Scalanie naturalne

1. Wstęp:

Celem projektu było napisanie programu sortującego plik metodą scalania naturalnego. Projekt został wykonany w języku C++. Zastosowano metodę scalania naturalnego. Taśmy są realizowane w formie plików dyskowych. Zaimplementowano symulację odczytu i zapisu blokowego, która pozwala na odczyt i zapis pojedynczego rekordu podczas sortowania. Program daje możliwość losowego generowania rekordów, wprowadzenia rekordów z klawiatury lub wczytania rekordów z pliku testowego. Zawartość pliku jest wyświetlana przed i po sortowaniu, jednak istnieje możliwość jej wyświetlenia po każdej fazie sortowania. Po zakończeniu sortowania wyświetlana jest liczba odczytów i zapisów stron na dysk, a także liczba faz.

2. Opis zastosowanej metody:

Do sortowania zastosowano metodę scalania naturalnego w schemacie 2+2 z użyciem czterech taśm. Na początku konieczna jest dystrybucja rekordów z taśmy wejściowej na dwie taśmy pomocnicze. W kolejnych fazach serie są scalane z dwóch taśm i od razu dystrybuowane naprzemiennie na dwie taśmy. W schemacie 2+2 dystrybucja i scalanie są dokonywane jednocześnie. Na każdy rekord w każdej fazie przypadają dwie operacje: jedna operacja odczytu i jedna zapisu.

3. Opis rekordu:

Rekordy w pliku składają się z sześciu parametrów: a_0, a_1, a_2, a_3, a_4, x . Sortowanie rekordów polega na uporządkowaniu ich według wartości funkcji $g(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4$. Na potrzeby projektu przyjęto, że parametry a_0, a_1, a_2, a_3, a_4, x są typu int a ich wartości są większe lub równe zero. W rezultacie rekord jest stałej wielkości 24. bajtów. W przypadku losowego generowania rekordów przyjęto liczby z zakresu 0-49, tak, aby w przypadku wylosowania maksymalnej wartości z tego zakresu dla wszystkich parametrów nie nastąpiło przepełnienie przy obliczaniu wartości funkcji. Wartość funkcji $g(x)$ ma typ unsigned long long.

4. Specyfikacja zapisu i odczytu blokowego:

Na potrzeby projektu przyjęto, że bufor/strona dyskowa służąca do zapisu i odczytu blokowego jest określany w liczbie rekordów, ponieważ rekordy są stałej wielkości. Współczynnik blokowania jest ustalany w kodzie jako stała. Zapis do pliku następuje po całkowitym wypełnieniu bufora lub jego części, tylko gdy liczba rekordów jest niewystarczająca do jego całkowitego wypełnienia. Odczyt rekordów następuje po całkowitym wypełnieniu bufora lub jego części w przypadku, gdy plik liczba rekordów jest niewystarczająca do jego całkowitego wypełnienia.

5. Specyfikacja plików dyskowych i pliku testowego:

Pliki dyskowe są realizowane w formie plików binarnych z rozszerzeniem .bin. Przyjęto, że plik testowy może mieć format pliku binarnego z rozszerzeniem .bin, lub format tekstowy z rozszerzeniem .txt. W przypadku pliku tekstowego każdy rekord zapisany jest w jednym wierszu, a pola rekordu oddzielone są spacją. Wybór rodzaju pliku testowego jest dokonywany poprzez menu. Nazwa pliku testowego może być nie dłuższa niż 20 znaków.

6. Opis sposobu prezentacji wyników działania programu:

Wybór opcji z menu i prezentacja wyników odbywa się za pomocą konsoli. Po wybraniu sposobu wprowadzenia rekordów (rekordy wygenerowane losowo i podane z klawiatury są zapisywane do pliku *input.bin*) i ew. nazwy pliku wejściowego, zostanie wyświetlony komunikat o możliwości wyświetlenia pliku po każdej fazie sortowania. Następnie zostanie wyświetlona zawartość pliku do posortowania i pliku po posortowaniu (w każdym wierszu parametry i wartość rekordu oddzielone spacją: a_0, a_1, a_2, a_3, a_4, x, $value$). W przypadku wyświetlania pliku po każdej fazie zostanie wyświetlona też zawartość taśm *file1* i *file2* po pierwszej dystrybucji a także taśm *file3* i *file4* po scaleniu. Na końcu wyświetlona zostanie nazwa pliku z posortowanymi rekordami, liczba odczytów i zapisów, a także liczba faz sortowania.

7. Eksperyment:

Celem eksperymentu jest zbadanie zależności między liczbą rekordów i wielkością strony dyskowej a liczbą faz i operacji dyskowych. Do eksperymentu zostaną użyte pliki dyskowe o różnej liczbie wygenerowanych losowo rekordów i różnej wielkości stron dyskowych. Wyniki eksperymentu zostaną porównane z wynikami teoretycznymi. Do obliczeń posłużono się arkuszem kalkulacyjnym.

Oczekiwana liczba serii w losowym pliku o N elementach wynosi:

$$r = \frac{N}{2}$$

W pliku o N rekordach maksymalna liczba faz wyraża się wzorem:

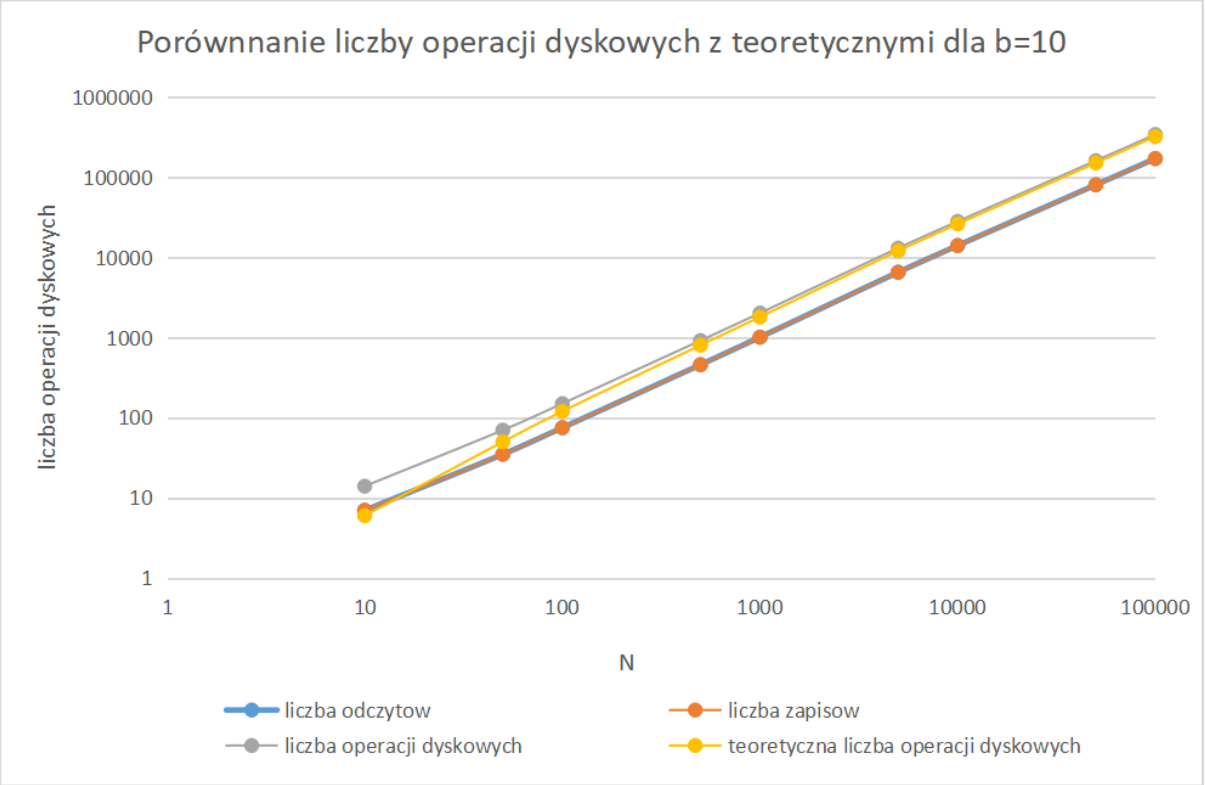
$$\lceil log_2 r \rceil$$

W schemacie 2+2 teoretyczna liczba operacji dyskowych wyrażana jest wzorem:

$$\frac{2N\lceil log_2 r \rceil}{b}, gdzie: \; b - współczynnik\ blokowania$$

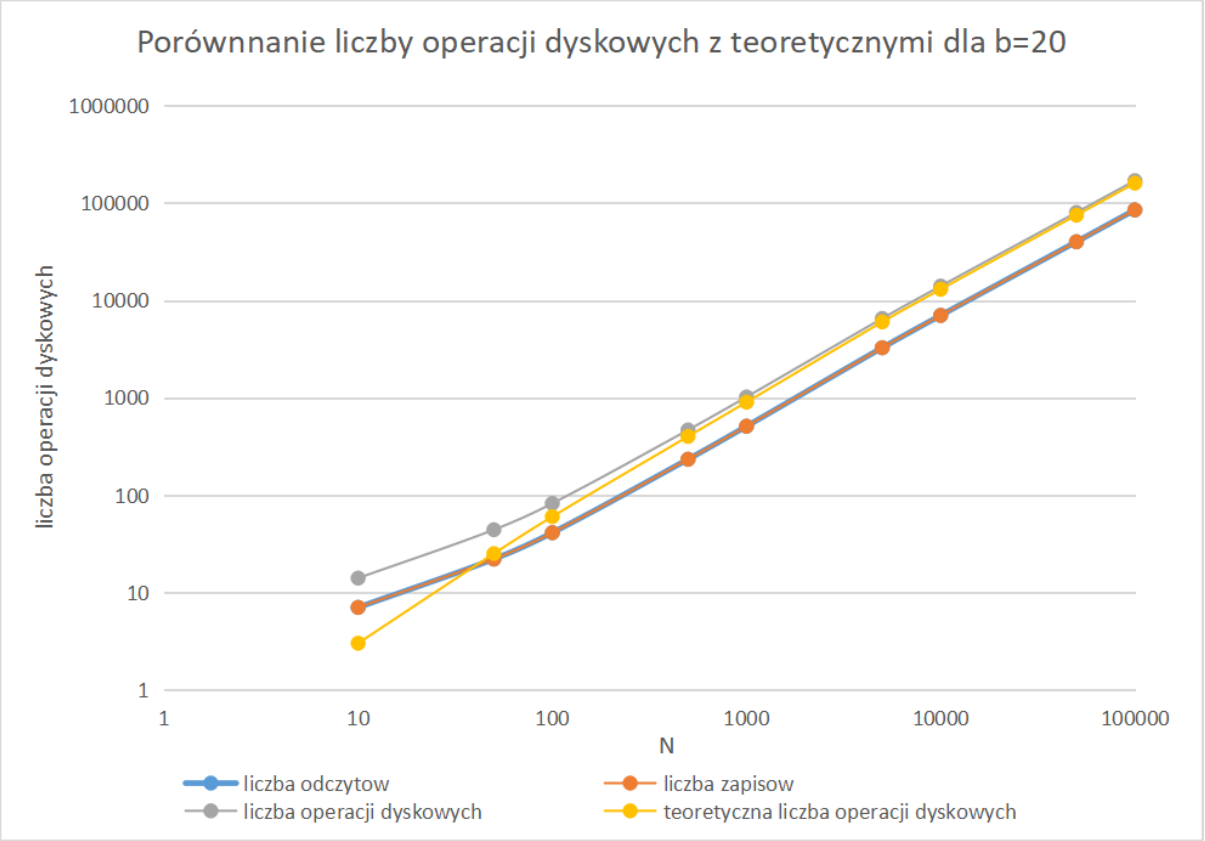
Wynik eksperymentu dla Nrekordów i współczynnika blokowania $b = 10$

N	liczba faz	liczba odczytów	liczba zapisów	liczba operacji dyskowych	teoretyczna liczba faz	teoretyczna liczba operacji dyskowych
10	3	7	7	15	3	6
50	5	35	35	71	5	50
100	6	75	75	151	6	120
500	8	458	458	917	8	800
1000	9	1009	1009	2019	9	1800
5000	12	6510	6510	13021	12	12000
10000	13	14011	14011	28023	13	26000
50000	15	80015	80015	160031	15	150000
100000	16	170013	170013	340027	16	320000



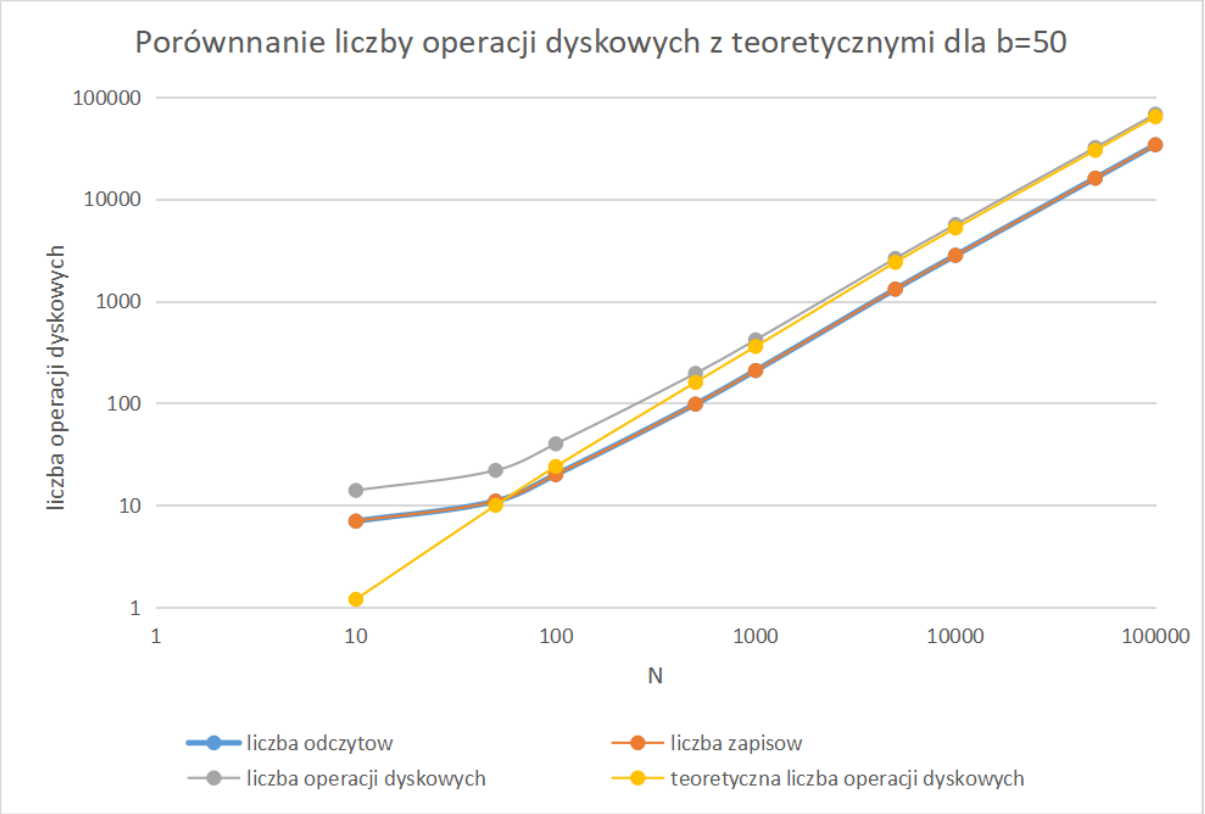
Wynik eksperymentu dla Nrekordów i współczynnika blokowania $b = 20$

N	liczba faz	liczba odczytów	liczba zapisów	liczba operacji dyskowych	teoretyczna liczba faz	teoretyczna liczba operacji dyskowych
10	3	7	7	15	3	3
50	5	22	22	45	5	25
100	6	41	41	83	6	60
500	8	233	233	467	8	400
1000	9	509	509	1019	9	900
5000	12	3261	3261	6523	12	6000
10000	13	7011	7011	14023	13	13000
50000	15	40015	40015	80031	15	75000
100000	16	85015	85015	170031	16	160000



Wynik eksperymentu dla Nrekordów i współczynnika blokowania $b = 50$

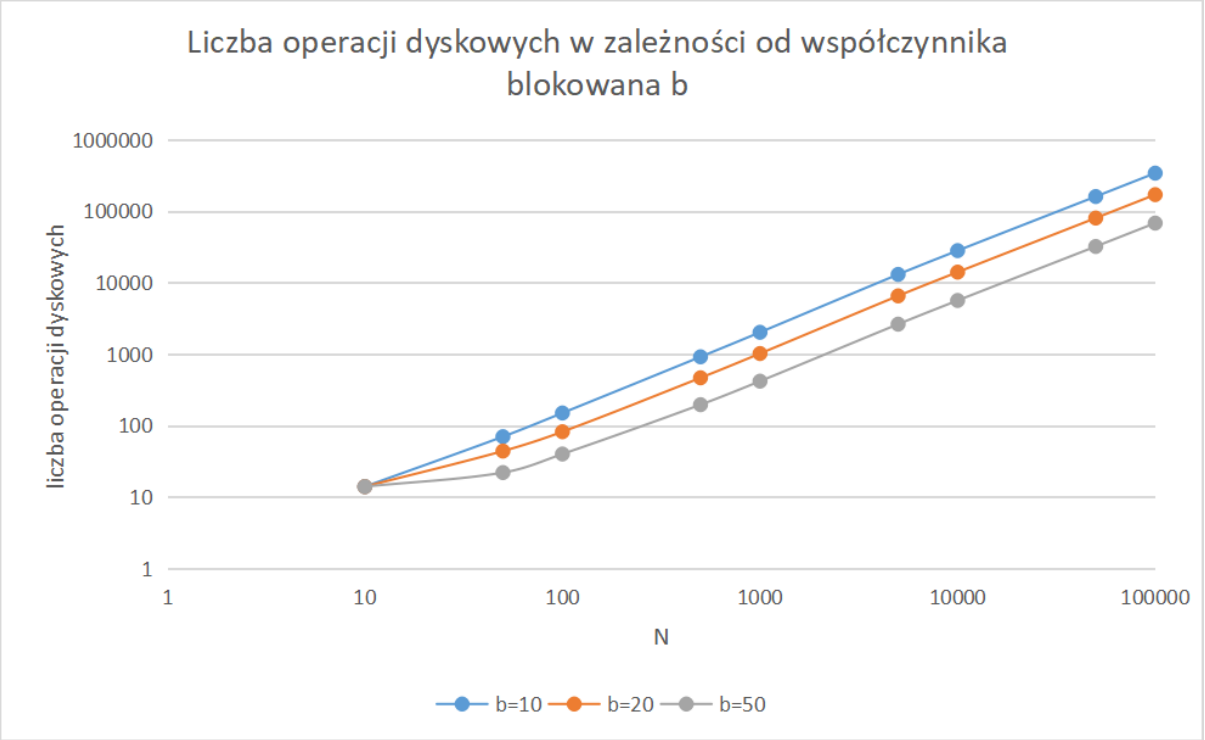
N	liczba faz	liczba odczytów	liczba zapisów	liczba operacji dyskowych	teoretyczna liczba faz	teoretyczna liczba operacji dyskowych
10	3	7	7	15	3	1,2
50	5	11	11	23	5	10
100	6	20	20	41	6	24
500	8	98	98	197	8	160
1000	9	209	209	419	9	360
5000	12	1311	1311	2623	12	2400
10000	13	2813	2813	5627	13	5200
50000	15	16015	16015	32031	15	30000
100000	16	34015	34015	68031	16	64000



Liczba faz wyznaczonych eksperymentalnie jest równa z wynikami teoretycznymi. Na liczbę faz ma wpływ sposób rozmieszczenia rekordów. W przypadku pliku posortowanego liczba faz ograniczyłaby się do jednej, dlatego liczba faz może być mniejsza od liczby wyznaczonej teoretycznie. Liczba faz może być mniejsza od maksymalnej liczby również z powodu łączenia się serii. Wyniki są zgodne z oczekiwaniami.

Liczba operacji dyskowych różni się od oczekiwanej. Wartość wyznaczona teoretycznie jest mniejsza. Powodem może być nierównomierny rozkład serii, różna długość serii, losowa dystrybucja rekordów i sposób implementacji. Wykresy prezentują jednak, że różnica między liczbą operacji dyskowych wyznaczoną eksperymentalnie a teoretyczną liczbą, staje się mniejsza wraz z zwiększającą się liczbą rekordów. Dla dużej liczby rekordów wykresy się pokrywają. Analizowane wielkości są logarytmiczne i przedstawione są na skali logarytmicznej. Liczba operacji dyskowych rośnie liniowo względem liczby rekordów.

Wraz z większą liczbą rekordów, liczba operacji rośnie. Liczba zapisów stron dyskowych jest równa liczbie odczytów, o czym świadczą pokrywające się wykresy. Dzięki operacjom na rozmiarze i aktualnej pozycji w pliku, uniknięto dodatkowych odczytów.



Można zaobserwować, że liczba operacji dyskowych maleje wraz ze wzrostem wielkości strony dyskowej. Wyjaśnieniem tego zjawiska jest rzadszy dostęp do pliku spowodowany większym buforem.