



Ministerstwo Nauki
i Szkolnictwa Wyższego

**Regionalna Inicjatywa Doskonałości w Dyscyplinach
Informatyki, Elektrotechniki, Elektroniki, Automatyki i Robotyki
na Politechnice Częstochowskiej**



Projekt w ramach programu Regionalna Inicjatywa Doskonałości, decyzja nr 020/RID/2018/19

Architektura Systemów Komputerowych

Analiza architektur i ocena wydajności przykładowych systemów komputerowych

dr hab. inż. Łukasz Szustak, prof. PCz

lszustak@icis.pcz.pl

Katedra Informatyki

Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska



Wydział Inżynierii
Mechanicznej
i Informatyki
The Faculty of Mechanical Engineering
and Computer Science



Ministerstwo Nauki
i Szkolnictwa Wyższego

Regionalna Inicjatywa Doskonałości w Dyscyplinach Informatyki, Elektrotechniki, Elektroniki, Automatyki i Robotyki na Politechnice Częstochowskiej



Projekt w ramach programu Regionalna Inicjatywa Doskonałości, decyzja nr 020/RID/2018/19

**Zamieszczane materiały służą wyłącznie do celów
indywidualnego kształcenia. Nie wyrażam zgody na ich
utrwalanie przekazywanie osobom trzecim ani
rozpowszechnianie.**

dr hab. inż. Łukasz Szustak, prof. PCz
lszustak@icis.pcz.pl

Katedra Informatyki
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska



Wydział Inżynierii
Mechanicznej
i Informatyki
The Faculty of Mechanical Engineering
and Computer Science

Wydajność procesora

- Obecnie teoretyczna wydajność systemów komputerowych mierzona jest w liczbie wykonywanych operacji zmiennoprzecinkowych na sekundę
- Flop/s („FLOPS”) jest jednostką mocy obliczeniowej komputerów
- Flop/s (floating point operations per second) – liczba operacji zmiennoprzecinkowych na sekundę
- Obecnie zwykle używa się wielokrotności:
 - Mflop/s = 10^6 x flop/s
 - Gflop/s = 10^9 x flop/s
 - Tflop/s = 10^{12} x flop/s
- Obecne „standardowe” architektury komputerowe cechuje wydajność rzędu kilkudziesięciu- kilkuset Gflop/s

Wydajność procesora

- Wydajność procesora:
 - Teoretyczna – wydajność na jaką w teorii pozwala dana architektura komputerowa
 - Rzeczywista – wydajność osiągnięta w praktyce dla konkretnej aplikacji

Wydajność procesora

- Wydajność procesora:
 - Teoretyczna – wydajność na jaką w teorii pozwala dana architektura komputerowa
 - Rzeczywista – wydajność osiągnięta w praktyce dla konkretnej aplikacji



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,299,072	415,530.0	513,854.7	28,335
2	Summit - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096

<https://top500.org/lists/top500/2020/06/>

Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * i$$

- f - częstotliwość taktowania procesora
- i – „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na **procesor**

Teoretyczna wydajność procesora

- Architektury wielordzeniowe:
Wykład 4 A, slajd 15



Teoretyczna wydajność procesora

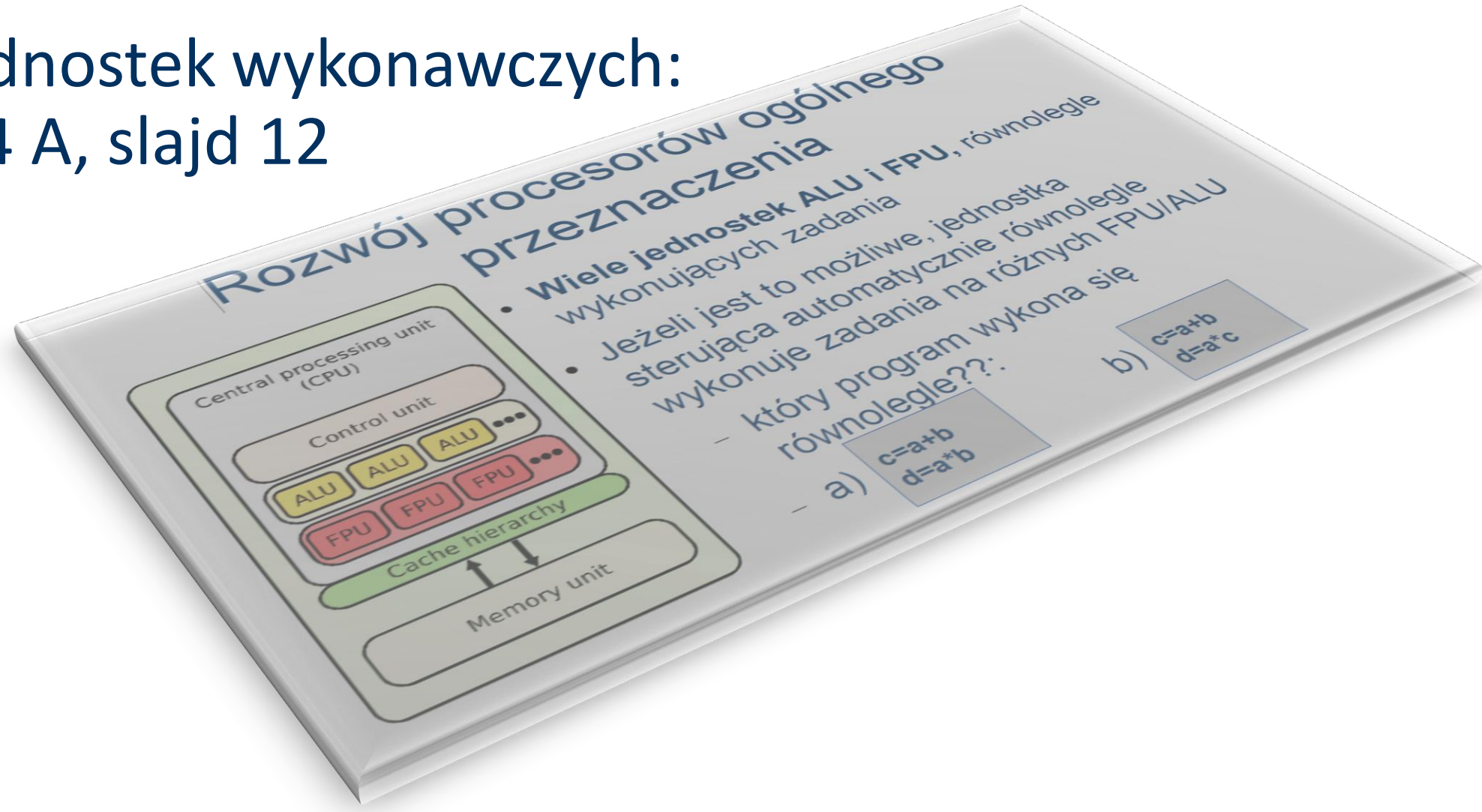
- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * c * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na rdzeń

Teoretyczna wydajność procesora

- Wiele jednostek wykonawczych:
Wykład 4 A, slajd 12

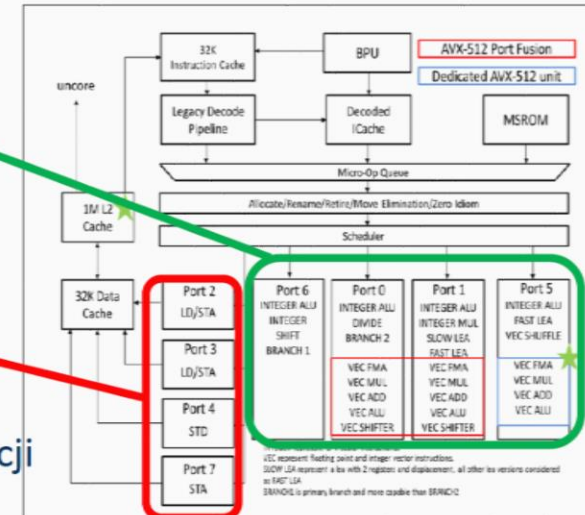


Teoretyczna wydajność procesora

- Wiele jednostek wykonawczych: Wykład 5, slajdy 27 - 32

Jednostka centralna CPU: budowa - Out-of-Order Engine

- Dyspozytor zasobów może maksymalnie uruchamiać osiem mikro-operacji w każdym cyklu, z użyciem 8 portów
- Cztery porty wykorzystują zasoby obliczeniowe: każdy z tych portów odnosi się do innej jednostki ALU, dwa z nich odnoszą się do dedykowanych jednostek FMA, jak również jednostek SIMD oraz zmiennopozycyjnych
- Pozostałe cztery porty determinują wykonywanie instrukcji na pamięci
- Uwzględniając przetwarzanie potokowe, the out-of-order engine dla Skylake umożliwia jednoczesne przetwarzanie aż 224 mikrooperacji (Haswell - 192, Sandy Bridge - 168)



Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * c * u * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na **jednostkę**

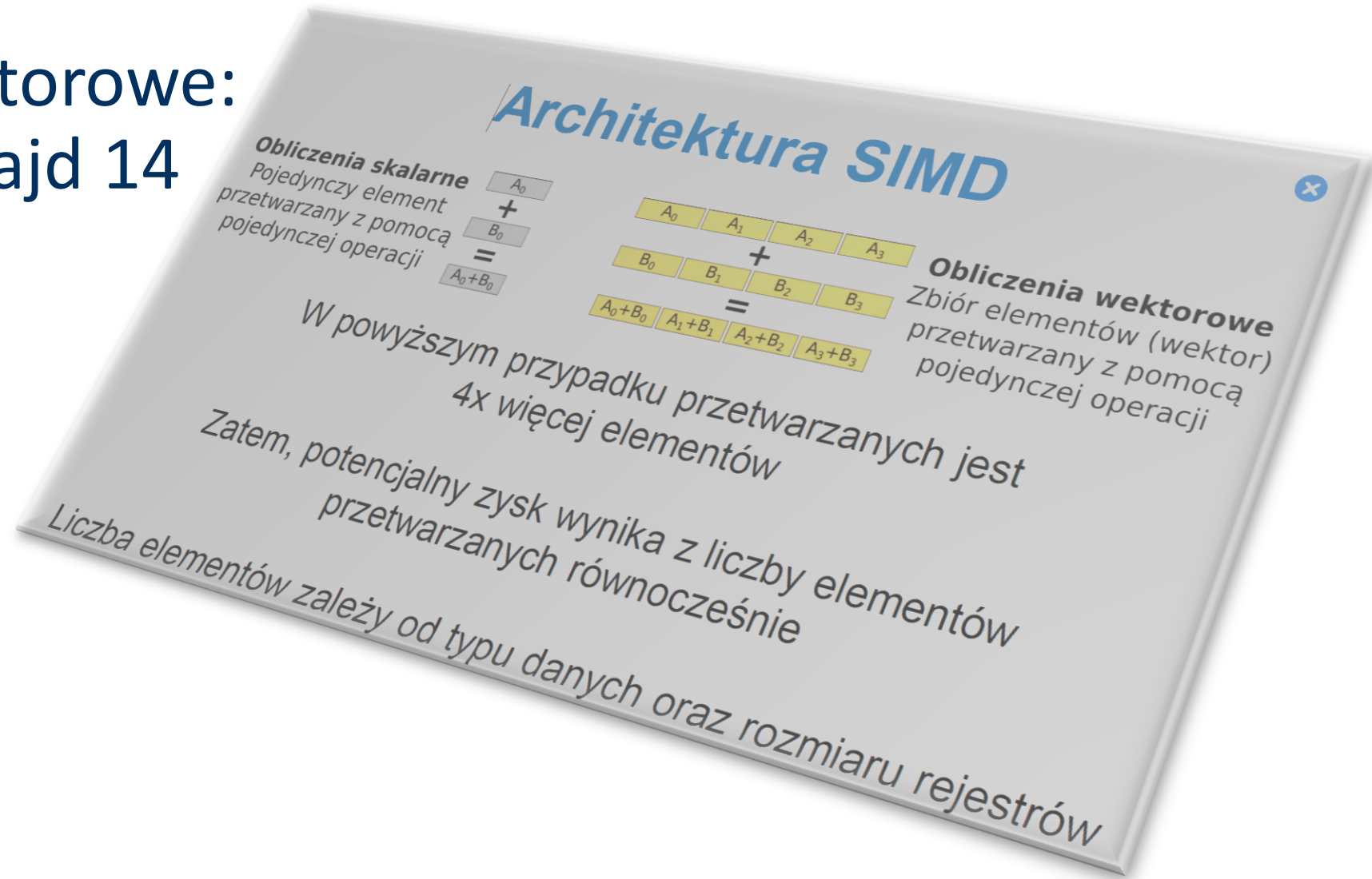
Teoretyczna wydajność procesora

- Jednostki wektorowe:
Wykład 4 A, slajd 16



Teoretyczna wydajność procesora

- Jednostki wektorowe:
Wykład 4 B, slajd 14



Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- v - liczba jednocześnie przetwarzanych elementów w wektorze
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na **jednostkę**

Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

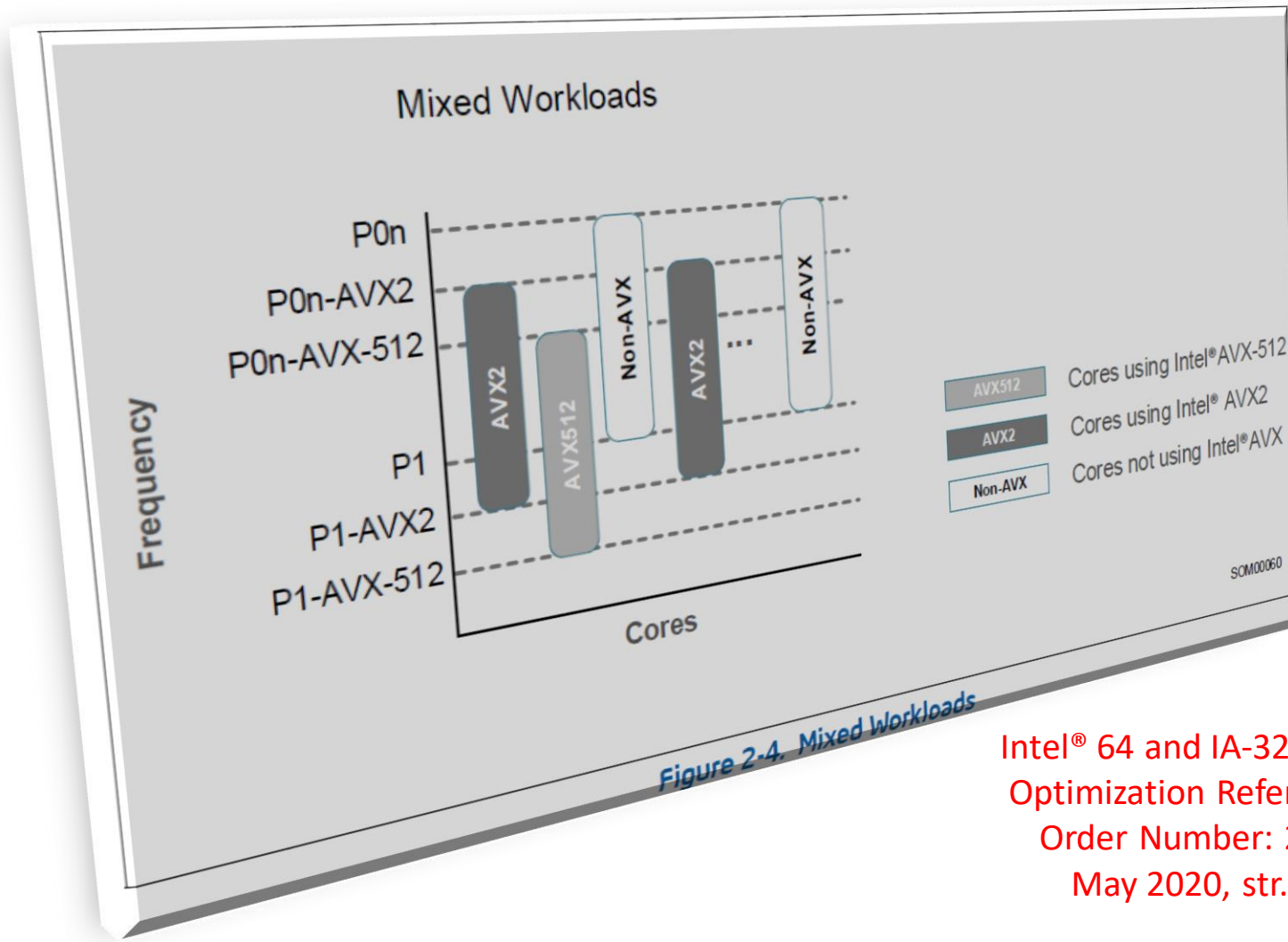
$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- v - liczba jednocześnie przetwarzanych elementów w wektorze
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na jednostkę

Teoretyczna wydajność procesora: częstotliwość procesora

- Częstotliwość bazowa określa wartość referencyjną „gwarantowaną” przez producenta procesorów
- Jednakże, obecne procesory oferują bardzo dynamiczny zakres częstotliwości, której wartość w praktyce zależy od różnych czynników:
 - Głównym czynnikiem ograniczającym częstotliwość jest temperatura procesora, dla której przekroczenie wartości bezpiecznej dla procesora skutkuje redukcją częstotliwość
 - Turbo – dodatkowy przyrost częstotliwość pod warunkiem spełnienia warunków termicznych
 - Każdy rdzeń może mieć różną częstotliwość
 - Typ i wielkość obciążenia

Teoretyczna wydajność procesora: częstotliwość procesora



- Częstotliwość bazowa określa wartość referencyjną „gwarantowaną” przez producenta procesorów
- Jednakże, obecne procesory oferują bardzo dynamiczny zakres częstotliwości, której wartość w praktyce zależy od różnych czynników:
 - Głównym czynnikiem ograniczającym częstotliwość jest temperatura procesora, dla której przekroczenie wartości bezpiecznej dla procesora skutkuje redukcją częstotliwość
 - Turbo – dodatkowy przyrost częstotliwość pod warunkiem spełnienia warunków termicznych
 - Każdy rdzeń może mieć różną częstotliwość
 - Typ i wielkość obciążenia

Intel® 64 and IA-32 Architectures
Optimization Reference Manual
Order Number: 248966-043
May 2020, str. 50 (2-14)

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

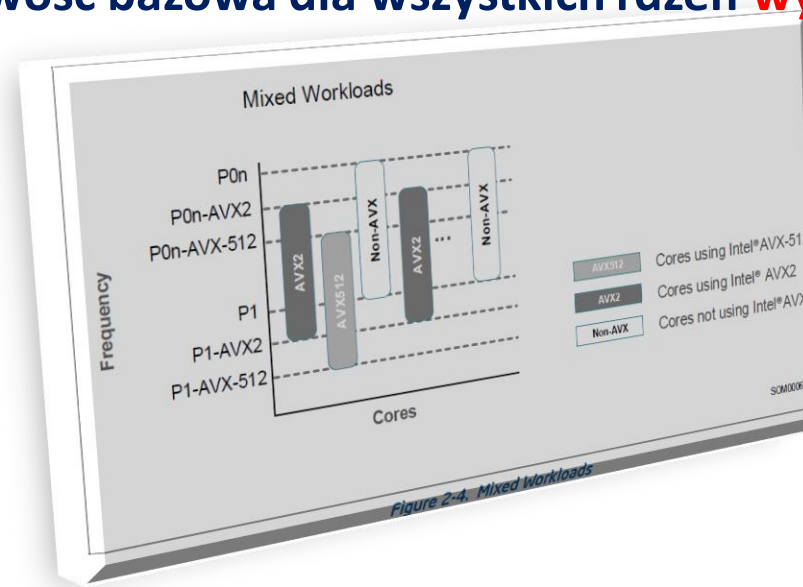
- *„Processor Base Frequency: 2.7 GHz”*
- *„Max Turbo Frequency 4.0 GHz”*
- *„Recommended Customer Price \$10009”*

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):
 - „*Processor Base Frequency: 2.7 GHz*”
 - „*Max Turbo Frequency 4.0 GHz*”
 - „*Recommended Customer Price \$10009*”
- W praktyce:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni **wykonujących obliczenia nie wektorowe**

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>



Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):
 - „*Processor Base Frequency: 2.7 GHz*”
 - „*Max Turbo Frequency 4.0 GHz*”
 - „*Recommended Customer Price \$10009*”
- W praktyce:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni wykonujących obliczenia nie wektorowe
 - 4.0 GHz częstotliwość turbo dla **aktywnych maksymalnie 2 rdzeni (pozostałe rdzenie nic nie robią) wykonujących obliczenia nie wektorowe**

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):
 - „Processor Base Frequency: 2.7 GHz”
 - „Max Turbo Frequency 4.0 GHz”
 - „Recommended Customer Price \$10009”
- W praktyce:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni wykonujących obliczenia nie wektorowe
 - 4.0 GHz częstotliwość turbo dla **aktywnych maksymalnie 2 rdzeni (pozostałe rdzenie nic nie robią) wykonujących obliczenia nie wektorowe**

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

					# of active cores / maximum core frequency in turbo mode (GHz)																											
SKU	Cores	LLC (MB)	TDP (W)	Base	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
				non-AVX Core Freq (GHz)																												
8280	28	38.5	205	2.7	4.0	4.0	3.8	3.8	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.5	3.5	3.5	3.5	3.3	3.3	3.3	3.3

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):
 - „*Processor Base Frequency: 2.7 GHz*”
 - „*Max Turbo Frequency 4.0 GHz*”
 - „*Recommended Customer Price \$10009*”
- W praktyce:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni wykonujących obliczenia nie wektorowe
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni (pozostałe rdzenie nic nie robią) wykonujących obliczenia nie wektorowe

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

					# of active cores / maximum core frequency in turbo mode (GHz)																											
SKU	Cores	LLC (MB)	TDP (W)	Base	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
				non-AVX Core Freq. (GHz)																												
8280	28	38.5	205	2.7	4.0	4.0	3.8	3.8	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.5	3.5	3.5	3.5	3.3	3.3	3.3	3.3

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

- „*Processor Base Frequency: 2.7 GHz*”
- „*Max Turbo Frequency 4.0 GHz*”
- „*Recommended Customer Price \$10009*”

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

- W praktyce – obliczenia nie wektorowe (skalarne)
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeń
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

- „Processor Base Frequency: 2.7 GHz”
- „Max Turbo Frequency 4.0 GHz”
- „Recommended Customer Price \$10009”

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

- W praktyce – obliczenia nie wektorowe
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni
- W praktyce – obliczenia wektorowe **AVX-512 (512 bitów)**

				# of active cores / maximum core frequency in turbo mode (GHz)																												
SKU	Cores	LLC (MB)	TDP (W)	Base AVX-512 Core Freq. (GHz)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
8280	28	38.5	205	1.8	2.7	3.7	3.5	3.5	3.4	3.4	3.4	3.4	3.3	3.3	3.3	3.3	2.9	2.9	2.9	2.9	2.7	2.7	2.7	2.7	2.5	2.5	2.5	2.5	2.4	2.4	2.4	2.4

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

- „*Processor Base Frequency: 2.7 GHz*”
- „*Max Turbo Frequency 4.0 GHz*”
- „*Recommended Customer Price \$10009*”

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

- W praktyce – obliczenia nie wektorowe
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeń
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni
- W praktyce – obliczenia wektorowe **AVX-512 (512 bitów)**
 - 1.8 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 2.4 GHz częstotliwość turbo dla wszystkich rdzeni

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

- „Processor Base Frequency: 2.7 GHz”
- „Max Turbo Frequency 4.0 GHz”
- „Recommended Customer Price \$10009”

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

- W praktyce – obliczenia nie wektorowe:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni
- W praktyce – obliczenia wektorowe **AVX-2.0 (256 bitów)**

				# of active cores / maximum core frequency in turbo mode (GHz)																												
SKU	Cores	LLC (MB)	TDP (W)	Base AVX 2.0 Core Freq. (GHz)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
8280	28	38.5	205	2.2	3.8	3.8	3.6	3.6	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.5	3.3	3.3	3.3	3.3	3.0	3.0	3.0	3.0	2.9	2.9	2.9	2.9

Teoretyczna wydajność procesora: częstotliwość procesora - przykład

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):

- „*Processor Base Frequency: 2.7 GHz*”
- „*Max Turbo Frequency 4.0 GHz*”
- „*Recommended Customer Price \$10009*”

<https://ark.intel.com/content/www/us/en/ark/products/192478/intel-xeon-platinum-8280-processor-38-5m-cache-2-70-ghz.html>

- W praktyce – obliczenia nie wektorowe:
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeń
 - 4.0 GHz częstotliwość turbo dla aktywnych maksymalnie 2 rdzeni
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni
- W praktyce – obliczenia wektorowe **AVX-2.0 (256 bitów)**
 - 2.2 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 2.9 GHz częstotliwość turbo dla wszystkich rdzeni

Teoretyczna wydajność procesora:

częstotliwość procesora - **podsumowanie**

- Procesor Intel Xeon Platinum 8280 Processor (28 rdzeni):
 - Obliczenia nie wektorowe (skalarne):
 - 2.7 GHz częstotliwość bazowa dla wszystkich rdzeń
 - 3.3 GHz częstotliwość turbo dla aktywnych wszystkich rdzeni
 - Obliczenia wektorowe AVX-2.0 (256 bitów)
 - 2.2 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 2.9 GHz częstotliwość turbo dla wszystkich rdzeni
 - Obliczenia wektorowe AVX-512
 - 1.8 GHz częstotliwość bazowa dla wszystkich rdzeni
 - 2.4 GHz częstotliwość turbo dla wszystkich rdzeni

Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

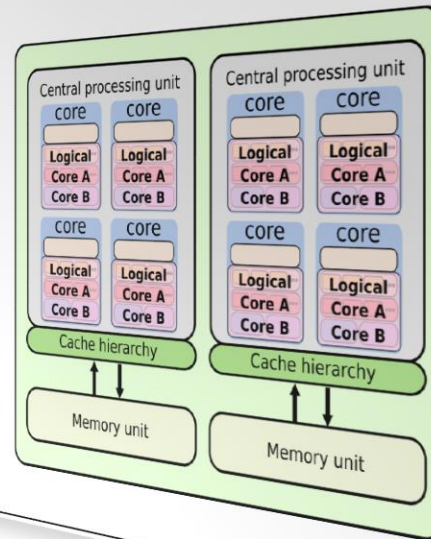
$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- v - liczba jednocześnie przetwarzanych elementów w wektorze
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na jednostkę

Teoretyczna wydajność procesora: liczba rdzeni

- SMT – sprzętowa wielowątkowość:
Wykład 4 A Slajd 19
- Intel Xeon Platinum 8280 Processor posiada:
 - 28 fizycznych rdzeni
 - 56 logicznych rdzeni

Rozwój procesorów ogólnego przeznaczenia

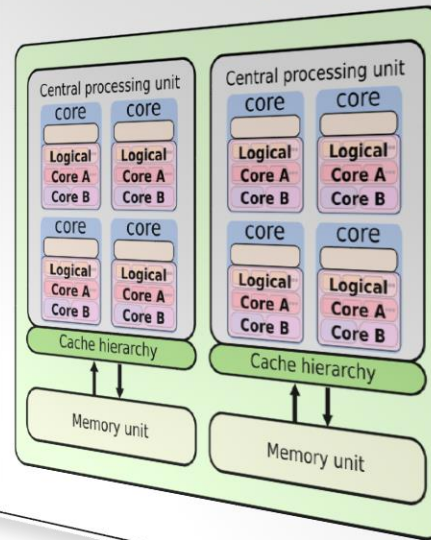


- Obecne procesory wspierają sprzętową wielowątkowość (Simultaneous multithreading, SMT)
- Oznacza to, że jeden fizyczny rdzeń traktowany jest jako kilka logicznych rdzeni (sprzętowe rdzenie),
- system operacyjny oraz użytkownik widzi dwa rdzenie, a w rzeczywistości jest tylko jeden
- np.. procesory firm Intel i AMD oferują procesory CPU wspierające dwa logiczne rdzenie w obrębie pojedynczego rdzenia fizycznego
- SMT ma na celu zwiększenie wydajności równoległe (współbieżnie) wykonywanych zadań
- Uwaga, SMT nie oznacza duplikowanie głównych zasobów rdzenie
- Pomimo że logiczne procesory pozwalają na jednoczesne wykonanie zadań, tylko jedno zadanie zużywa zasoby w danym cyklu procesora

Teoretyczna wydajność procesora: liczba rdzeni

- SMT – sprzętowa wielowątkowość:
Wykład 4 A Slajd 19
- Intel Xeon Platinum 8280 Processor posiada:
 - 28 fizycznych rdzeni
 - 56 logicznych rdzeni
- Z punktu widzenia teoretycznej wydajności interesuje nas liczba fizycznych rdzeni

Rozwój procesorów ogólnego przeznaczenia



- Obecne procesory wspierają sprzętową wielowątkowość (Simultaneous multithreading, SMT)
- Oznacza to, że jeden fizyczny rdzeń traktowany jest jako kilka logicznych rdzeni (sprzętowe rdzenie),
- system operacyjny oraz użytkownik widzi dwa rdzenie, a w rzeczywistości jest tylko jeden
- np.. procesory firm Intel i AMD oferują procesory CPU wspierające dwa logiczne rdzenie w obrębie pojedynczego rdzenia fizycznego
- SMT ma na celu zwiększenie wydajności równoległe (współbieżnie) wykonywanych zadań
- Uwaga, SMT nie oznacza duplikowanie głównych zasobów rdzenia
- Pomimo że logiczne procesory pozwalają na jednoczesne wykonanie zadań, tylko jedno zadanie zużywa zasoby w danym cyklu procesora

Teoretyczna wydajność procesora

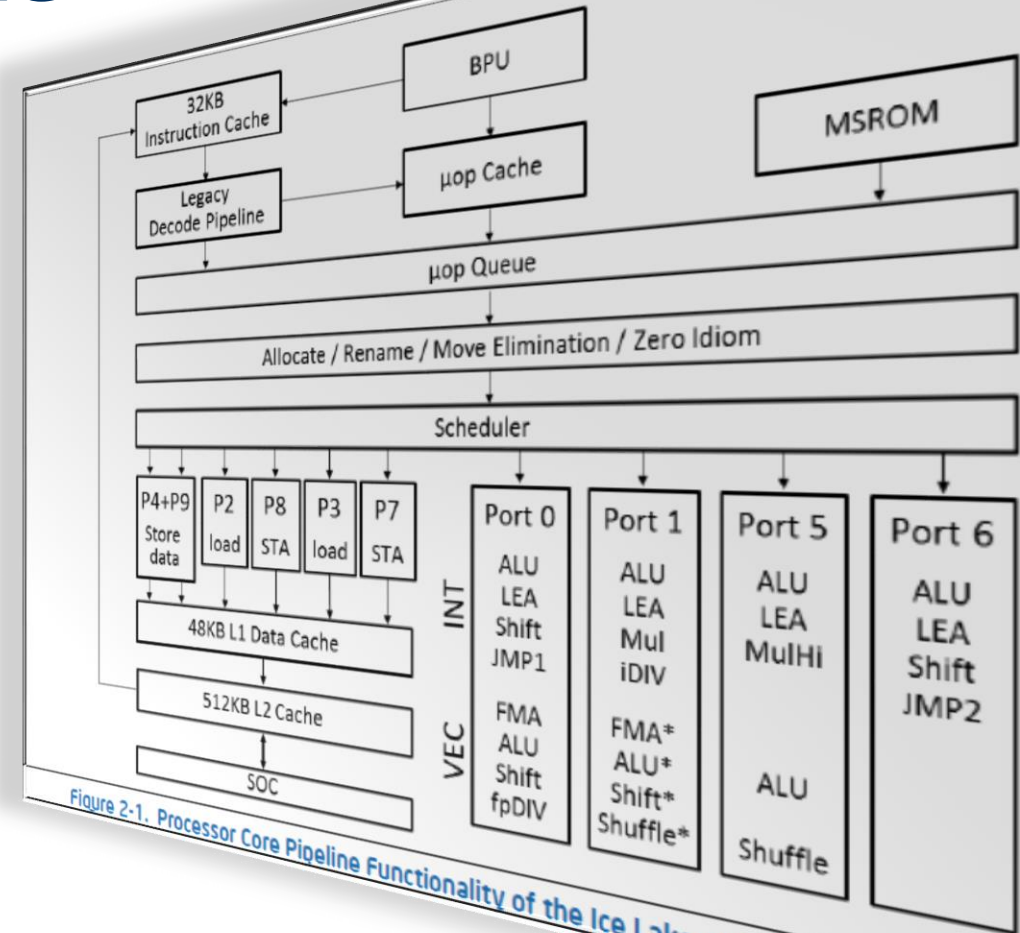
- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- **u - liczba jednostek wykonawczych realizujących daną operację**
- v - liczba jednocześnie przetwarzanych elementów w wektorze
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na jednostkę

Teoretyczna wydajność procesora: jednostki wykonawcze

- Liczba jednostek zależy od rodzaju operacji i architektury
- Przykładowo dla Intel Xeon Platinum 8280 operacja **dodawania**:
 - 2 jednostki dla FPU w sposób skalarny lub
 - 2 jednostki dla FPU w sposób wektorowy lub
 - 4 jednostki ALU w sposób skalarny lub
 - 2 jednostki ALU w sposób wektorowy
- Przykładowo dla Intel Xeon Platinum 8280 operacja **dzielenia**:
 - 1 jednostka dla FPU w sposób skalarny lub
 - 1 jednostki dla FPU w sposób wektorowy



Intel® 64 and IA-32
Architectures Optimization
Reference Manual

Teoretyczna wydajność procesora

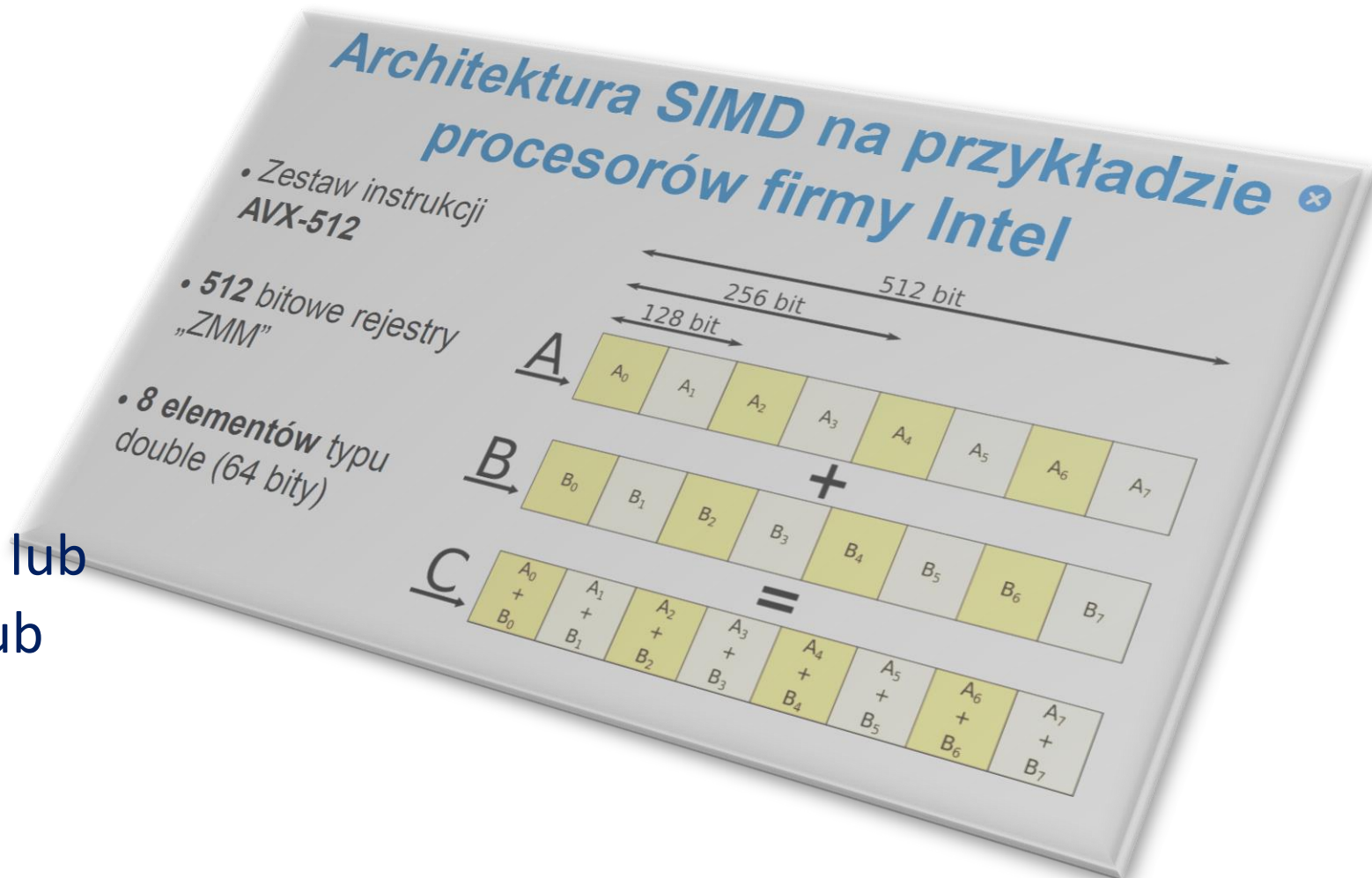
- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- **v - liczba jednocześnie przetwarzanych elementów w wektorze**
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na jednostkę

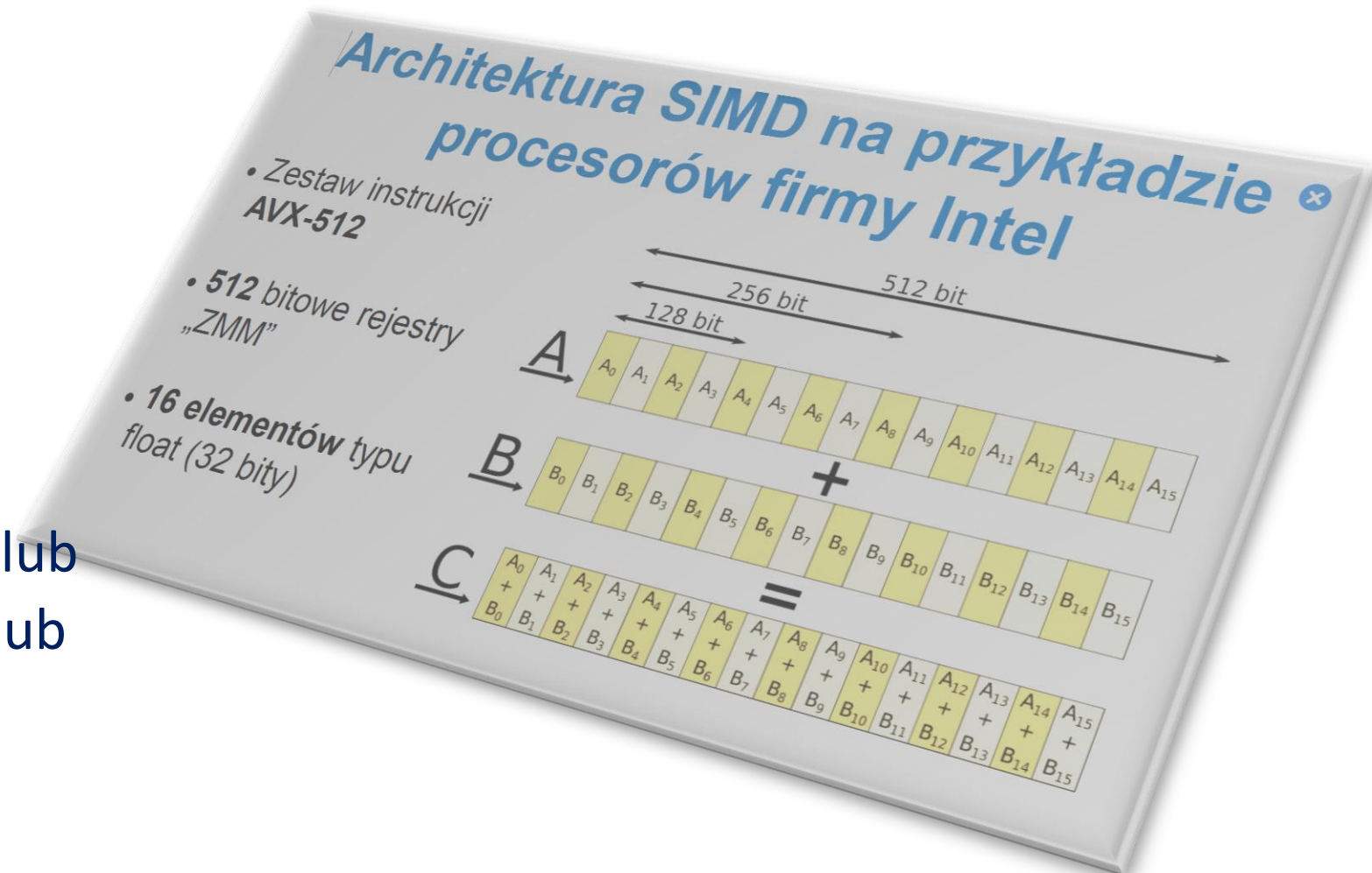
Teoretyczna wydajność procesora liczba elementów w wektorze

- Liczba elementów zależy od:
 - **Typu danych**
 - **Rozmiaru rejestrów**
- Wykład 4 B
slajdy 24-29
- Przykładowo dla Intel Xeon Platinum 8280 operacja:
 - **8 elementów typu double, lub**
 - 16 elementów typu float, lub
 - 16 elementów typu int



Teoretyczna wydajność procesora liczba elementów w wektorze

- Liczba elementów zależy od:
 - **Typu danych**
 - **Rozmiaru rejestrów**
- Wykład 4 B
slajdy 24-29
- Przykładowo dla Intel Xeon Platinum 8280 operacja:
 - 8 elementów typu double, lub
 - **16 elementów typu float**, lub
 - **16 elementów typu int**



Teoretyczna wydajność procesora

- Teoretyczną wydajność procesorów R_p można wyznaczyć zgodnie ze wzorem:

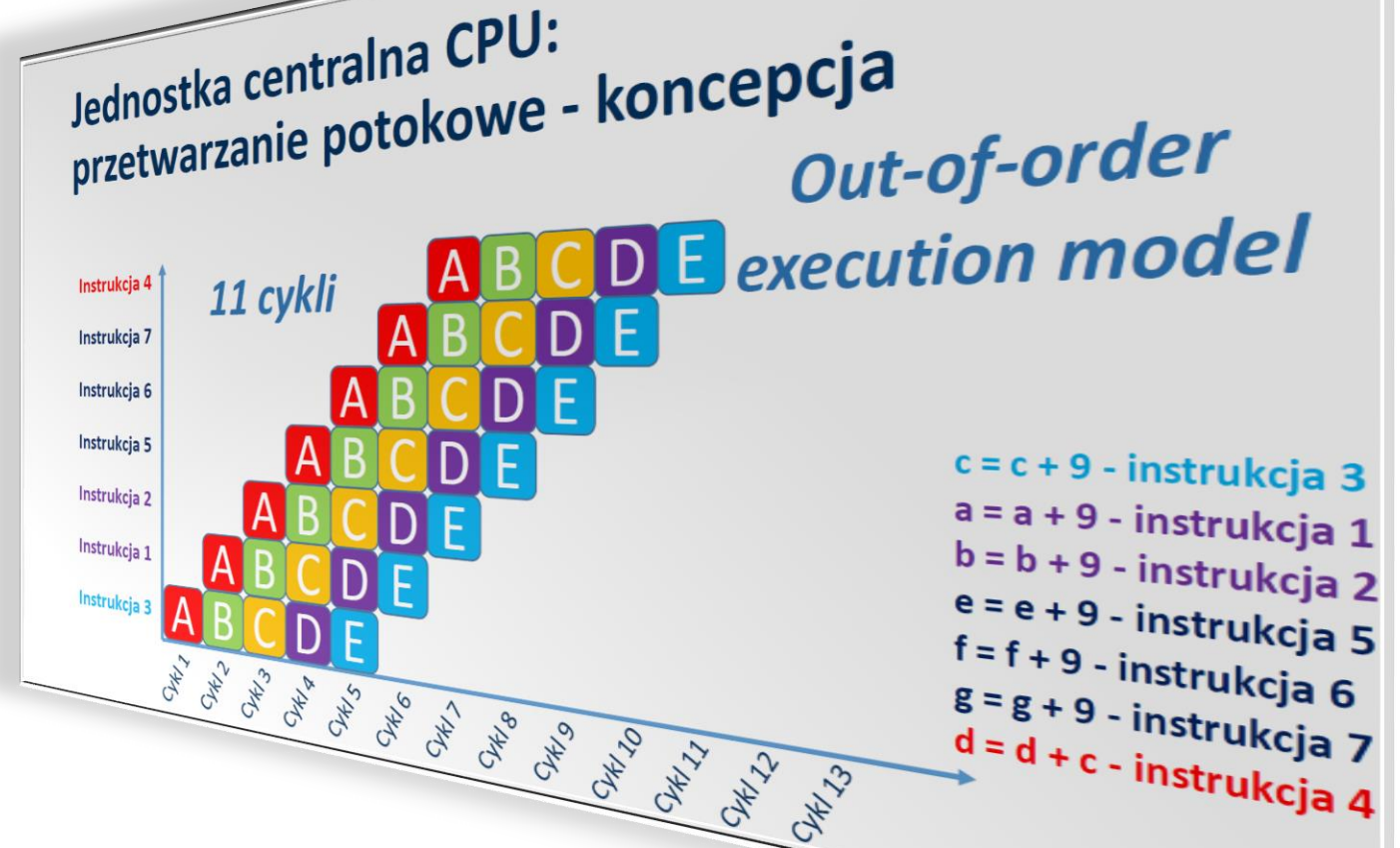
$$R_p = f * c * u * v * i$$

- f - częstotliwość taktowania procesora
- c - liczba rdzeni
- u - liczba jednostek wykonawczych realizujących daną operację
- v - liczba jednocześnie przetwarzanych elementów w wektorze
- i - „statystyczna” liczba operacji np. dodawania wykona w jednym cyklu zegarowym na jednostkę

Teoretyczna wydajność procesora

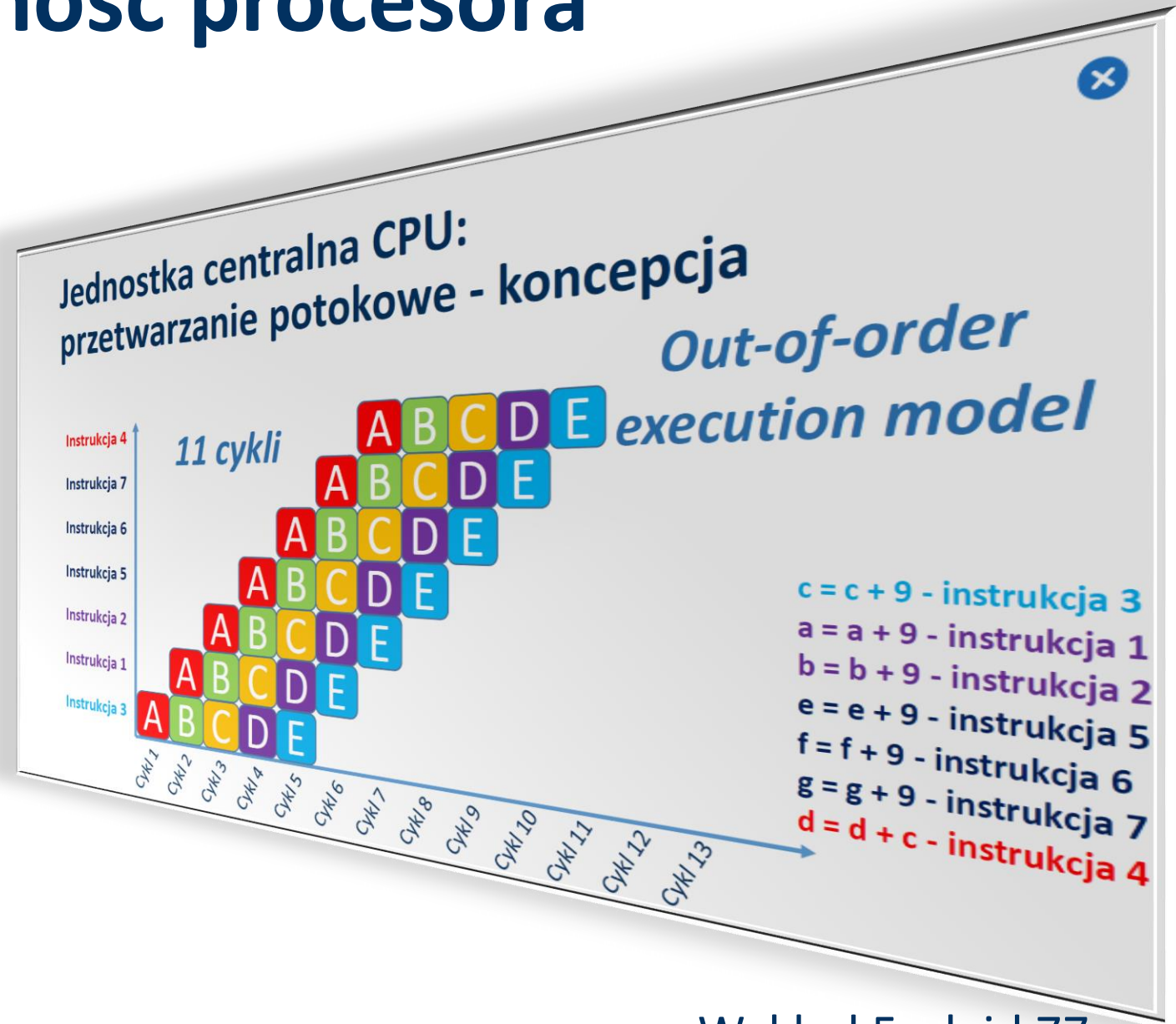
liczba operacji

- Ze względu na możliwość przetwarzanie potokowe w każdym cyklu procesora uruchamiamy/kończymy operację np. dodawania
- Zatem „statystyczna” liczba operacji np. dodawania lub mnożenia wynosi **jeden**



Teoretyczna wydajność procesora liczba operacji

- Jednakże, niektóre procesory mają jednostki wyspecjalizowane do wykonania operacji mnożenia i dodawania w pojedynczej instrukcji procesora FMA
- W tym przypadku „statystyczna” liczba operacji (instrukcja FMA) wynosi **dwa**
 - Jednostka procesora **FMA** wykonuje jednocześnie dwie operacje matematyczne
 - Uwaga, procesory mogą mieć kilka jednostek FMA



Teoretyczna wydajność procesora

podsumowanie

$$R_p = f * c * u * v * i$$

Wydajność dla Intel Xeon Platinum 8280:

- Opcja Turbo, wszystkie rdzenie, operacja dodawania (bez FMA), skalarnie (double lub float):

$$R_p = 3,3 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 1 \text{ skalarnie} * 1 \text{ operacja}$$

Teoretyczna wydajność procesora

podsumowanie

$$R_p = f * c * u * v * i$$

Wydajność dla Intel Xeon Platinum 8280:

- Opcja Turbo, wszystkie rdzenie, operacja dodawania (bez FMA), wektorowo AVX-512 (512 bitów) typ **double**:

$$R_p = 2,4 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 8 \text{ SIMD double} * 1 \text{ operacja}$$

- Opcja Turbo, wszystkie rdzenie, operacja dodawania (bez FMA), wektorowo AVX-512 (512 bitów) typ **float**:

$$R_p = 2,4 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 16 \text{ SIMD float} * 1 \text{ operacja}$$

Teoretyczna wydajność procesora

podsumowanie

$$R_p = f * c * u * v * i$$

Wydajność dla Intel Xeon Platinum 8280:

- Opcja Turbo, wszystkie rdzenie, operacja dodawania (bez FMA), wektorowo AVX-2.0 (256 bitów) typ **double**:

$$R_p = 2,9 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 4 \text{ SIMD double} * 1 \text{ operacja}$$

- Opcja Turbo, wszystkie rdzenie, operacja dodawania (bez FMA), wektorowo AVX-2.0 (256 bitów) typ **float**:

$$R_p = 2,9 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 8 \text{ SIMD float} * 1 \text{ operacja}$$

Teoretyczna wydajność procesora

podsumowanie

$$R_p = f * c * u * v * i$$

Wydajność dla Intel Xeon Platinum 8280:

- Opcja Turbo, wszystkie rdzenie, operacja dodawania i mnożenia (**FMA**), skalarnie (double lub float):

$$R_p = 3,3 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 1 \text{ skalarnie} * 2 \text{ operacje}$$

- Opcja Turbo, wszystkie rdzenie, operacja dodawania i mnożenia (**FMA**), wektorowo AVX-512 (512 bitów) typ double:

$$R_p = 2,4 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 8 \text{ SIMD double} * 2 \text{ operacje}$$

- Opcja Turbo, wszystkie rdzenie, operacja dodawania i mnożenia (**FMA**), wektorowo AVX-512 (512 bitów) typ float:

$$R_p = 2,4 \text{ GHz} * 28 \text{ rdzenie} * 2 \text{ jednostki FPU} * 16 \text{ SIMD float} * 2 \text{ operacje}$$

Teoretyczna wydajność procesora

podsumowanie - **pomoc**

$$R_p = f * c * u * v * i$$

- Liczba v jednocześnie przetwarzanych elementów w wektorze

Typ danych	Skalarnie	SSE (128 bity)	AVX (256 bity)	AVX-512 (512 bity)
float (32 bity)	1	4	8	16
double (64 bity)	1	2	4	8
int (32 bity)	1	4	8	16

- Liczba i operacji wykona w jednym cyklu zegarowym na jednostkę

Rodzaj operacji	Liczba operacji
FMA	2
no FMA	1

Wydajność procesora

- Wydajność procesora:
 - Teoretyczna – wydajność na jaką w teorii pozwala dana architektura komputerowa
 - **Rzeczywista – wydajność osiągnięta w praktyce dla konkretnej aplikacji**

Rzeczywista wydajność procesora

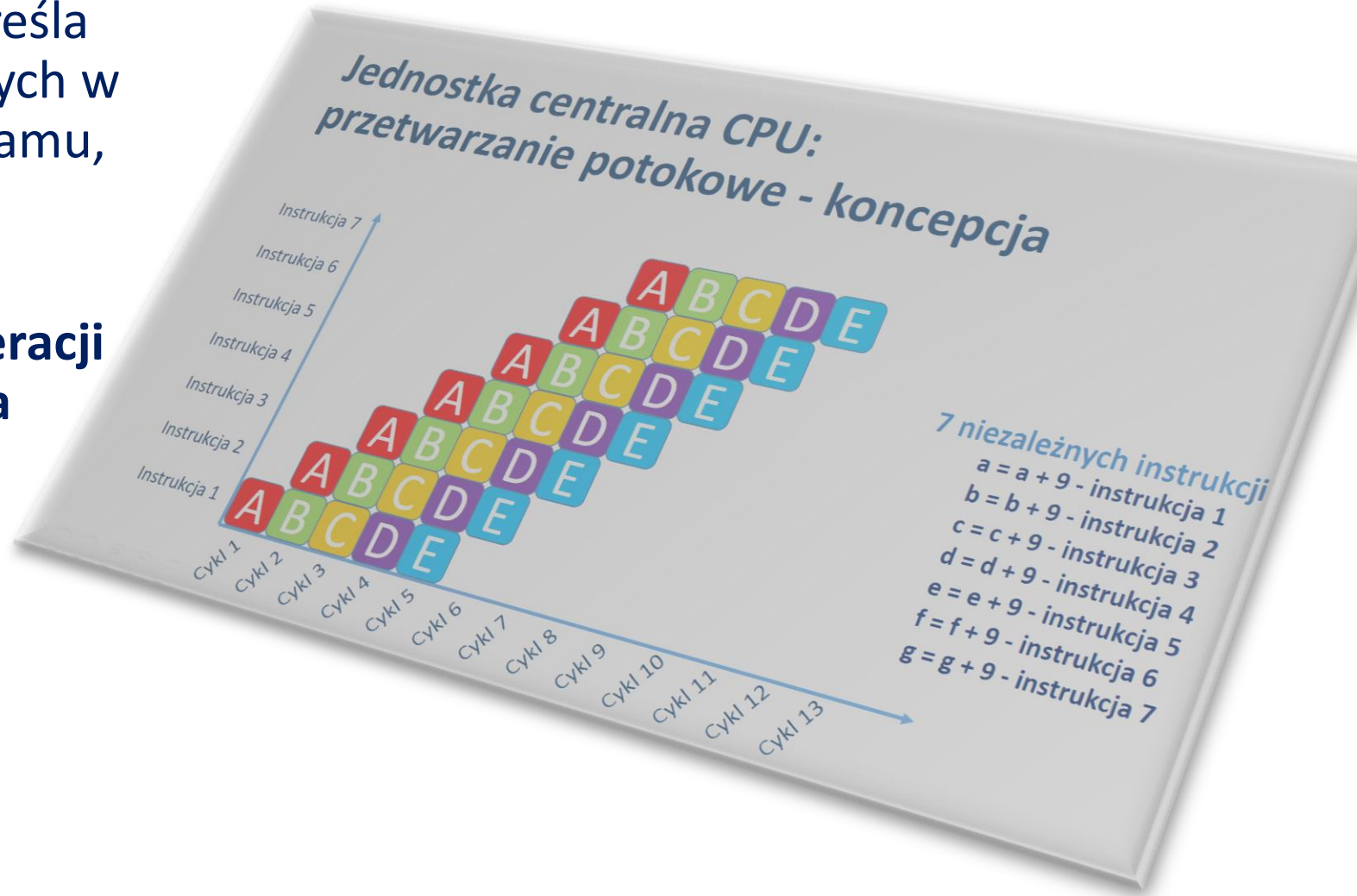
- Rzeczywista wydajność jest miarą określającą z jaką liczbą operacji zmiennoprzecinkowych na sekundę przetwarzana jest konkretna aplikacja
- Rzeczywistą wydajność procesorów R_m można wyznaczyć zgodnie ze wzorem:

$$R_m = \frac{l_o}{t_p}$$

- l_o - Liczba operacji w programie
- t_p – czas wykonania programu mierzony w sekundach

Rzeczywista wydajność procesora liczba operacji

- Liczba operacji zazwyczaj określa liczbę operacji matematycznych w wybranym fragmencie programu, dla którego wyznaczana jest rzeczywista wydajność
- W przykładzie widzimy **7 operacji matematycznych dodawania** zamienionych na 7 instrukcji dodawania procesora



Rzeczywista wydajność procesora

liczba operacji

- W poniższym przykładzie widzimy: **$n*2$** operacji matematycznych mnożenia zamienionych na **$n*2$** instrukcji mnożenia procesora

```
for(int i = 0; i<n; ++i) {  
    A[i] = a[i]* c[i] * b[i]  
}
```

Rzeczywista wydajność procesora

liczba operacji

- W poniższym przykładzie widzimy: **$n*2$** operacji matematycznych dodawania i mnożenia zamienionych
 - na **$n*2$** instrukcji dodawania i mnożenia procesora, lub

```
for(int i = 0; i<n; ++i) {  
    A[i] = a[i] * c[i] + b[i]  
}
```

Rzeczywista wydajność procesora

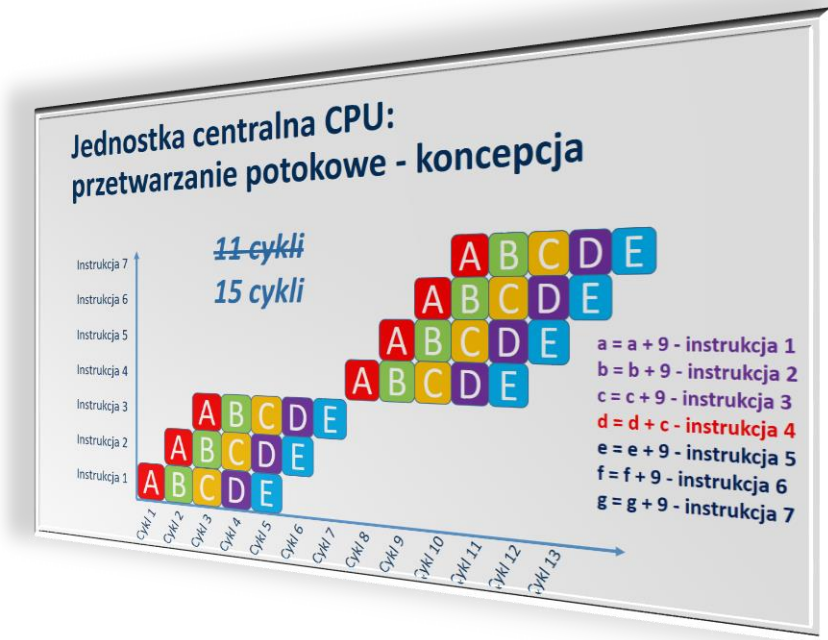
liczba operacji

- W poniższym przykładzie widzimy: $n*2$ operacji matematycznych dodawania i mnożenia zamienionych
 - na $n*2$ instrukcji dodawania i mnożenia procesora, lub
 - $n*2/2$ instrukcji FMA dodawania i mnożenia

```
for(int i = 0; i<n; ++i) {  
    A[i] = a[i] * c[i] + b[i]  
}
```

Rzeczywista wydajność procesora czas wykonania programu

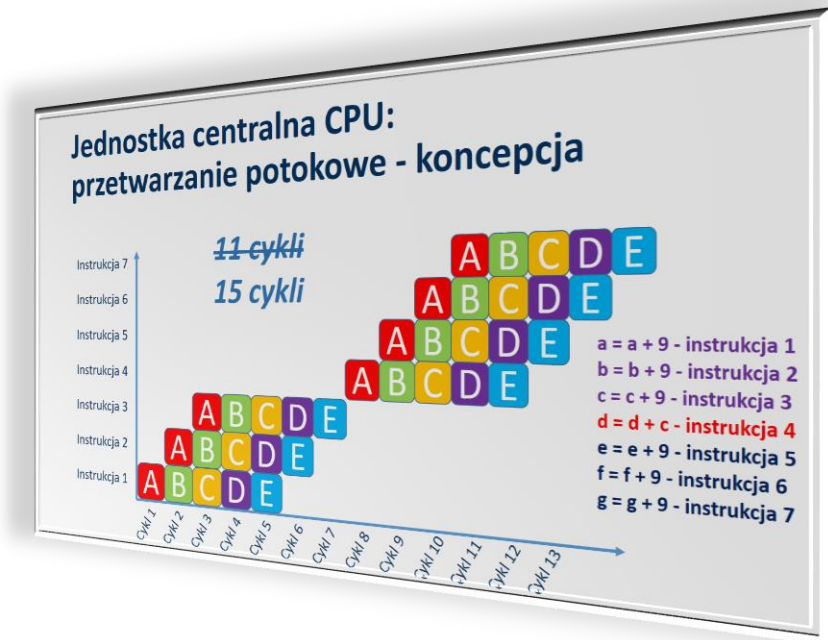
- Czas wykonania programu mierzony w sekundach można uzyskać na dwa sposoby:
 - Rzeczywisty pomiar czasu wykonania programu lub jego fragmentu, dla którego wyznaczana jest wydajność
 - Na podstawie liczby cykli i częstotliwość taktowania procesora, przykładowo:
 - Skoro procesor taktowany zegarem 2.0GHz przetwarza $2 * 10^9$ cykli w czasie 1 sekundy to 1 cykl trwa $t_{1cykl} = \frac{1}{2 * 10^9} [s]$



Rzeczywista wydajność procesora czas wykonania programu

- Czas wykonania programu mierzony w sekundach można uzyskać na dwa sposoby:
 - Rzeczywisty pomiar czasu wykonania programu lub jego fragmentu, dla którego wyznaczana jest wydajność
 - Na podstawie liczby cykli i częstotliwość taktowania procesora, przykładowo:
 - Skoro procesor taktowany zegarem 2.0GHz przetwarza $2 * 10^9$ cykli w czasie 1 sekundy to 1 cykl trwa $t_{1cykl} = \frac{1}{2 * 10^9} [s]$
 - Zatem czas programu t_p , którego liczba cykli $l_c=15$ wynosi: $t_p = t_{1cykl} * l_c$

$$t_p = \frac{1}{2 * 10^9} * 15[s] = 7.5 * 10^{-9}[s]$$



Rzeczywista wydajność procesora

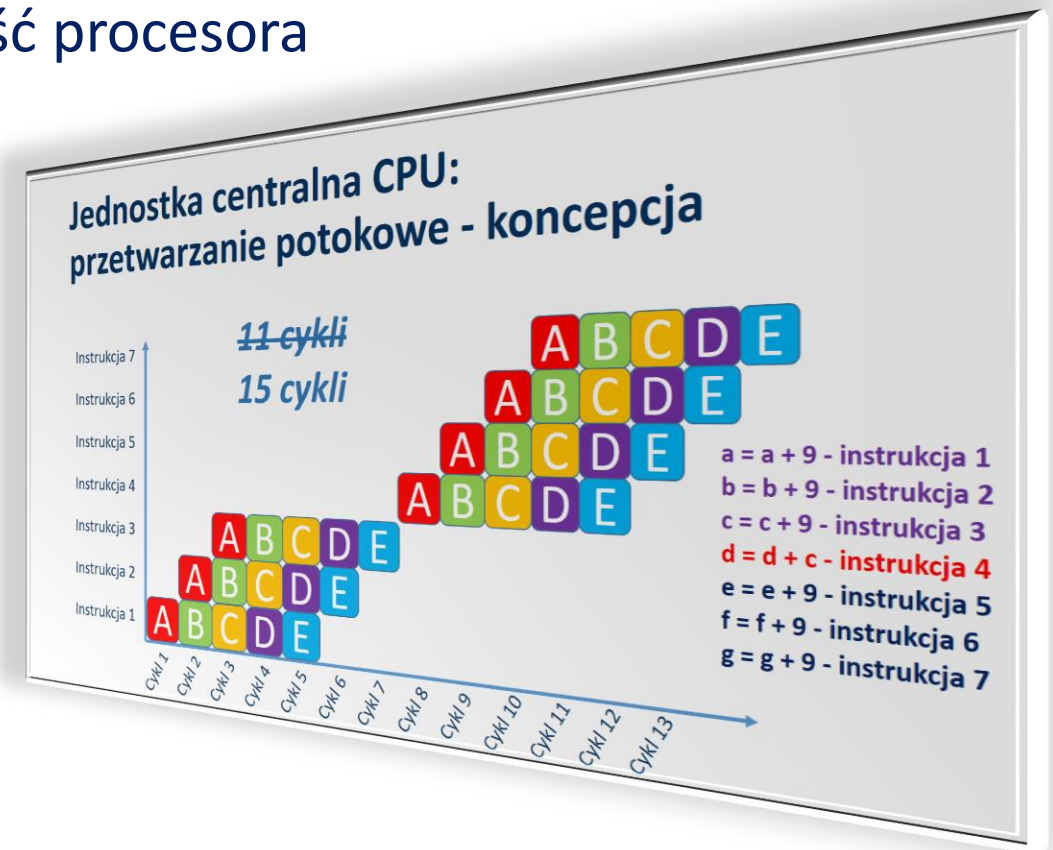
podsumowanie

- W poniższym przykładzie, program wykonał $l_o = 7$ operacji matematycznych w czasie $l_c = 15$ cykli uzyskując następującą rzeczywistą wydajność procesora taktowanego zegarem **2.0 GHz**:

$$t_{1cykl} = \frac{1}{2 * 10^9} [s]$$

$$t_p = t_{1cykl} * l_c = \frac{1}{2 * 10^9} * 15 [s] = 7.5 * 10^{-9} [s]$$

$$R_m = \frac{l_o}{t_p} = \frac{7}{7.5 * 10^{-9}} \left[\frac{flop}{s} \right] = 0.9333 \left[\frac{Gflop}{s} \right]$$



Dziękuję za uwagę



Kontakt:

dr hab. inż. Łukasz Szustak, prof. PCz

lszustak@icis.pcz.pl

Katedra Informatyki

Wydział Inżynierii Mechanicznej i Informatyki

Politechnika Częstochowska