15장. 외부정렬 (보충자료)

외부정렬

- 내부정렬: 정렬하고자 하는 데이터를 주기억 장치 내에 기억시켜 놓고 재배열하여 순서화시킴

- 외부정렬: 데이터의 양이 많아서 주기억 장치에 모두 수용할 수 없을 때, 외부 보조 기억 장치를 이용하여 정렬
  - 정렬해야 하는 리스트는 디스크상에 있다고 가정하자
  - 디스크의 판독/기록에 영향을 미치는 세가지 요소
  (1) 탐색시간(seek time): 읽기/쓰기 헤드가 해당 실린더를 찾는데 걸리는 시간
  (2) 회전 지연 시간(rotation delay time): 읽기/쓰기 헤드가 트랙위의 해당 섹터에 올 때까지 걸리는 시간
  (3) 전송 시간(transmission time): 디스크와 주기억장치간에 데이터가 전송되는데 걸리는 시간

k-way 병합

- 합병정렬을 이용한 두 단계 처리

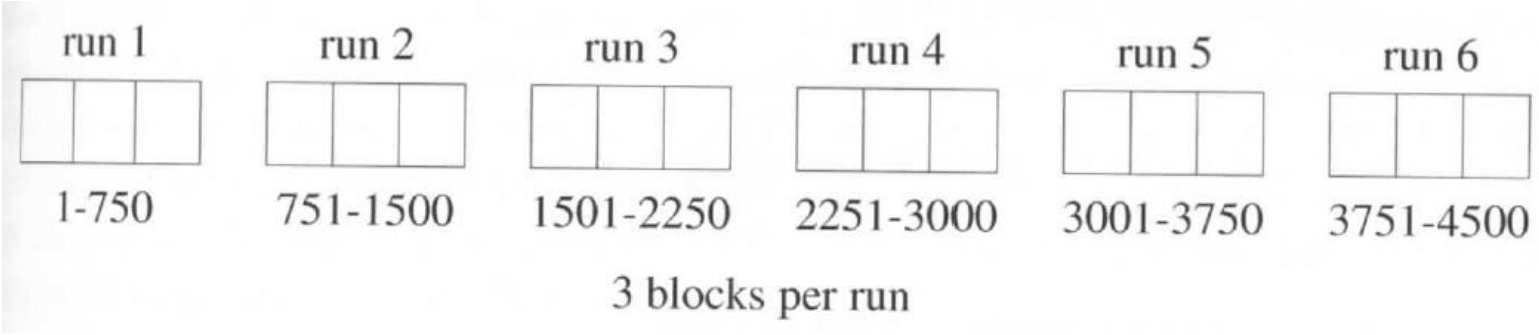
(1) 입력 리스트의 여러 세그먼트들을 좋은 내부정렬방법으로 정렬. 이 정렬된 세그먼트들을 '런(run)'또는 '서브파일'이라고 하는데, 런이 생성되면 외부 저장 장치에 기록

(2) 첫단계에서 만들어진 런들을 하나의 런이 될 때까지 합병한다

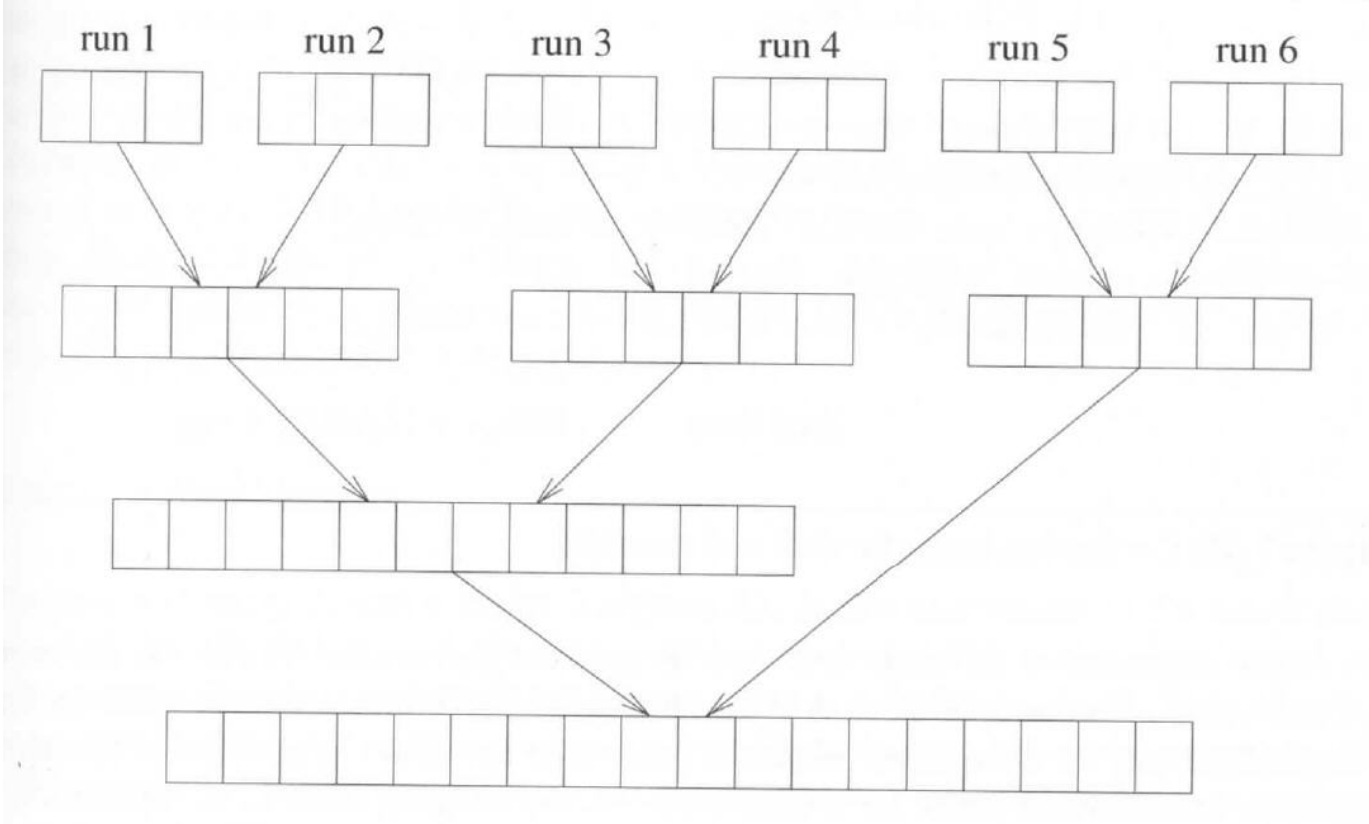
- 블럭(block): 한 번에 디스크로부터 읽거나 쓸 수 있는 데이터의 단위

최대 750개의 레코드만 정렬할 수 있는 내부메모리를 가진 컴퓨터를 이용하여 4500개의 레코드로 된 리스트를 정렬할 때. 입력리스트는 디스크에 저장되어 있고 250 레코드의 블럭 길이를 가진다

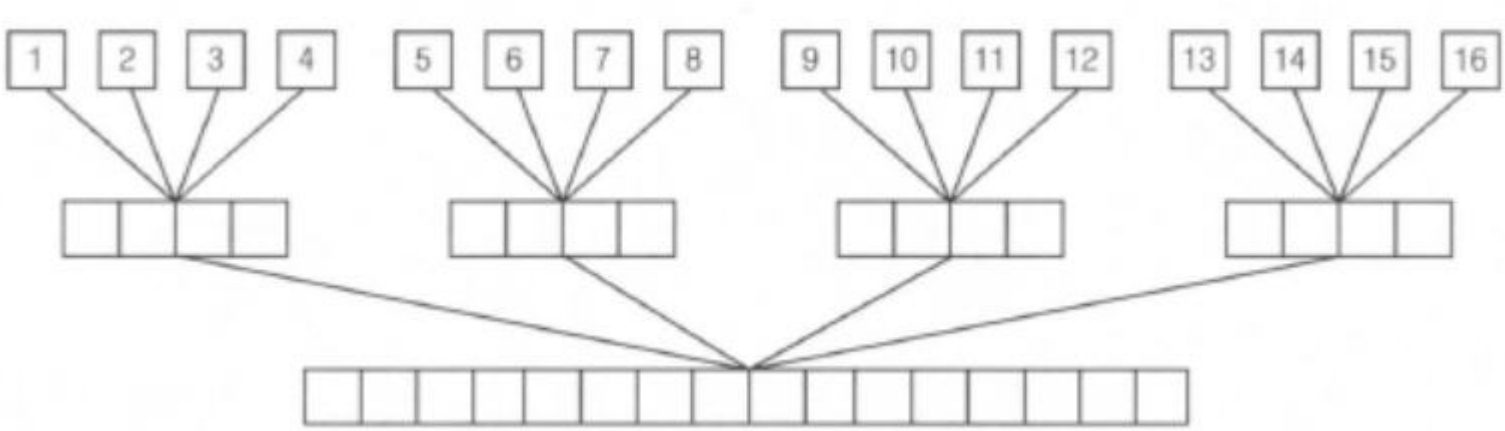
Phase 1



Phase 2



16개의 서브파일에 대한 4-way 병합

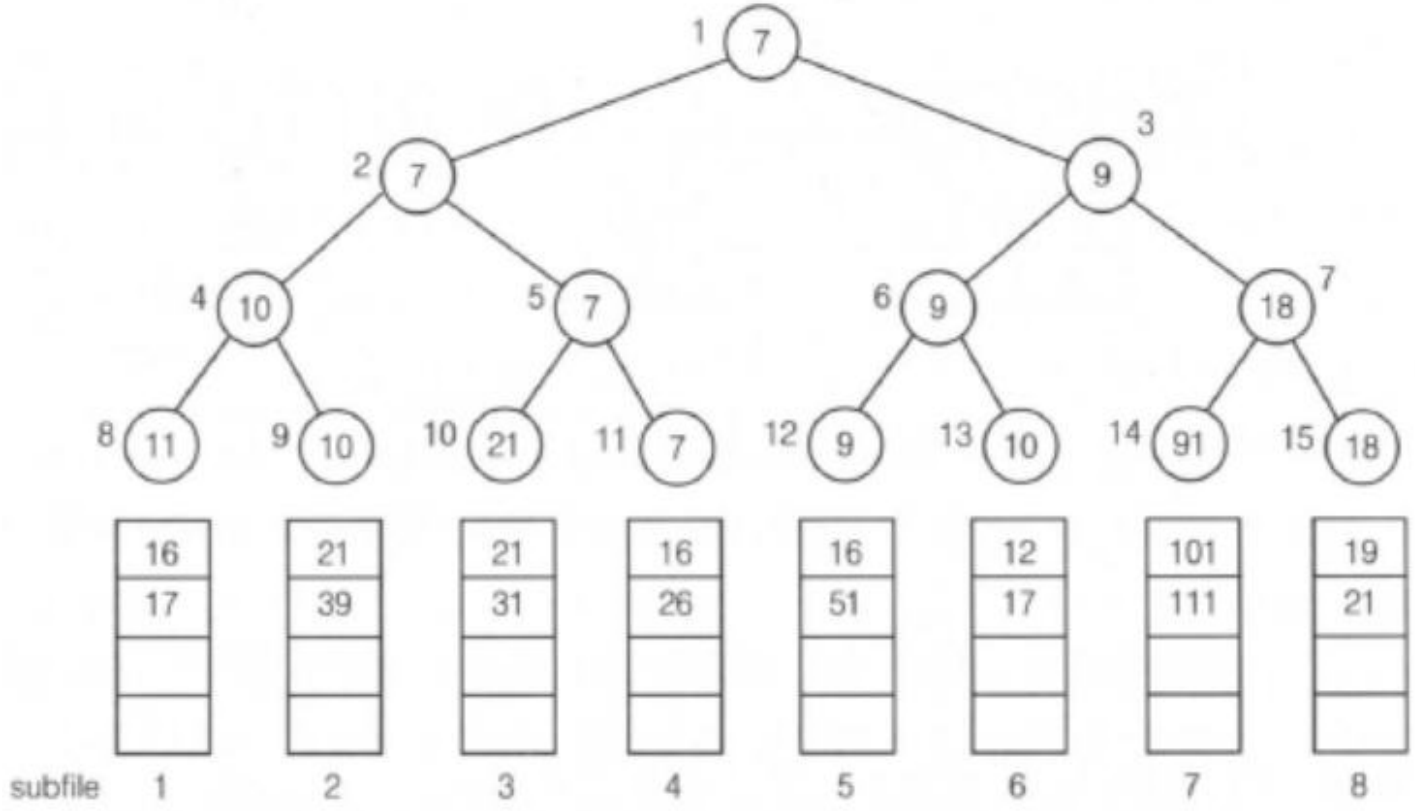


k-way 병합: m개의 서브파일에 대해 $\lceil log_k m \rceil$회의
자료처리를 하므로 고차병합을 하면 그 만큼
입출력 시간인 전송시간을 줄인다.

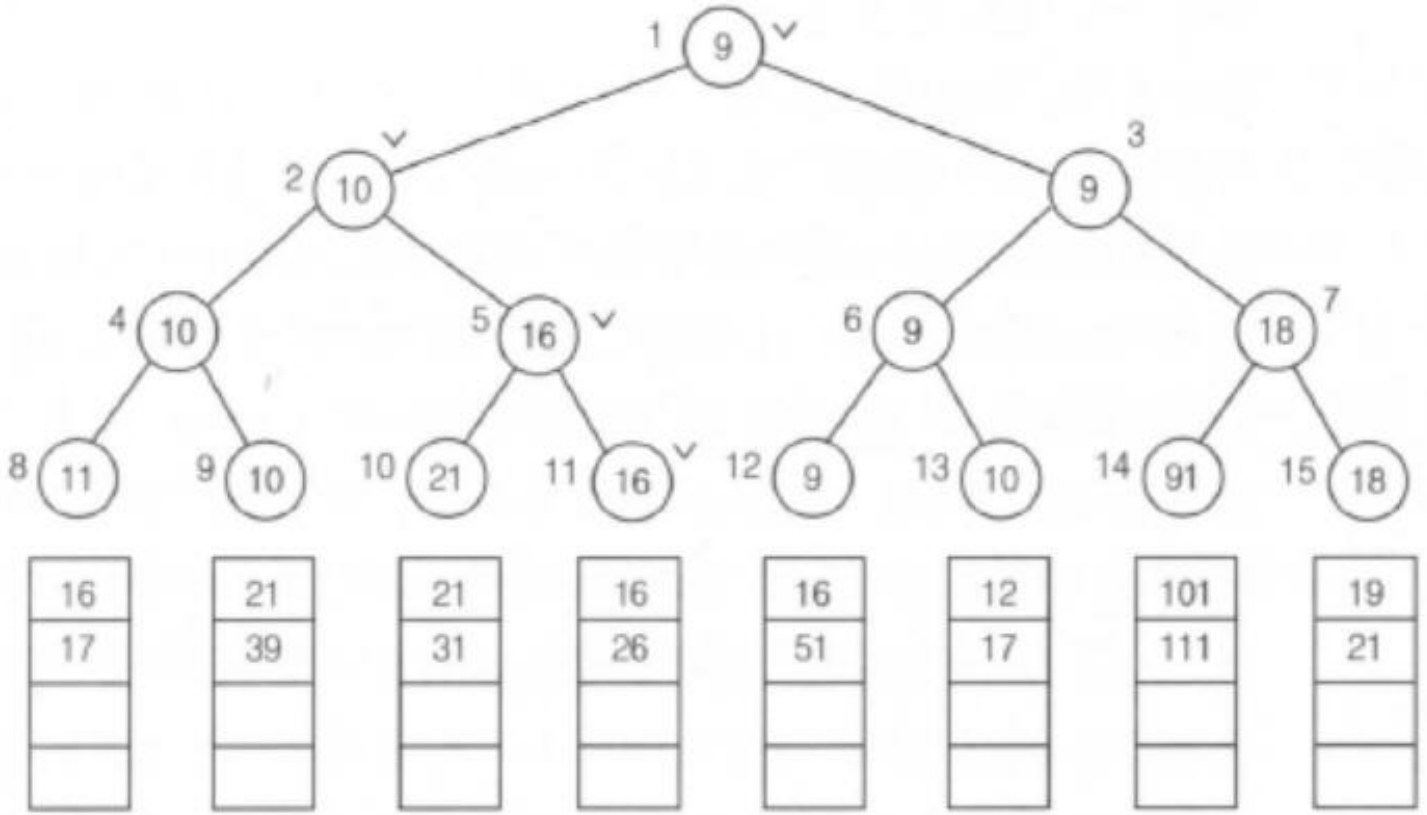
k개의 자료를 비교해 가장 작은 것을 선택하는데 걸리는 시간은
선택 트리 기법 (또는 패자 트리)을 이용해 줄일 수 있다

k-way 병합에 대한 선택트리

- 선택트리: 각 노드가 두 자식노드보다 더 작은 값을 갖도록 구성된 이진트리



〈그림 11.6〉 8-way 병합에 대한 선택 트리



〈그림 11.7〉 하나의 데이터를 출력한 후의 선택 트리(V표는 바꾸어진 노드)

k-way 병합에 대한 선택트리

- 소요 시간

- 처음에 선택트리 만들 때: $O(k)$

- n개의 데이터를 병합하는데: $O(n\log_2 k)$

- m개의 서브파일에 대한 레벨 수는 $o(\log_k m)$이므로 합병에 걸리는 내부 처리시간은 $O(n\log_2 k * \log_k m)=O(n\log_2 m)$

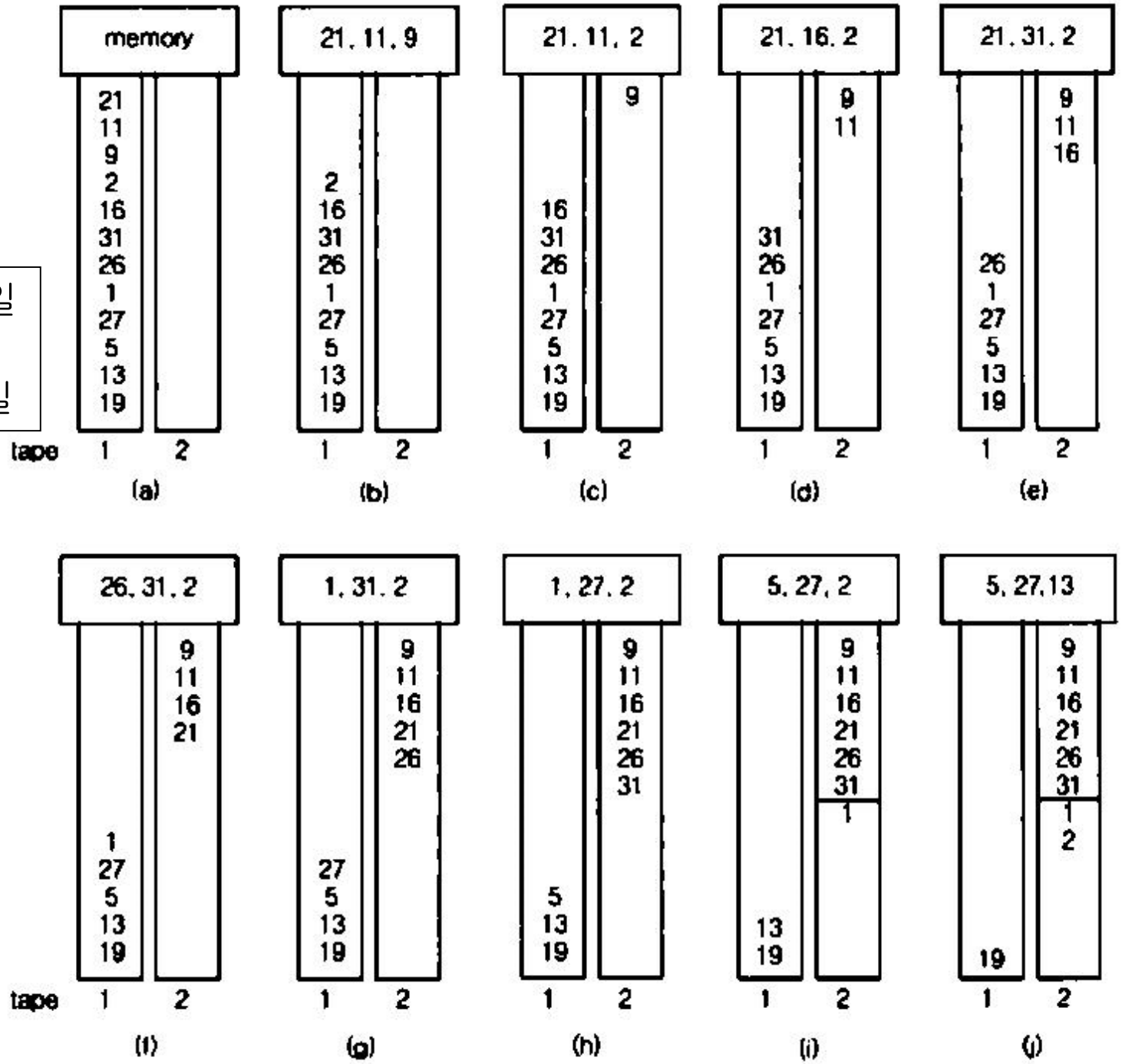
서브파일의 생성

- 내부정렬을 통해 생성되는 서브파일의 크기는 메모리용량보다 클 수 없다
- 대치 선택(replace-selection) 알고리즘 이용하면 평균적으로 약 2배크기의 서브파일을 생성할수 있다

- 대치 선택 알고리즘은 각기 다른 크기의 순서화된 서브파일을 만들어 내기 때문에 최소 병합 트리(minimal merge tree)를 형성하여 병합작업을 수행함으로써 전체 정렬시간을 단축하게 된다

메모리크기:3
tape1: input
tape2: output

12/3=4 서브파일
vs
2 개의 서브파일

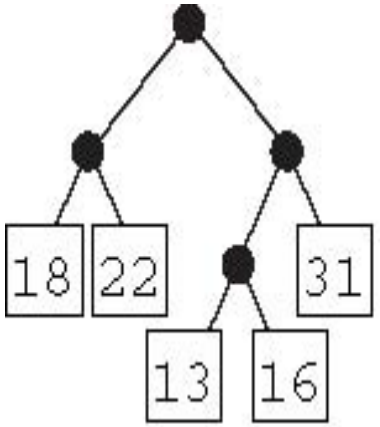


최적병합트리

- 대치 선택 알고리즘에 의해 서브 리스트의 길이가 18,22,13,16,31인 리스트들이 생성되었다고 하자
- 아래 병합트리에서 데이터의 총 이동회수는?

병합트리의 각 레벨은 한 번의 이동을 필요로 하므로 총이동 회수는 $\sum_{1 \le i \le r} q_i k_i$ 이다

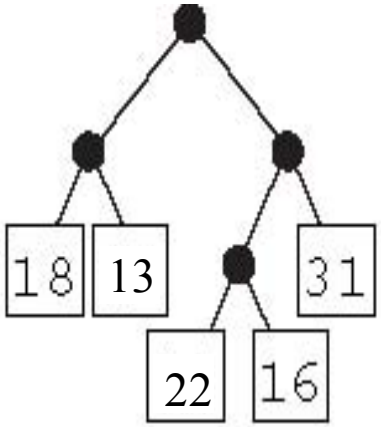
(k_i는 루트노드에서 가중값 q_i를 갖는 외부노드까지의 거리)



18+22+31=71

13+16=29

WEPL = 2(71) + 3(29) = 229



18+13+31=62

22+16=38

WEPL = 2(62) + 3(38) = 238

C-way병합을 위한 최적병합 C-진 트리

- 최소값을 가진 C개의 노드의 값들이 더해지고 그 합계는 C개의 노드를 대체하는 하나의 노드가 된다.
- 이러한 과정을 반복 적용하여 하나의 노드가 남게 될 때 트리의 구성은 끝난다

예) 28, 25, 13, 10, 8, 7, 6, 2, 1 에 대한 3-진 트리?