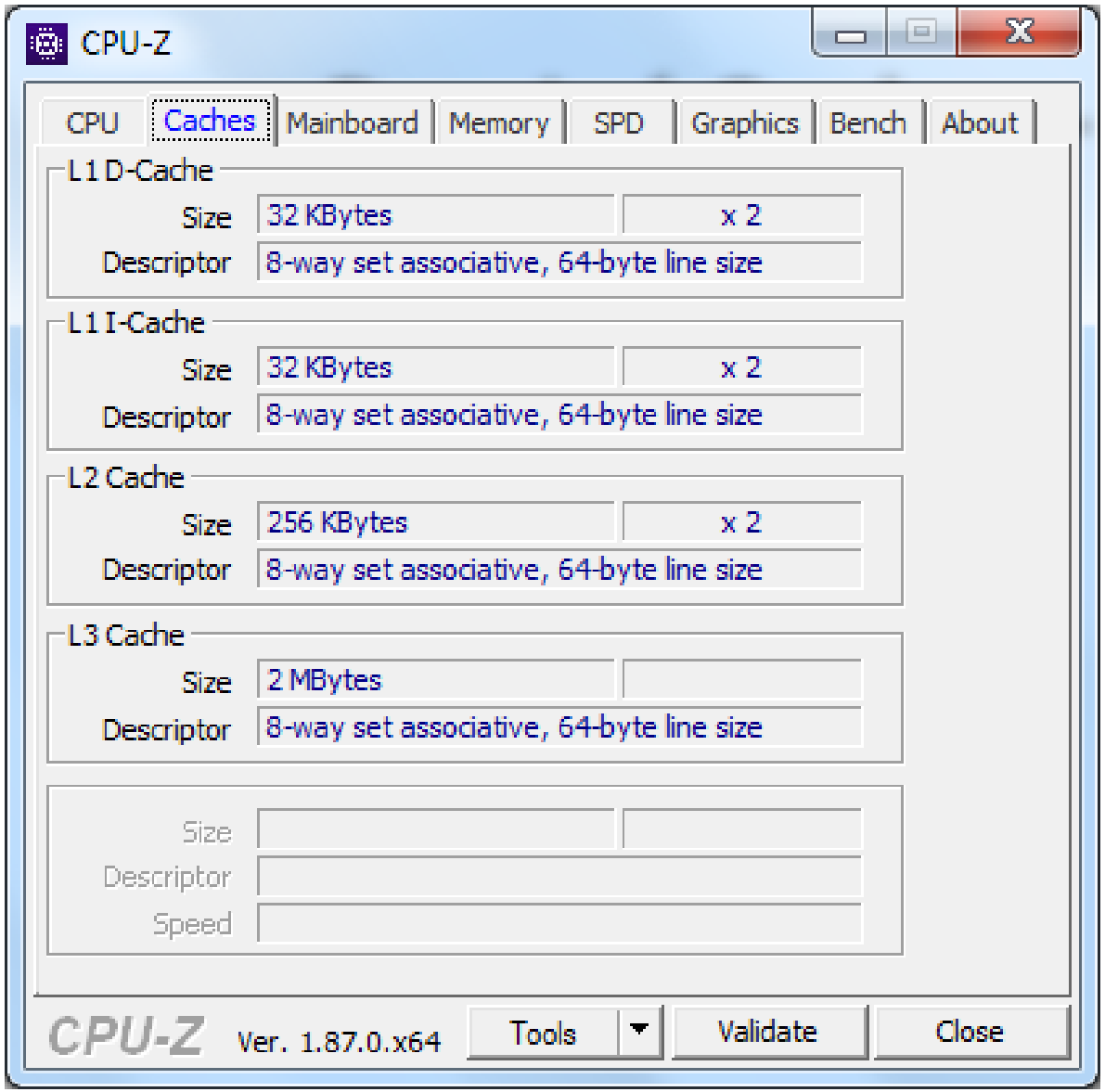
IBM's zEnterprise 196 CPU



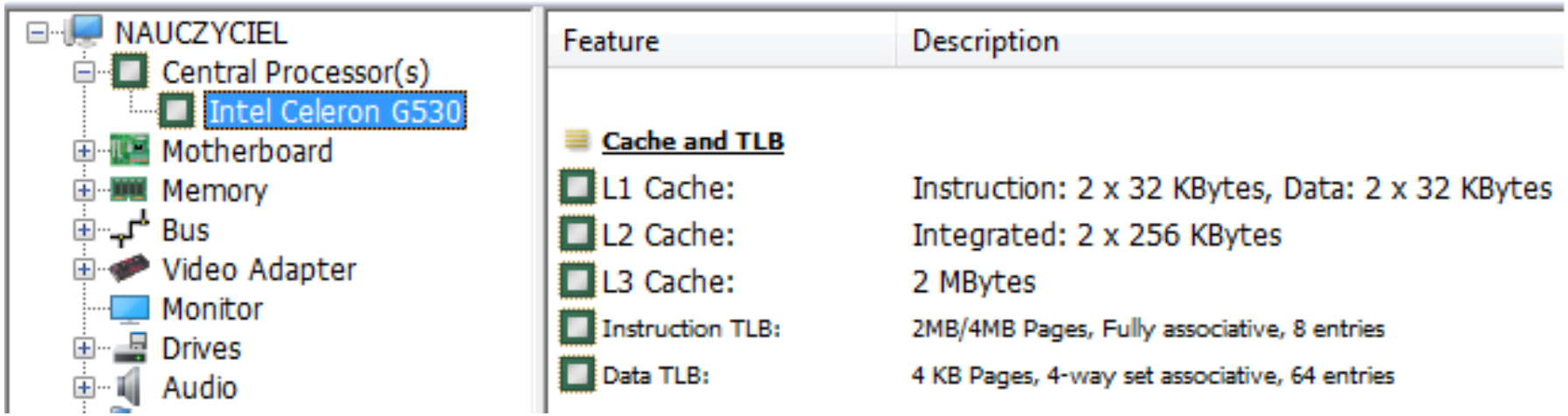
Pamięć Cache – CPU-Z



Pamięć Cache – CPU-Z



Pamięć Cache – HWinfo

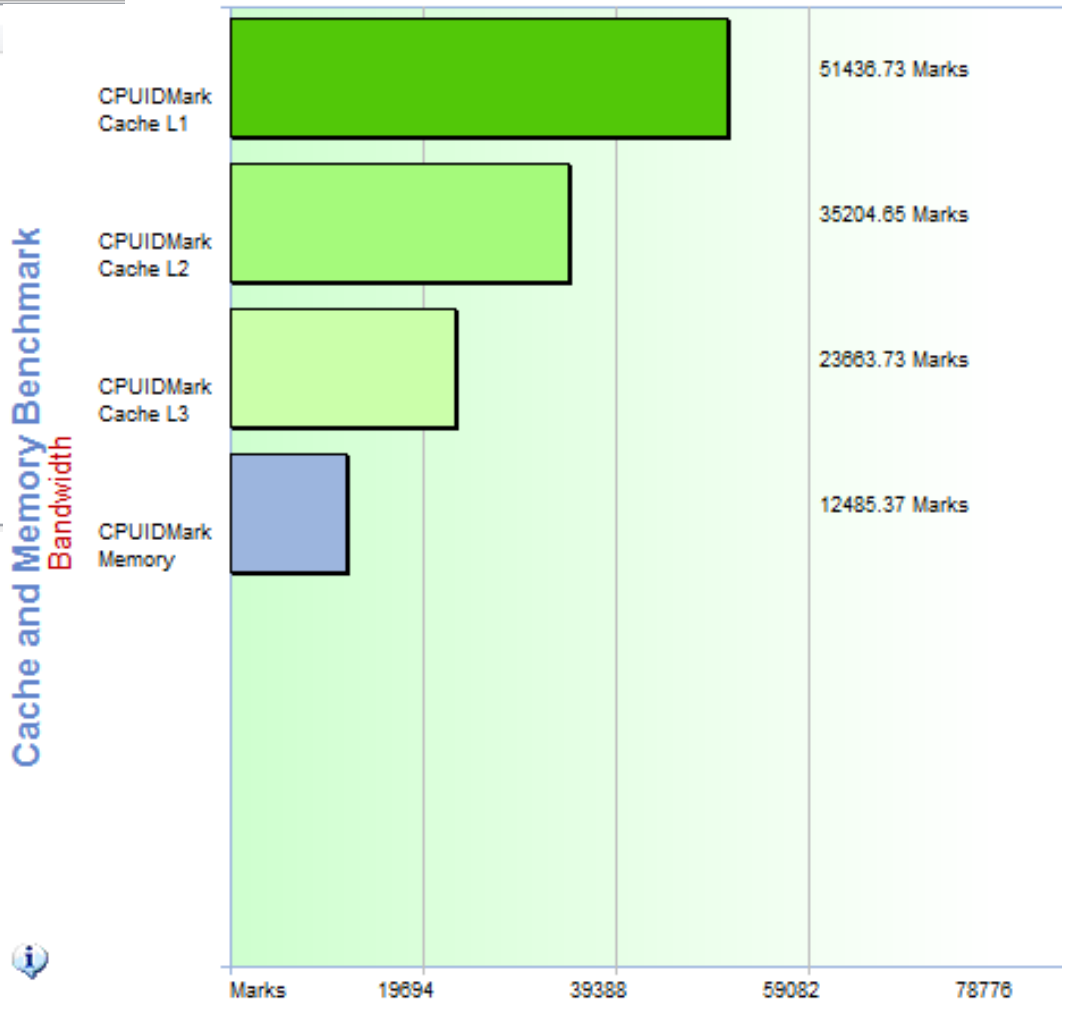


Porównanie osiągów pamięci Cache

Cache and Memory Benchmark

Item	Description
CPUIDMark Cache L1	51436.73 Marks
CPUIDMark Cache L2	35204.65 Marks
CPUIDMark Cache L3	23663.73 Marks
CPUIDMark Memory	12485.37 Marks
Latency Cache L1	4 cycles
Latency Cache L2	12 cycles
Latency Memory	122 cycles

CPUIDMark Memory :
30698

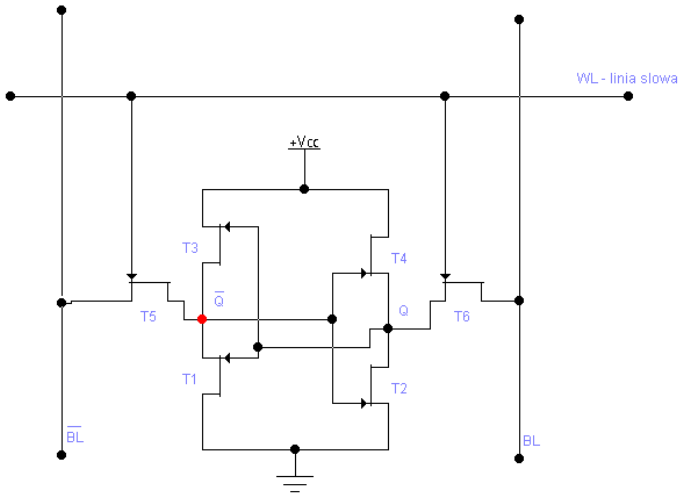


Memory Information :

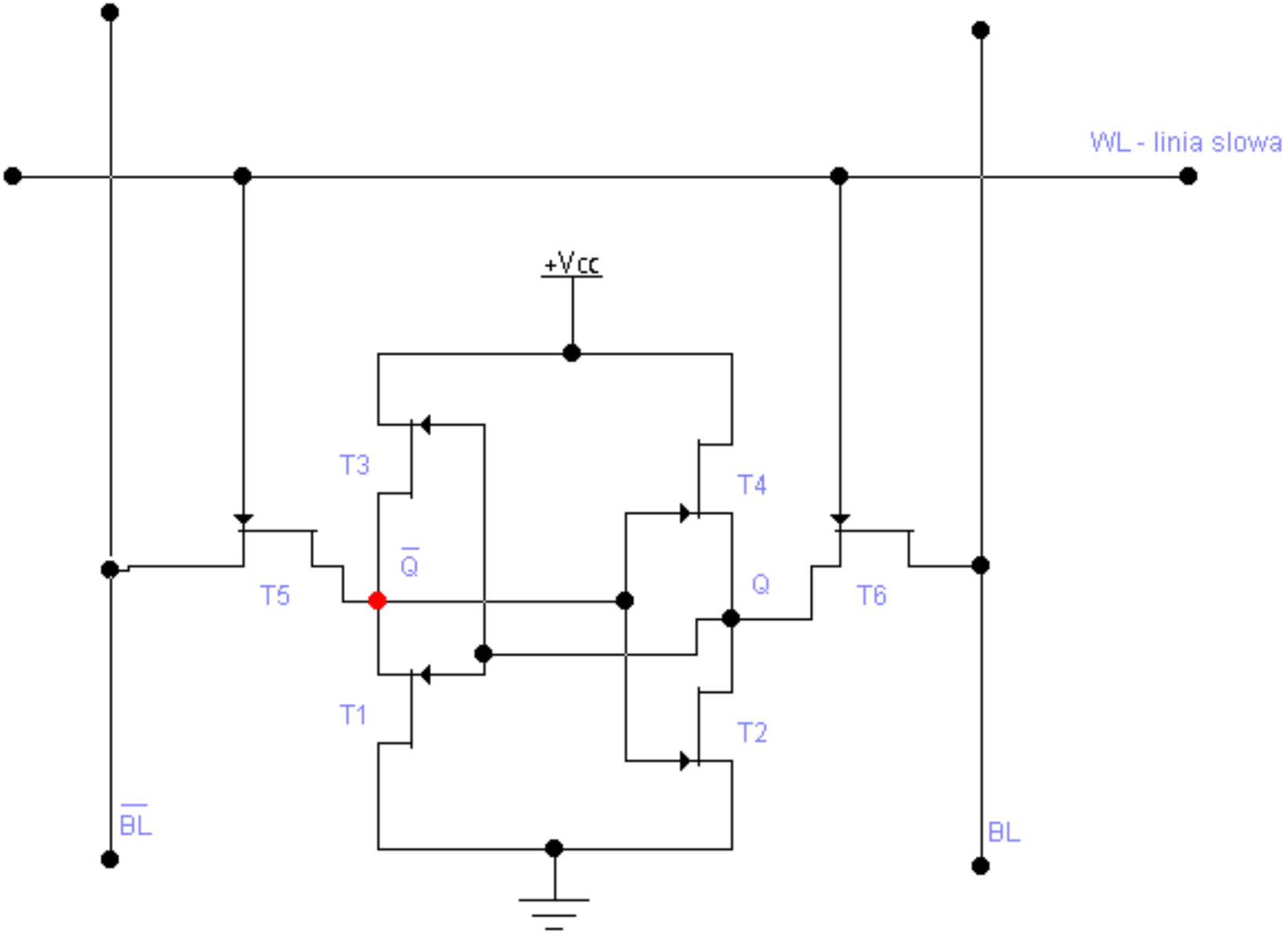
L1 Cache :	2 x 64 KB
L2 Cache :	2 x 256 KB
L3 Cache :	2048 KB
Total Memory :	4016 MB

Pamięć statyczna – S-RAM

- Pamięć statyczna – S-RAM (ang. Static Random Access Memory)
  - Pamięć S-RAM przechowuje dane tak długo, jak długo włączone jest zasilanie.
- Każdy bit przechowywany jest w układzie z czterech tranzystorów, które tworzą przerzutnik, oraz z dwóch tranzystorów sterujących.
- Taka struktura umożliwia szybkie odczytanie bitu i nie wymaga odświeżania.



Pamięć statyczna – S-RAM

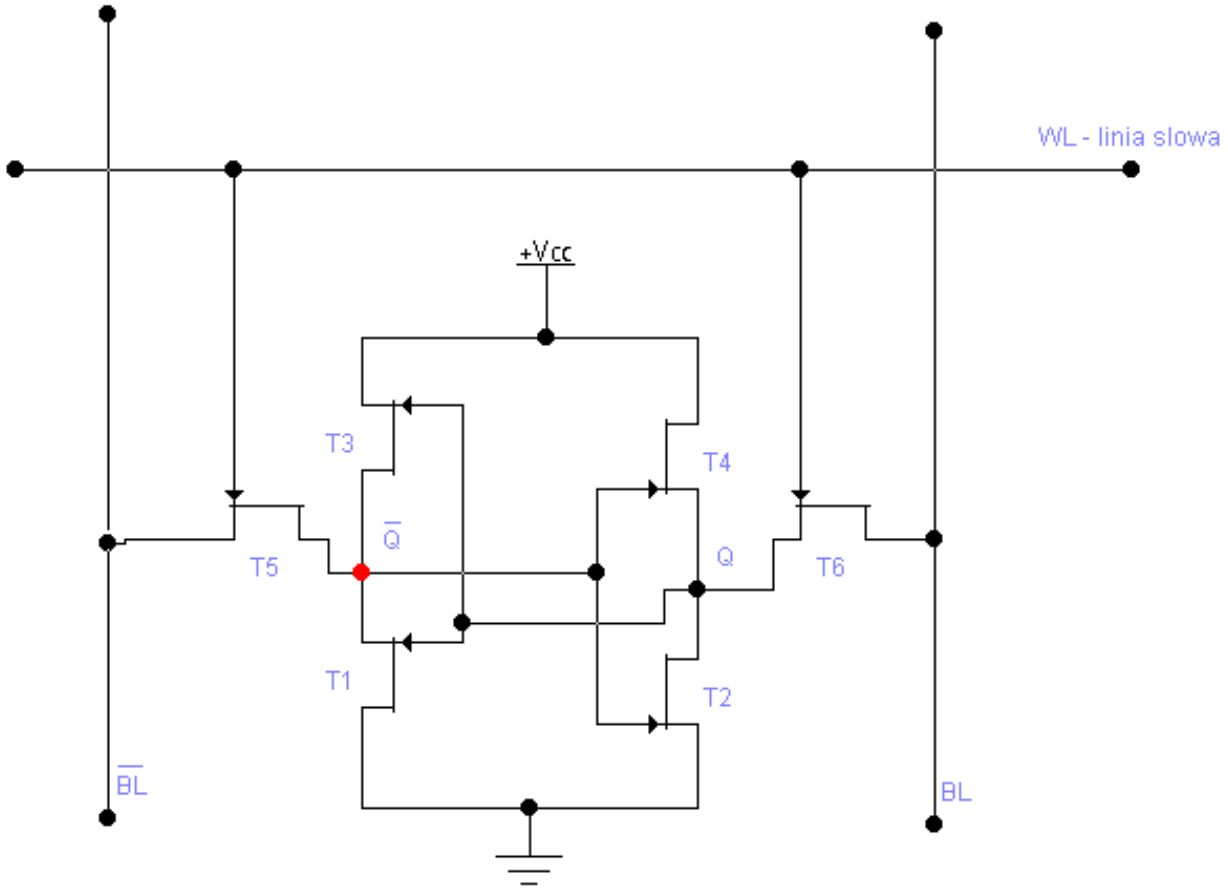


Pamięć statyczna – S-RAM

- Do przechowywania każdego bitu w pamięci statycznej RAM wykorzystywane są dwa krzyżowo sprzężone inwertory, zbudowane z tranzystorów polowych CMOS oznaczonych jako T_1, T_2 i T_3, T_4.
  - Inwertory tworzą prosty przerzutnik bistabilny, który posiada dwa stabilne stany wykorzystywane do zapisu poziomów logicznych 0 i 1.
- Dwa tranzystory T_5 i T_6 służą do sterowania dostępem do komórki podczas zapisu i odczytu danych. Są one podłączone do linii słowa WL.
- Odpowiednio wysterowane sygnałem na tej linii tranzystory T_5 i T_6 łączą wyjście Q przerzutnika z linią bitu BL oraz wyjście komplementarne ~Q z komplementarną linią ~BL.
  - Dwie linie bitu zwiększają poziom sygnału w stosunku do szumów (sygnałów zakłócających), które pojawiają się w strukturze pamięci półprzewodnikowych.
- Komórka pamięci statycznej RAM może być w jednym z trzech stanów:
  - Oczekiwania
  - Zapisu
  - Odczytu.

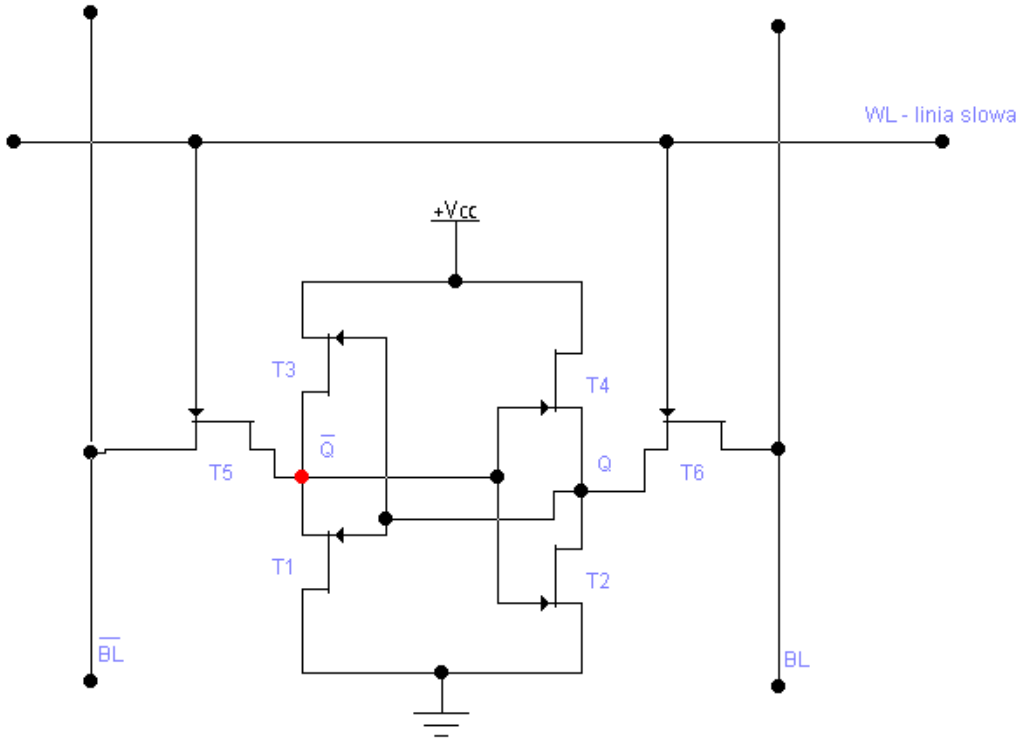
Stan oczekiwania S-RAM

- Gdy linia słowa WL nie jest wysterowana odpowiednim napięciem, tranzystory T_5 i T_6 separują wyjścia przerzutnika od linii bitów.
- Przerzutnik, zbudowany z dwóch sprzężonych wzajemnie inwertorów, pamięta swój stan wewnętrzny ustawiony przy poprzedniej operacji zapisu.



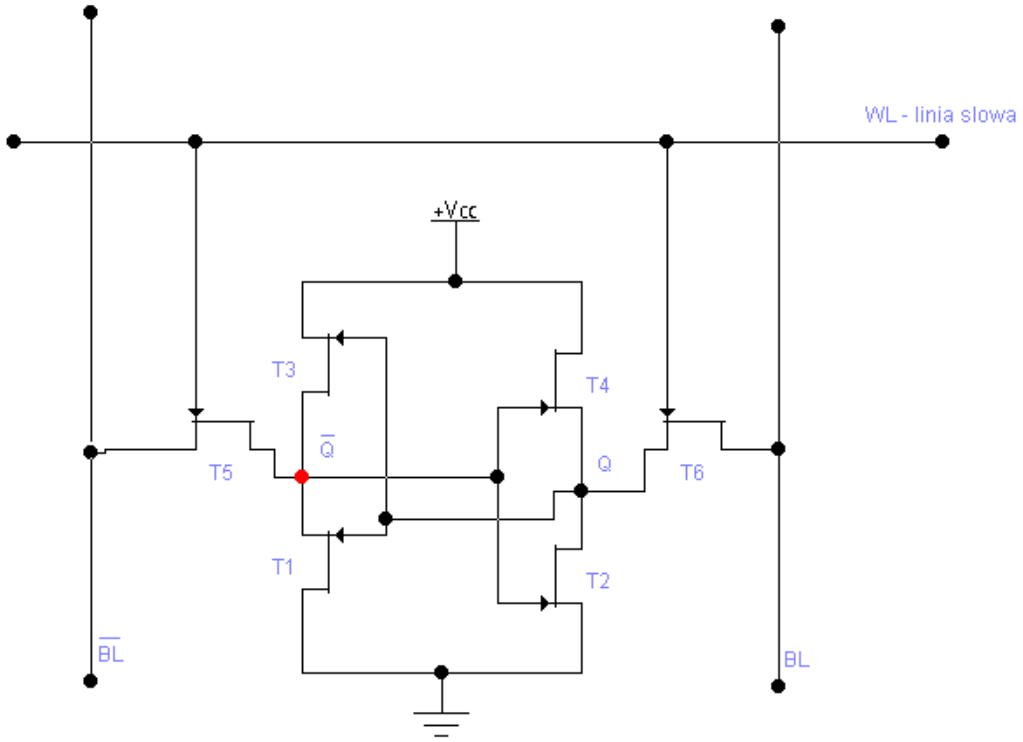
Zapis w S-RAM

- Przy zapisie ustawia się linie BL zgodnie z wartością zapisywanego bitu
  - (dla 1 BL = 1, BL = 0; dla 0 BL = 0, BL = 1).
- Następnie linia WL zostaje wysterowana i tranzystory T_5, T_6 łączą wejścia inwertorów z liniami BL powodując zapis informacji w przerzutniku.
- Jest to możliwe, ponieważ sygnał na liniach BL i BL jest wystarczająco mocny, aby wymusić zmianę stanu w tranzystorach T_1...T_4 przerzutnika, które są z reguły bardzo małe.



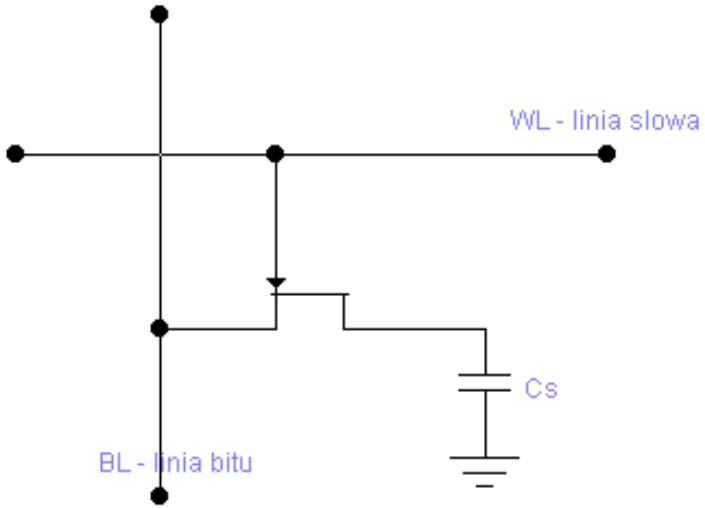
Odczyt z S-RAM

- Załóżmy, iż przerzutnik w komórce pamięci pamięta stan logiczny 1, czyli wyjście Q = 1, a Q = 0 (odblokowane tranzystory M_4 i M_1, zablokowane M_2 i M_3). Cykl odczytu rozpoczyna się przez naelektryzowanie obu linii bitów BL i BL do wartości logicznej 1, a następnie przez wysterowanie linii słowa WL, co spowoduje włączenie (odblokowanie) tranzystorów M_5 i M_6. Stan wyjść Q i Q zostaje przeniesiony na linie BL i BL. Linia BL pozostaje w stanie 1, gdyż odblokowane tranzystory M_4 i M_6 łączą ją z napięciem V_{dd}. Z kolei napięcie linii BL zostanie rozładowane do zera, ponieważ tranzystory M_1 i M_5 łączą ją z masą układu. W efekcie na linii BL pojawi się 1, a na BL 0. Jeśli przerzutnik przechowuje wartość 0, to otrzymamy sytuację odwrotną.

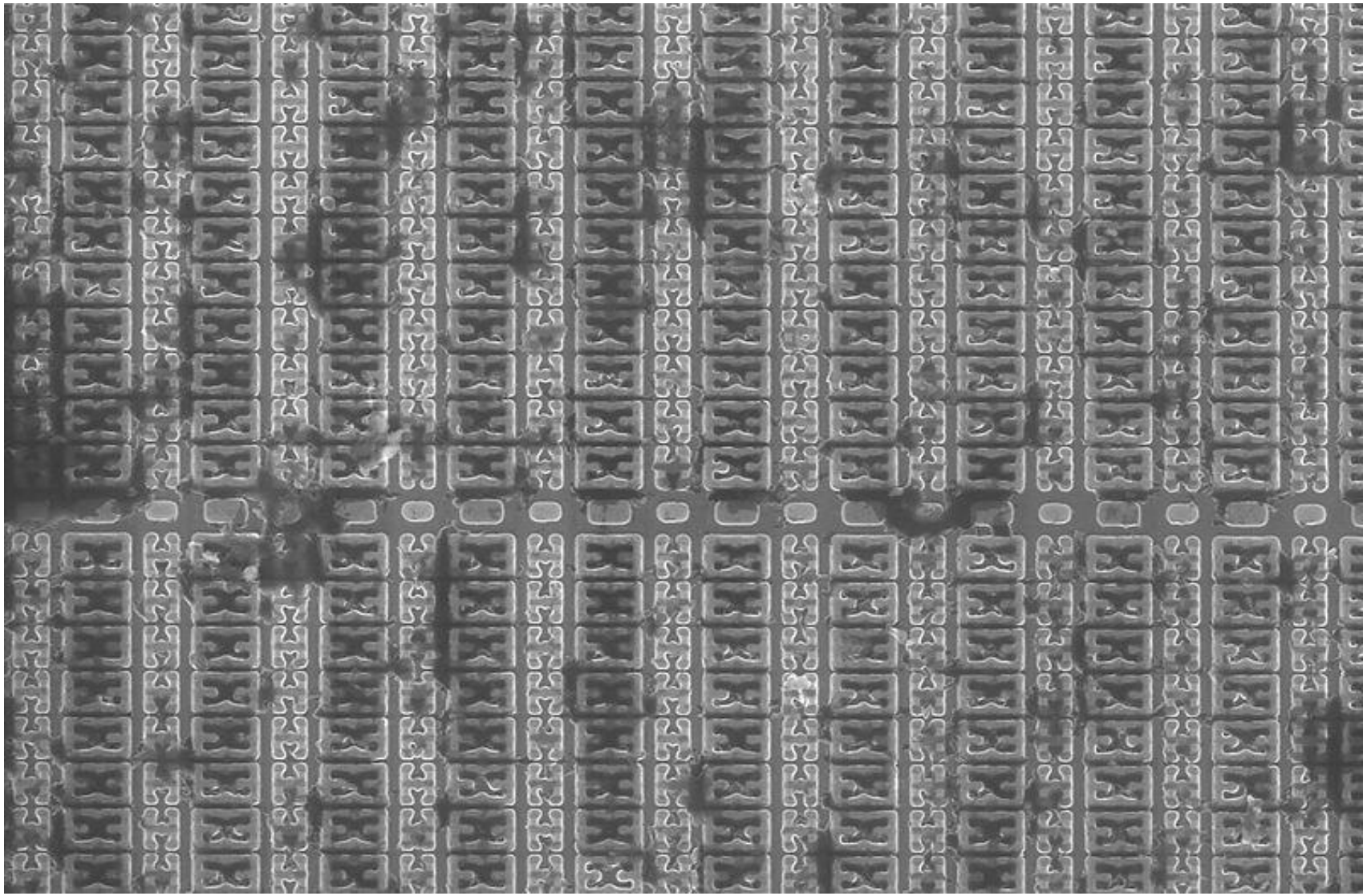


Odświeżanie S-RAM

- http://eduinf.waw.pl/inf/alg/002_struct/0043.php#Komorka



S-RAM



Historia pamięci Cache w PC

- Pamięć cache pojawiła się po raz pierwszy w procesorach Intel386 DX. Była zlokalizowana na zewnątrz CPU. Pamięć ta miała 64KB lub 128KB.
  - Niektóre płyty nie były w stanie obsłużyć jej, co pokazywało wyraźną różnicę w wydajności dzięki zastosowaniu pamięci podręcznej.
  - Model zapisu danych to write through gdzie procesor zapisywał wynik do pamięci podręcznej a z niej robiono od razu kopię do pamięci RAM.
- Kolejna generacja 486 miała wewnątrz procesora niewielką ilość (8KB) pamięci cache do dyspozycji CPU. Była to pamięć L1 (poziom1) lub wewnętrzna odróżniało ją to od pamięci zewnętrznej określanej jako L2. Ta druga miała 128KB lub 256KB.
  - Stosowano to architekturę write back. Wyniki operacji są przechowywane w pamięci cache ale uaktualnienie zawartości pamięci operacyjnej odbywa się w określonych momentach. Ten model stosowany jest do dziś.
- Pierwsze procesory Pentium miały 2 oddzielne pamięci cache L1-jedna do instrukcji, druga do danych. Miały one po 8KB.
  - Pamięć L2 w dalszym ciągu znajdowała się na płycie komputera. Rodzaj kości zależał od producenta płyty. Typowe wartości jej wynosiły 256KB lub 512KB.
  - AMD K5, K6 i K6-2 miały taką samą architekturę.
- AMD K6-III wprowadził dodatkowy poziom cache - L3 (level 3).
- Intel w procesorach P6 przesunął pamięć cache L2 do wnętrza CPU.
  - Przyśpieszyło to pracę pamięci. Była taktowana wewnętrznym zegarem procesora, szybszym od taktowania zegarem bazowym.
- W 2012 roku Intel wprowadził pamięć cache poziomu 4 – L4. Była to pamięć eDRAM.

Algorytm zastępowania

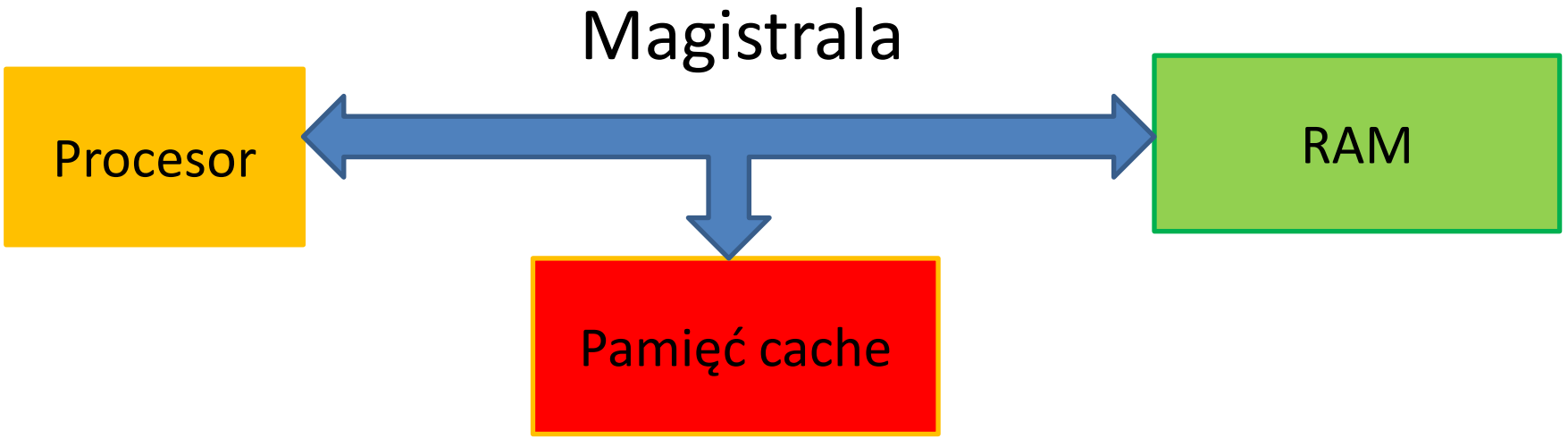
- W przypadku metody skojarzeniowej i sekcyjno-skojarzeniowej wymagany jest algorytm zastępowania.
- **Algorytm „najmniej ostatnio używany"** (ang. least-recently used -LRU).
  - Należy zastąpić ten blok w sekcji, który pozostawał w pamięci podręcznej najdłużej bez odwoływania się do niego.
- **Algorytm „pierwszy wchodzi - pierwszy wychodzi"** (ang. fist in –first out- FIFO).
  - Polega na zastępowaniu tego bloku w sekcji, który najdłużej pozostawał w pamięci podręcznej.
- **Algorytm „najrzadziej używany"** (ang. least frequently used-LFU).
  - Zastępowany jest blok w sekcji, którego miał najmniej odniesień.
- Metodą odmienną od ilości odniesień jest **losowy wybór wśród kandydujących wierszy**.

TOPOLOGIE PAMIĘCI CACHE

Topologie pamięci cache

- Look- Aside (dostęp bezpośredni)
  - Procesor odwołuje się do cache wykorzystując magistralę pamięciową.
- Look - Throgh (dostęp „przez”)
  - Układ pamięci podręcznej pośredniczy w dostępie procesora do RAM.
- Backside (dostęp „z tyłu”)
  - Układ pamięci podręcznej jest dołączony do procesora przez oddzielną magistralę nazywaną BSB (Back Side Bus).

Look - Aside (dostęp bezpośredni)

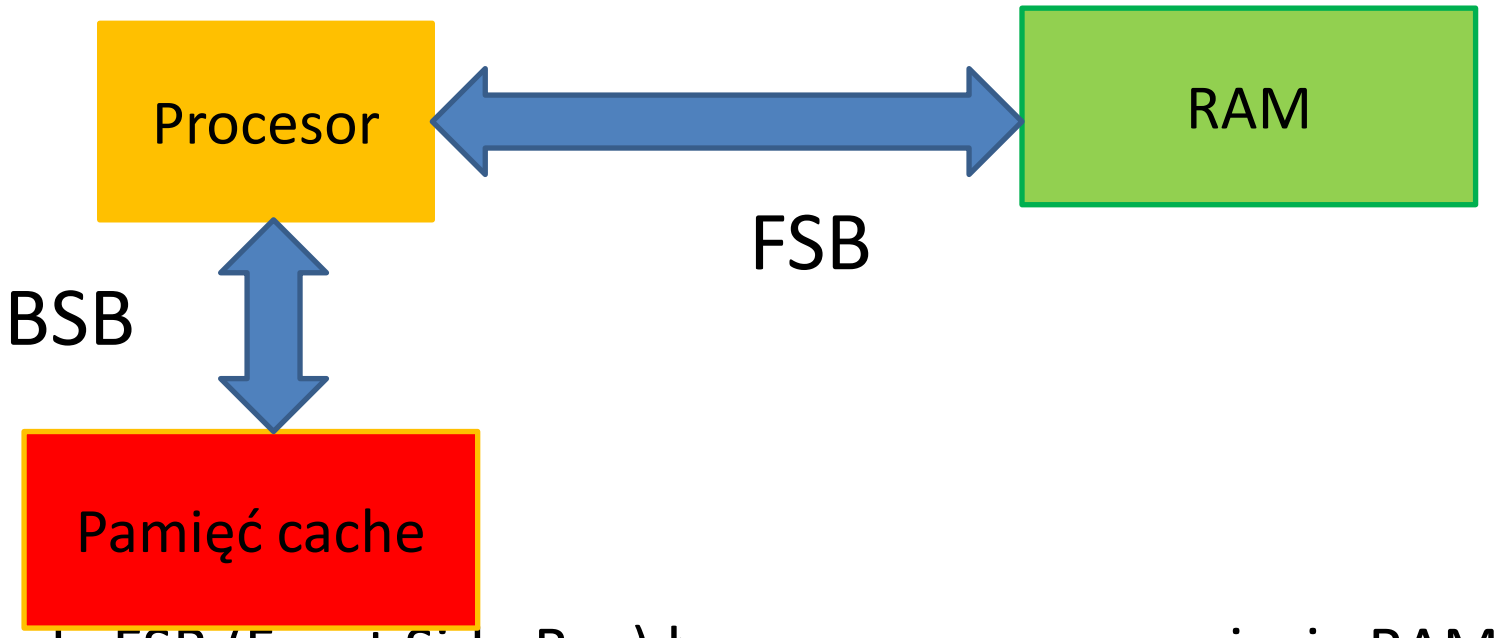


- Procesor odwołuje się do cache wykorzystując magistralę pamięciową.
- Pamięć podręczna jest podłączona równolegle z pamięcią operacyjną RAM.
  - częstotliwość pracy obu pamięci jest taka sama (komunikacja odbywa się po wspólnej magistrali),
  - tylko czas dostępu dzięki szybkości cache może ulec skróceniu.
- Wykorzystanie tej samej magistrali nie jest korzystne. Jest blokowana przy każdym dostępie procesora do cache i nie może być w tym samym czasie udostępniona innym urządzeniom.

Look - Throgh (dostęp „przez”)

- Układ pamięci podręcznej pośredniczy w dostępie procesora do RAM.
- Procesor odwołuje się do układu cache, natomiast ten układ jest dołączony przez magistralę pamięciową do RAM.

Backside (dostęp „z tyłu”)



- Magistrala FSB (Front Side Bus) łączy procesor z pamięcią RAM.
- Pamięć podręczna jest dołączona do procesora przez oddzielną magistralę BSB (Back Side Bus).
- Częstotliwości obu magistral są niezależne od siebie.
- Możliwe jest wykorzystanie magistrali FSB przez inne urządzenia zapisujące do pamięci RAM, w czasie gdy procesor komunikuje się z pamięcią podręczną po BSB.