Лекція 13.
Зовнішнє сортування

§ 1. Постановка задачі зовнішнього сортування

Зовнішнім сортуванням називають сортування послідовних файлів, які розташовані в зовнішній пам’яті і занадто великі, для того, щоб можливо було повністю перемістити їх в основну пам’ять і застосувати один з методів внутрішнього сортування.

Найчастіше зовнішнє сортування використовується в системах управління базами даних при виконанні запитів, і від ефективності методів, які застосовуються, суттєво залежить продуктивність СУБД.

Послідовні файли – це файли, які можна читати запис за записом в послідовному режимі, а писати можна тільки після останнього запису.

§ 2. Пряме злиття

Нехай є послідовний файл A, який містить записи a_1, a_2...a_n (для спрощення, покладемо *n* дорівнює степінь числа 2). Вважатимемо, що кожний запис складається рівно з одного елемента, що є ключем сортування. Для сортування використовуються два допоміжних файли B і C розміром *n*/2.

Сортування складається з послідовних кроків, в кожному з яких виконується розподіл вмісту файлу A в файли B і C, а потім злиття файлів B і C у файл A.

Для виконання зовнішнього сортування методом прямого злиття в основній пам’яті необхідно розташувати лише дві змінні – для розміщення записів з файлів B і C. Файли A, B, C будуть O(logn) раз прочитати і стільки ж раз записані.

На першому кроці для розподілу послідовно зчитується файл А і записи a_1, a_3 ...a_{n-1} записуються а файл В, а записи a_2, a_4 ...a_n у файл С (початковий розподіл). Початкове злиття виконується над парами (a_1, a_2), (a_3, a_4), ...(a_{n-1}, a_n), і результат записується в файл А.

На другому кроці знову послідовно зчитується файл А, і в файл В записуються послідовні пари з непарними номерами, а в файл С – з парними. При злитті утворюються і записуються в файл А впорядковані четвірки записів. І т. д. Перед виконанням останнього кроку файл А буде містити дві впорядковані послідовності розміром n/2 кожна.

При розподілі перша з них потрапить в файл В, а друга – в файл С. Після злиття файл А буде містити повністю впорядковану послідовність записів.

Приклад. Відсортувати файл A(8, 23, 5, 65, 44, 33, 1, 6), використовуючи зовнішнє сортування прямим злиттям.

Наведемо розв’язання в таблиці.

Початковий стан файлу A	8, 23, 5,65, 44, 33, 1, 6
Крок 1. Розподіл	
Файл B	8 5 44 1 (a_1, a_3, a_5, a_7)
Файл C	23 65 33 6 (a_2, a_4, a_6, a_8)
Файл A	(a_1, a_2) (a_3, a_4) (a_5, a_6) (a_7, a_8)
	8, 23, 5, 65, 33, 44, 1, 6

Крок 2. Розподіл Файл B Файл C Файл A	 8 23 33 44 (a_1, a_3) 5 65 1 6 (a_2, a_4) (a_1, a_2) (a_3, a_4) 5 8 23 65 1 6 33 44
Крок 3. Розподіл Файл B Файл C Файл A	 5 8 23 65 (a_1) 1 6 33 44 (a_2) 1 5 6 8 23 33 44 65

При використанні методу прямого злиття не враховується часткове відсортування файлу.

§ 3. Природне злиття

Метод природного злиття полягає в розпізнаванні серій при розподілі і їх використання при послідовному злитті.

Серією називається послідовність записів a_1, a_2,.. a_i, a_j, така що $a_i < a_{i+1}$ і $a_j > a_{j-1}$, кінець серії $a_j > a_{j+1}$.

Файл – 3, 7, 5, 15, 3, 6, 9, 0.

Серія 1 – 3, 7.

Серія 2 – 5, 15.

Серія 3 – 3, 6, 9.

Серія 4 – 0.

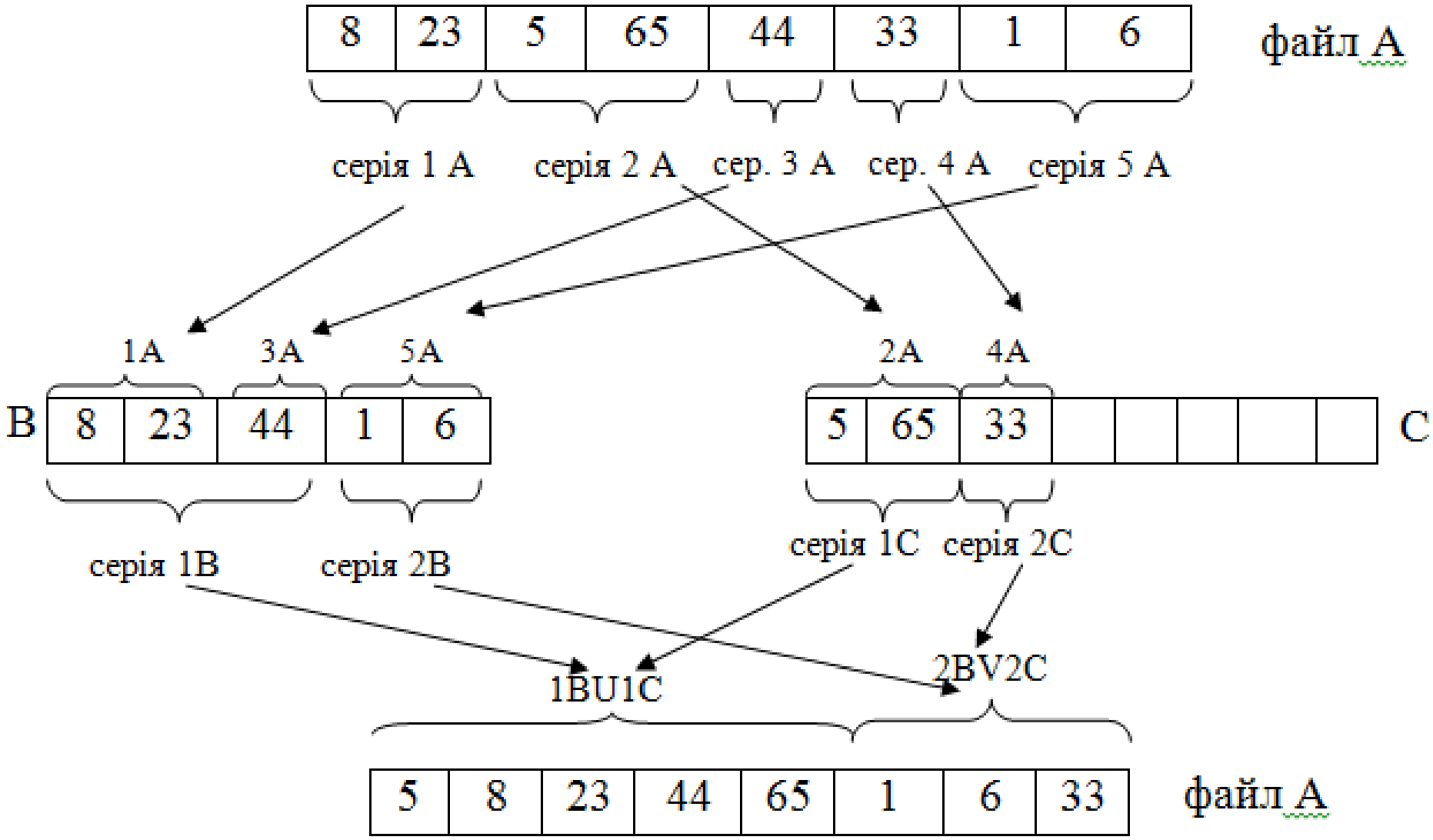
Сортування виконується за декілька кроків, в кожному з яких спочатку виконується розподілення файлу A по файлах B і C, а потім злиття B і C в файл A.

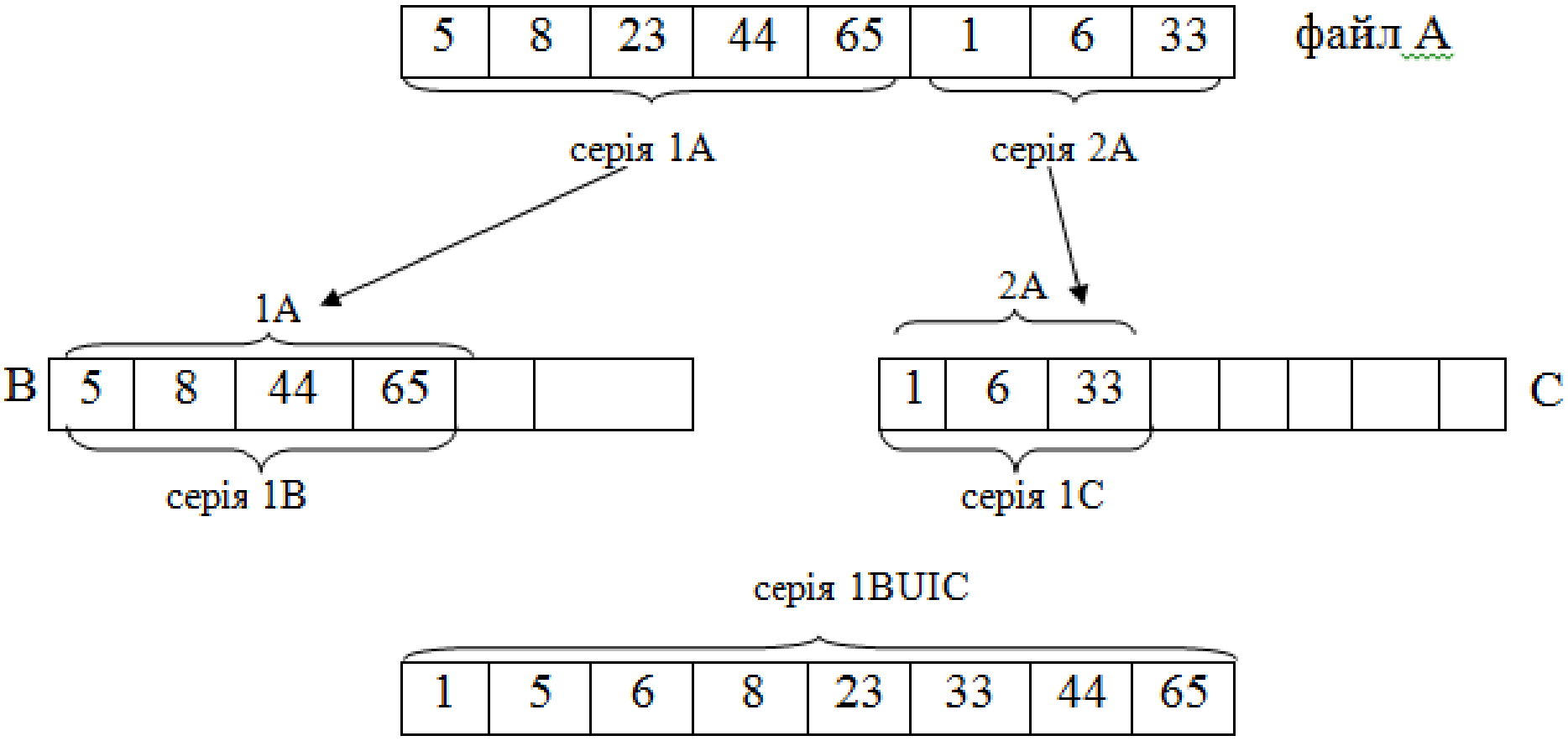
При розподілі розпізнається перша серія записів і переписується в файл B, друга – в файл C і т. д. При злитті перша серія записів файлу B зливається з першою серією файлу C, друга серія B з другою серією C і т. д. Якщо перегляд одного файлу закінчується раніше ніж перегляд другого (з причини різної кількості серій), то залишок недопереглянутого файлу цілком переноситься в кінець файлу A.

Процес завершується коли в файлі A залишається лише одна серія.

Оскільки довжина серій може бути довільною, то максимальний розмір файлів B і C може бути близьким до розміру файлу A.

Приклад. Відсортувати файл A(8, 23, 5, 65, 44, 33, 1, 6), використовуючи зовнішнє сортування природним

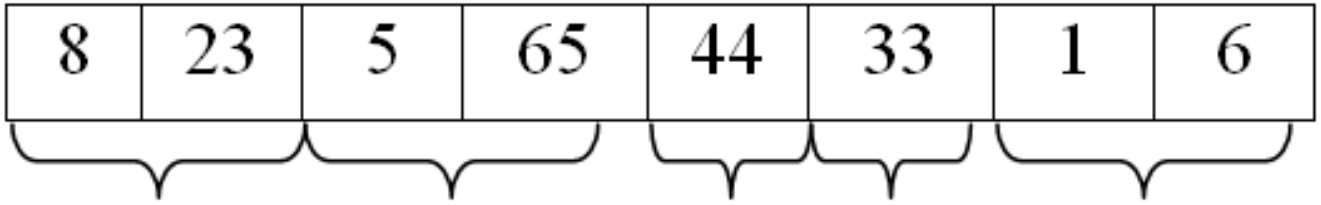


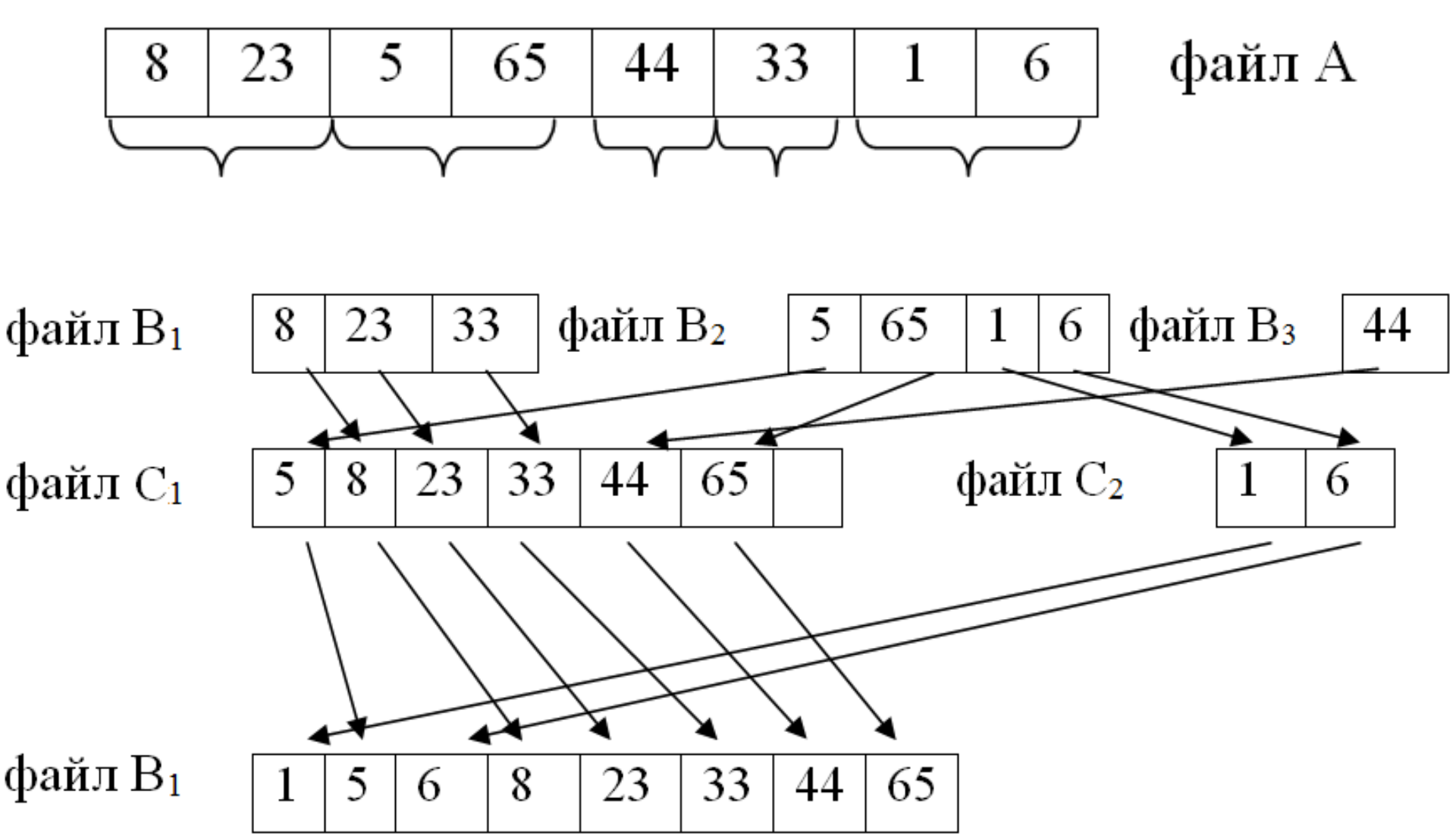

5 8 23 44 65 1 6 33
файл A
серія 1A
серія 2A
1A
2A
B
5 8 44 65
1 6 33
C
серія 1B
серія 1C
серія 1BUIC
1 5 6 8 23 33 44 65

§ 4. Збалансоване багатоканальне злиття

В основі методу зовнішнього сортування збалансованим багатоканальним злиттям є розподіл серій вхідного файлу по *m* допоміжних файлах B_1, B_2, ..., B*m* і їх злиття в *m* допоміжних файлів C_1, C_2, ..., C*m*. На наступному кроці відбувається злиття файлів C_1, C_2, ..., C*m* в файли B_1, B_2, ..., B*m* і т. д., доки в B_1 або C_1 не утвориться одна серія.

Приклад. Відсортувати файл A(8, 23, 5, 65, 44, 33, 1, 6), використовуючи трьохканальне злиття.




8 23 5 65 44 33 1 6
файл A
файл B1
8 23 33
файл B2
5 65 1 6
файл B3
44
файл C1
5 8 23 33 44 65
файл C2
1 6
файл B1
1 5 6 8 23 33 44 65

§ 4. Багатофазне сортування

Ідея багатофазного сортування полягає в тому, що з m допоміжних файлів (m-1) файл використовується для вводу послідовностей, що зливаються, а один для виводу утворених серій.

Як тільки один з файлів вводу стає порожнім, його починають використовувати для виводу серій, отриманих при злитті серій нового набору (m-1) файлів.

Таким чином, маємо перший крок, при якому серії початкового файлу A розподіляються по m-1 допоміжних файлах, а потім виконується багатофазне злиття серій з (m-1) файлів, доки в одному з них не утвориться одна серія.

Крок	B_1	B_2	B_3
n	1	0	0
n-1	0	1	1
n-2	1	2	0
n-3	3	0	2
n-4	0	3	5
n-5	5	8	0
n-6	13	0	8
…	…	…	…
1	…	…	…

Цей приклад показує, що метод трьохфазного зовнішнього сортування дає бажаний результат і працює максимально ефективно (на кожному етапі відбувається злиття максимальної кількості серій), якщо початковий розподіл серій між допоміжними файлами описується сусідніми числами Фібоначчі.

Приклад. Відсортувати файл A(8, 23, 5, 65, 44, 33, 1, 6), використовуючи багатофазне злиття.

Крок	B_1	B_2	B_3
1	{8, 23} {5, 65} {44}	{33} {1, 6}	0
2	{44}	0	{8, 23, 33} {1, 5, 6, 65}
3	0	{8, 23, 33, 44}	{1, 5, 6, 65}
4	{1, 5,6,8, 23, 33, 44,65}	0	0