# Sortowanie Polifazowe

Kamil Śliwiński

November 2024

## 1 Opis

Program został stworzony w celu sortowania w podanej kolejności rekordów reprezentujacych pary liczb rzeczywistych. Kryterium sortowania jest odległość punktu od poczatku układu współrzednych(0,0) obliczana wzorem:

$$d = \sqrt{x^2 + y^2} \tag{1}$$

W implementacji wykorzystano algorytm sortowania polifazowego przy użyciu trzech taśm, który działa w sposób efektywny na dużych zbiorach danych dzieki optymalnej dystrybucji poczatkowych rekordów na taśmy. Proces sortowania składa sie z etapów dzielenia pliku na serie i ich iteracyjnego łaczenia, aż do uzyskania jednego posortowanego pliku.

Zastosowanie trzech taśm umożliwia równoczesne przechowywanie serii oraz operacje odczytu i zapisu, co optymalizuje proces sortowania przy ograniczonej ilości dostepnej pamieci.

## 2 Struktura pliku z danymi (taśmy) – Format binarny

W eksperymencie dane wejściowe oraz dane tymczasowe (na taśmach) sa zapisywane w formacie binarnym, co zapewnia wieksza wydajność i szybkość działania programu oraz mniejsza zajetość pliku. Rekordy sa zapisywane bezpośrednio po sobie, bez żadnych dodatkowych danych. Na rekord składaja sie dwie liczby rzeczywiste(double), każda o rozmiarze 8 bajtów, a wiec pojedynczy rekord ma rozmiar 16 bajtów.

## 3 Wyjście programu

Po zakończeniu sortowania wyniki sa zapisywane do pliku wyjściowego "result.csv" aby można było sprawdzić poprawność algorytmu. Plik wyjściowy jest zapisany w formacie csv i zawiera rekordy posortowane według podanego klucza. Po zakończeniu sortowania 20 pierwszych rekordów również jest wyświetlane wraz z liczba faz i liczba operacji dyskowych.

# 4 Eksperyment

Celem eksperymenty było zauważanie jaka jest tendencja zmian liczby operacji dyskowych w zależności od liczby rekordów w pliku i wielkości strony oraz ich odzwierciedlenie w danych teoretycznych.

## 4.1 Opis eksperymentu

W pierwszej cześci eksperyment polegał na wielokrotnym sortowaniu losowo generowanego pliku przy zadanych parametrach wstepnych(wielkość współczynnika blokowania b i ilość rekordów do posortowania N).Dla zadanych parametrów każde sortownie zostało powtórzone 100 razy i ze zwracanych wyników(ilość faz i ilość operacji dyskowych) zostały policzone średnie.Druga cześć eksperymentu polegała na tym samym jednak przed testowaniem algorytmu zbór rekordów został wcześniej posortowany w kolejności odwrotnej.

## 4.2 Wzór na ilość faz

Liczba faz sortowania powinna rosnać logarytmicznie wzgledem ilości rekordów w pliku. Można to opisać wzorem:

$$p = 1,45 \times \log_2 (r) \tag{2}$$

Oczekiwana poczatkowa ilość serii można oszacować jako N/2, a wiec powyższe równanie przyjmuje postać:

$$p = 1,45 \times \log_2 \left( \frac{N}{2} \right) \tag{3}$$

Gdzie:

- $N$ - liczba rekordów,
- $p$ - liczba faz,
- $r$ - liczba serii,

## 4.3 Wzór na ilość operacji dyskowych

Liczba dostepów do dysku powinna w przybliżeniu wynosić:

$$\frac{(2 \times N \times \ p + 2 \times N)}{b} \tag{4}$$

Gdzie:

- $N$ - liczba rekordów,
- $b$ - współczynnik blokowania,

- $p$ - liczba faz,

W pojedynczej fazie nie sa przetwarzane wszystkie rekordy, każdy rekord jest przetwarzany około

$$1,04 \times \log_2 \left( \frac{N}{2} \right) \tag{5}$$

razy w trakcie całego procesu sortowania. Uwzgledniajac ten fakt, podstawiamy nowy wzór i otrzymujemy:

$$\frac{\left(2,08 \times N \times \log_2 \left(\frac{N}{2}\right) + 2 \times N\right)}{b} \tag{6}$$

## 4.4 Wyniki eksperymentu

### 4.4.1 Rekordy generowane losowo

| N | Ilość faz | | Liczba operacji | | Błąd wzgledny |
|---|---|---|---|---|---|
| | Pomierzona | Teoretyczna | Pomierzona | Teoretyczna | |
| 10 | 3 | 3,36 | 10 | 0,18 | 0,98 |
| 100 | 8 | 8,18 | 20 | 5,36 | 0,73 |
| 1000 | 13 | 13,0 | 90 | 80,65 | 0,11 |
| 10000 | 18 | 17,8 | 1124 | 1076,5 | 0,04 |
| 50000 | 21 | 21,18 | 6391 | 6325,79 | 0,01 |
| 100000 | 23 | 22,6 | 13859 | 13464,08 | 0,03 |
| 200000 | 24 | 24,08 | 28842 | 28553,16 | 0,01 |

Table 1: Tabela dla współczynnika blokowania b = 256

| N | Ilość faz | | Liczba operacji | | Błąd wzgledny |
|---|---|---|---|---|---|
| | Pomierzona | Teoretyczna | Pomierzona | Teoretyczna | |
| 10 | 3 | 3,36 | 10 | 0,13 | 0,99 |
| 100 | 8 | 8,18 | 20 | 2,68 | 0,87 |
| 1000 | 13 | 13,0 | 58 | 40,32 | 0,30 |
| 10000 | 18 | 17,8 | 575 | 538,25 | 0,06 |
| 50000 | 21 | 21,18 | 3205 | 3162,89 | 0,01 |
| 100000 | 23 | 22,6 | 6945 | 6732,04 | 0,03 |
| 200000 | 24 | 24,08 | 14433 | 14276,58 | 0,01 |

Table 2: Tabela dla współczynnika blokowania b = 512

### 4.4.2 Rekordy posortowane w odwrotnej kolejności

| N | Ilość faz | | Liczba operacji | | Błąd względny |
|---|---|---|---|---|---|
| | Pomierzona | Teoretyczna | Pomierzona | Teoretyczna | |
| 10 | 5 | 3,36 | 14 | 0,18 | 0,99 |
| 100 | 10 | 8,18 | 24 | 5,36 | 0,78 |
| 1000 | 15 | 13,0 | 102 | 80,65 | 0,21 |
| 10000 | 19 | 17,8 | 1188 | 1076,5 | 0,09 |
| 50000 | 23 | 21,18 | 6946 | 6325,79 | 0,09 |
| 100000 | 24 | 22,6 | 14431 | 13464,08 | 0,07 |
| 200000 | 26 | 24,08 | 31141 | 28553,16 | 0,08 |

Table 3: Tabela dla współczynnika blokowania b = 256

### 4.4.3 Wykresy pokazujace zależność liczby operacji od liczby wejściowych rekordów

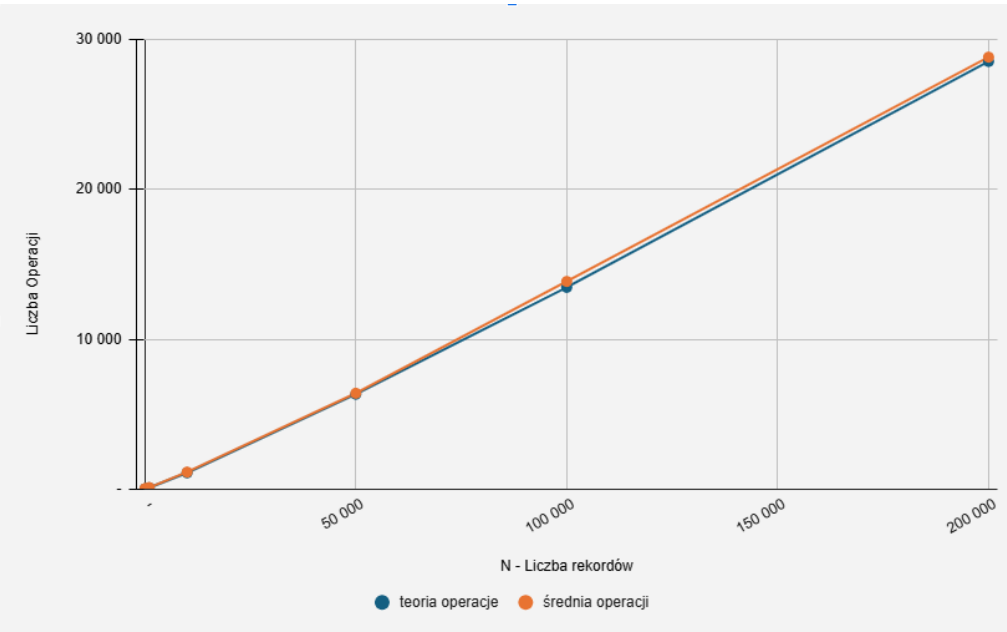


Figure 1: Wykres w skali liniowej na obu osiach dla losowych danych i b = 256

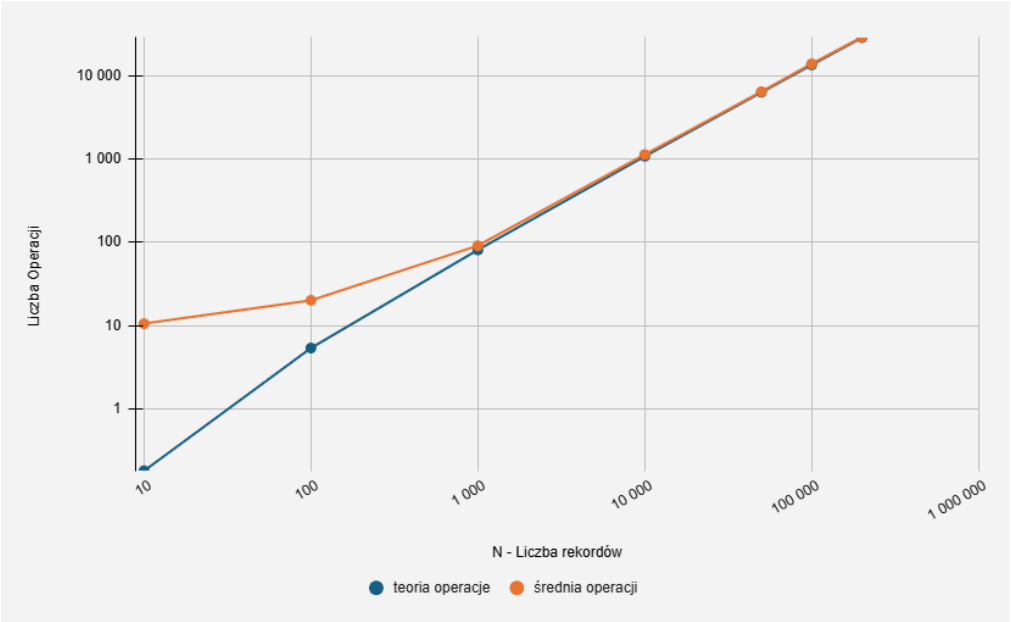


Figure 2: Wykres w skali logarytmicznej na obu osiach dla losowych danych i b = 256

## 4.5 Wnioski

Z eksperymentu wynika że wzór na oczekiwana liczbe faz jest w przecietnym przypadku bardzo dobrym przybliżeniem .

Natomiast wzór szacujacy ilość dostepów do dysku jest poprawny dla dużych wartości N. Dla małych N w porównaniu z współczynnikiem blokowania b ,bufor jest opróżniany jedynie pod koniec fazy bo nie jest nigdy zapełniony.Dlatego dla małych N wartości teoretyczne i zmierzone sa bardzo rozbieżne. Im wieksza wartość N tym faktyczna ilość operacji jest bardziej zbliżona do wyników teoretycznych co pokazuje kolumna bład wzgledy w powyższych tabelach.

Nieznacznie wieksza liczba operacji dyskowych w porównaniu z wartościami teoretycznymi jest spowodowana miedzy innymi tym bufor trzeba zapisać na koniec fazy na taśme na która scalamy rekordy trzeba zapisać zawartość bufora mimo że może być na nim mało wartości lub wcale. Podobnie jest z odczytem przy dystrybucji dane moga sie tak ułożyć że możemy odczytać kolejna strone dyskowa mimo tego że plik jest już pusty ale tego jeszcze nie wiemy.

Znaczenie ma też oszacowanie ilości serii opisywane jako r=N/2.Dla danych generowanych losowo przybliżenie jest właściwe ale przy pechowym losowaniu liczba operacji bedzie sie znaczaco różnić od wyników teoretycznych co widać w tabeli nr 3 ,która pokazuje wyniki dla pliku wstepnie odwrotnie posortowanego.Widać że wartości teoretyczne w sposób wyraźny różnia sie od zmierzonych i jest to spowodowane tym że w pliku odwrotnie posortowanym mamy

N serii czy 2 razy wiecej niż zakłada wzór.

Oba powyższe wykresy pokazuja że zarówno liczba rzeczywistych, jak i teoretycznych operacji I/O rośnie logarytmicznie wraz z liczba rekordów N. Wykres w skali logarytmicznej lepiej uwypukla różnice miedzy wynikami praktycznymi a teoretycznymi i potwierdza że wzór na teoretyczna liczbe operacji jest prawidłowy dla wiekszej liczby rekordów wejściowych.