Sprawozdanie Struktury Baz Danych Projekt 1: Scalanie polifazowe.

Radosław Piotrowicz 193251

1. Wprowadzenie:

Celem projektu było zaimplementowanie programu do sortowania plików sekwencyjnych, z metod sortowania były do wyboru: scalanie naturalne, scalanie przy użyciu wielkich buforów oraz scalanie polifazowe. W moim projekcie użyta została metoda polifazowa. Polega ona na podziale serii na taśmy według ciągu Fibonacciego, i następnym scalaniu k serii gdzie k jest długością krótszej serii. Scalanie w ten sposób redukuje liczbę faz. Metoda polifazowa ma też tą zaletę że faza dystrybucji występuję tylko raz.

2. Implementacja:

Program został zaimplementowany w języku C++.
Do sortowania użyte zostały 3 taśmy w schemacie 2+1, na które serie będą ładowane według ciągu Fibonacciego, w przypadku zbyt małej ilości serii dopełniam dłuższą taśmę seriami pustymi. Potencjalne problemy ze sklejającymi się seriami zostały rozwiązane przez zapamiętywanie końców serii.
Typem rekordu są ciągi znaków o długości 1-30 sortowane są one według porządku leksykograficznego, do porównanie używam funkcji wbudowanej. Rozmiar bufora wynosi 4 KB gdzie ostatnie 16 bajtów jest niewykorzystywane. Bufor jest wstanie pomieścić 136 rekordów.
Rekordy zapisane w pliku mają stałą długość (aby wyrównać długość dopisywane są znaki '\0', symulując niewykorzystane bajty), natomiast przy wczytywaniu pomijamy puste bajty, aby rekordy w logice programu były zmiennej długości. Taśmy są zrealizowane w postaci plików tekstowych gdzie w każdej linii znajduje się jeden rekord (w celu poprawienia czytelności).

Program posiada opcje wygenerowania nowej taśmy losowo, lub użytkownik może wprowadzić rekordy z klawiatury (natomiast wpisywanie ręczne jest zalecane dla małej liczby rekordów), po wygenerowaniu nowa taśma jest zapisywane do pliku. Program posiada też opcje wczytania danych z pliku testowego.

Program składa się z czterech klas:

- Sorter - warstwa sortowania, tutaj plik wejściowy jest rozdzielany na taśmy oraz przeprowadzane jest sortowanie, tutaj obliczane są też teoretyczne wartości ilości faz oraz liczby dostępów do pliku.
- App - zawiera menu, i tutaj wywołany był eksperyment oraz zapisane zostały jego wyniki.
- Tape - klasa będąca buforem dostarcza warstwie sortowania operacje odczytu i zapisu rekordu. Zapisuje też konice serii, oraz przy podawaniu kolejnego rekordu zapamiętuję poprzedni rekord, który warstwa sortowania wykorzystuje przy sortowaniu plików, podobnie jak końce serii. Przy zapisie / odczycie pliku następuje odwołanie do klasy FileManager.
- FileManager - ta klasa zapisuje i odczytuje do pliku po jednym bloku, pozwala na wygenerowanie pliku z losowymi rekordami. W tej klasie liczone są też dostępy do plików.

3. Eksperyment:

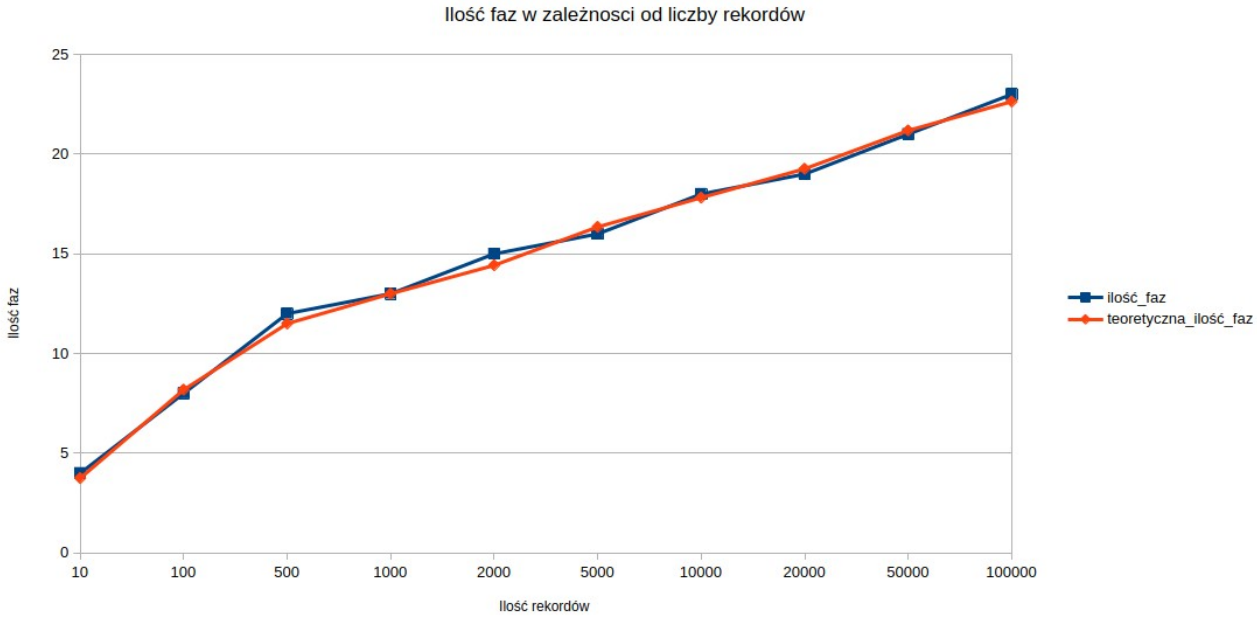
W eksperymencie porównana zostanie teoretyczna liczba faz oraz teoretyczna liczba dostępów do pliku z rzeczywistymi wartościami jakie zostały osiągnięte. Do eksperymentu wykorzystane zostanie 10 różnych taśmy o rosnących długościach. Wyniki teoretyczne zostały obliczone wewnątrz programu według wzorów podanych na wykładzie, a następnie zapisane do pliku. Wykresy zostały wygenerowane w arkuszu kalkulacyjnym, zawartym w pliku z projektem.

- Teoretyczna liczba fazy wynosi:
  1.45 * log2(r) - gdzie r to liczba serii
- Teoretyczna liczba dostępów do pliku wynosi
  2 * N * (1.04 * log2(r) + 1) / b - gdzie N to liczba rekordów, r to liczba serii a b to współczynnik blokowanie który wynosi 136 (4096 / 30 ~ 136).

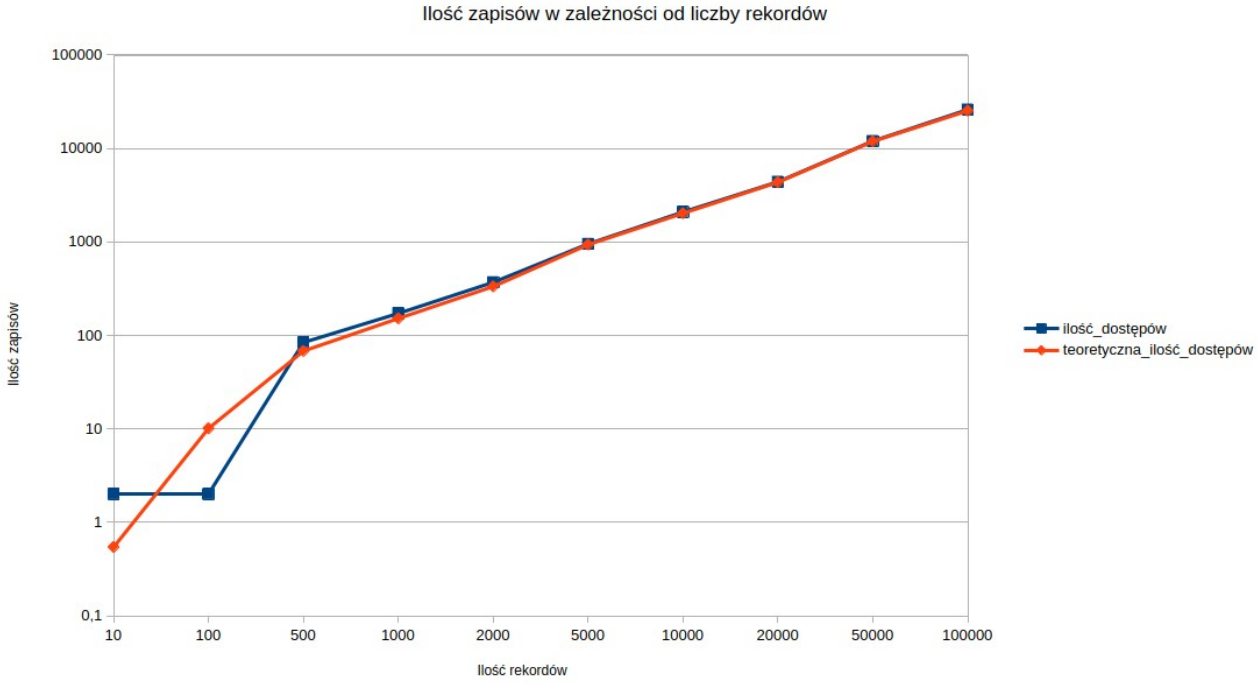
Obserwując wyniki można zauważyć że liczba faz jest prawie identyczna do teoretycznej liczby faz, natomiast dla teoretycznej liczby dostępów do plików i rzeczywistej ilości występują większe rozbieżności, zwłaszcza dla małej liczby rekordów. W przypadku gdy ilość rekordów jest mniejsza od wielkości bufora następuje tylko jeden zapis i jeden odczyt. Dla większej liczby rekordów stosunek teoretycznej a rzeczywistej liczby dostępów do pliku maleje. Nierówności mogą wynikać z nierównomiernego rozłożenia serii. Liczba serii też zgadza się z teorią i wynosi w okolicy N / 2. Można też dodać że dla dosyć sporego bufora (4 KB) liczba dostępów do pliku jest stosunkowo mała.

Ilość rekordów	Ilość serii	Oczekiwana liczba faz	Liczba faz	Oczekiwana liczba dostępów do pliku	Liczba dostępów do pliku
10	6	3.7482	4	0.542406	2
100	50	8.18359	8	10.1024	2
500	244	11.4996	12	67.9998	84
1000	500	13.0004	13	151.83	172
2000	988	14.4251	15	333.715	370
5000	2478	16.3487	16	935.732	953
10000	4999	17.8168	18	2026.31	2093
20000	9982	19.2634	19	4357.8	4397
50000	24994	21.1835	21	11907.1	12000
100000	50023	22.6349	23	25345.2	26068

Wykresy:



Na wykresie liczby faz widać że wartości są zbliżone.



Na wykresie ilości zapisów widać że dla małej liczby rekordów rozbieżności są większe.