External Sorting

- Basic Algorithms
  - External Sorting
  - Computing Projections
  - Computing Selection
  - Computing Joins

External Sorting

- Large amount of data in secondary storage
- Limited number of in memory space (memory buffers)
- Partial Sorting
- K-way merge sorted results

External Sorting
Simple Example with Integers

- Memory Space: 4, 5 word (each integer needs one word) memory buffers (i.e. there is only room for 20 integers in memory at one time)
- In the terms we will use later we have 4 memory buffers and a page holds 5 integers
- 223 Integers to sort
- In terms we will use later we have ceiling(223/5) = 45 pages to sort with 4 memory buffers
- Partial Sorting
- K-way merge sorted results

External Sorting
Simple Example with Integers

- Partial Sort
- Read 4 pages, sort them, write the 4 sorted pages to secondary storage
- Results in ceiling(45/4) = 12 sorted sequences

External Sorting
Simple Example with Integers

- Merge Sorted sequences (first pass)
- 3 input pages and 1 output page
- Results in ceiling(12/3) = 4 sorted sequences

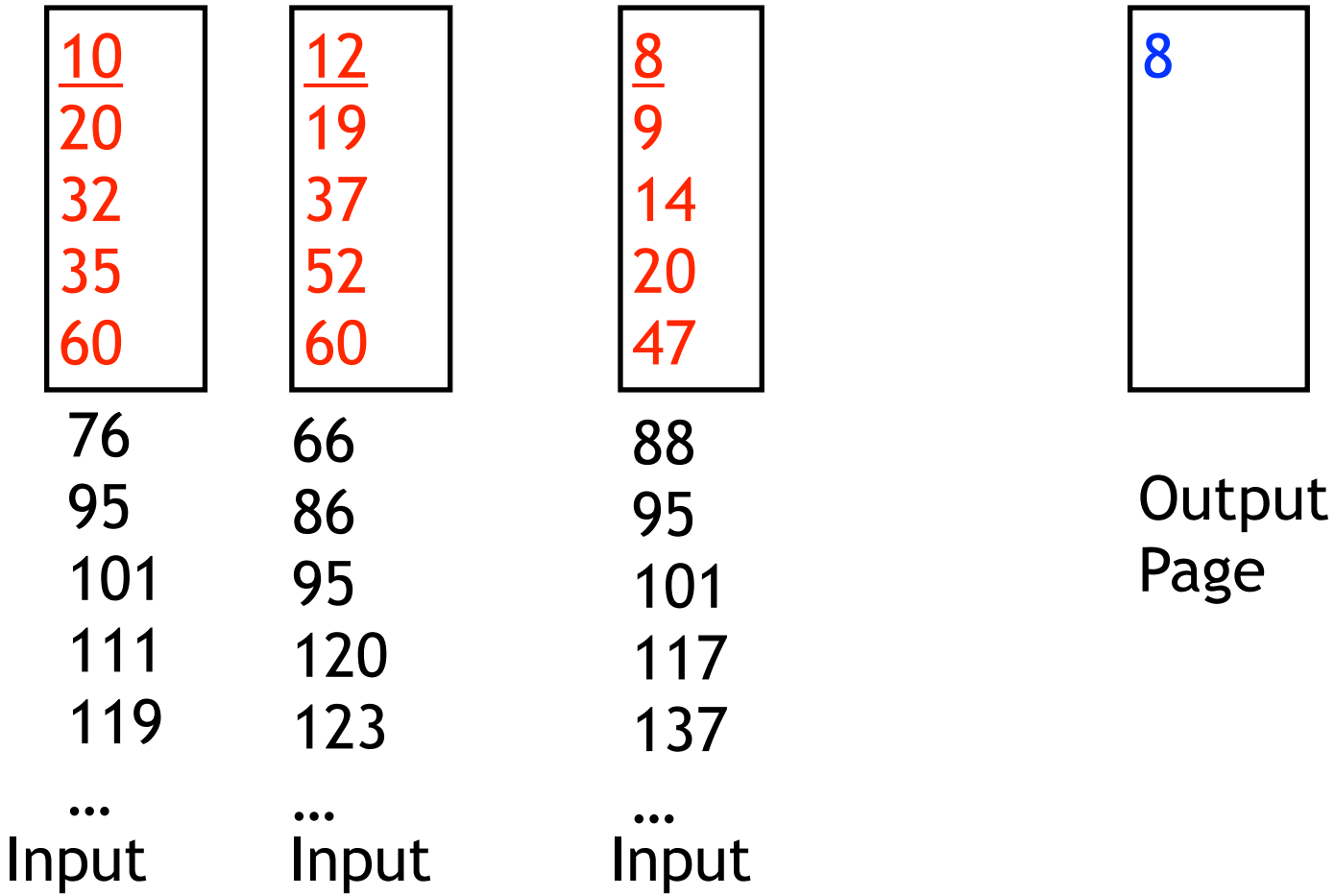
External Sorting
Simple Example with Integers

- Merge Sorted sequences (second pass)
- 3 input pages and 1 output page
- Results in ceiling(4/3) = 2 sorted sequences

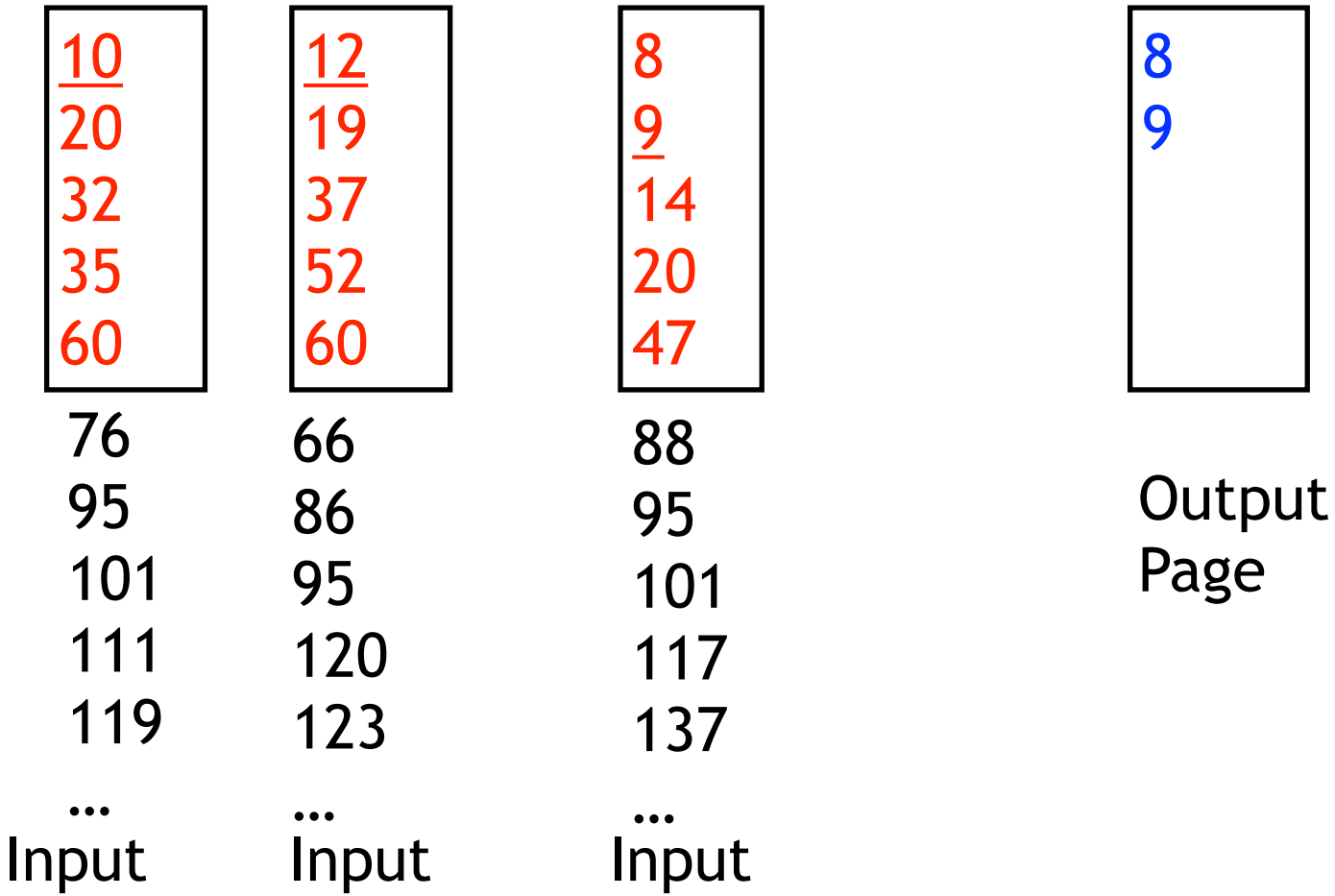
External Sorting
Simple Example with Integers

- Merge Sorted sequences (third pass)
- 3 input pages and 1 output page
- Results in ceiling(2/3) = 1 sorted sequences

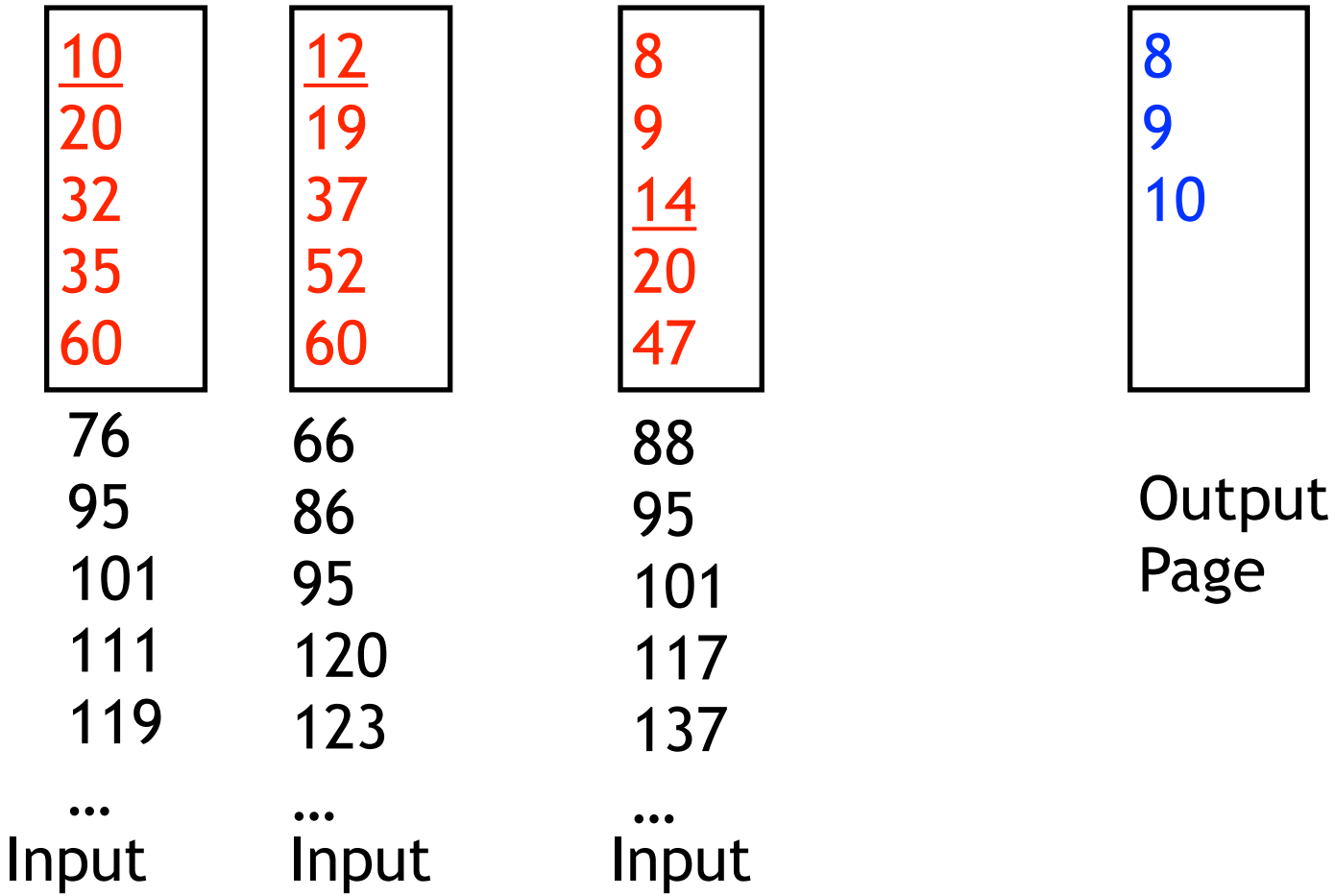
Merge Example



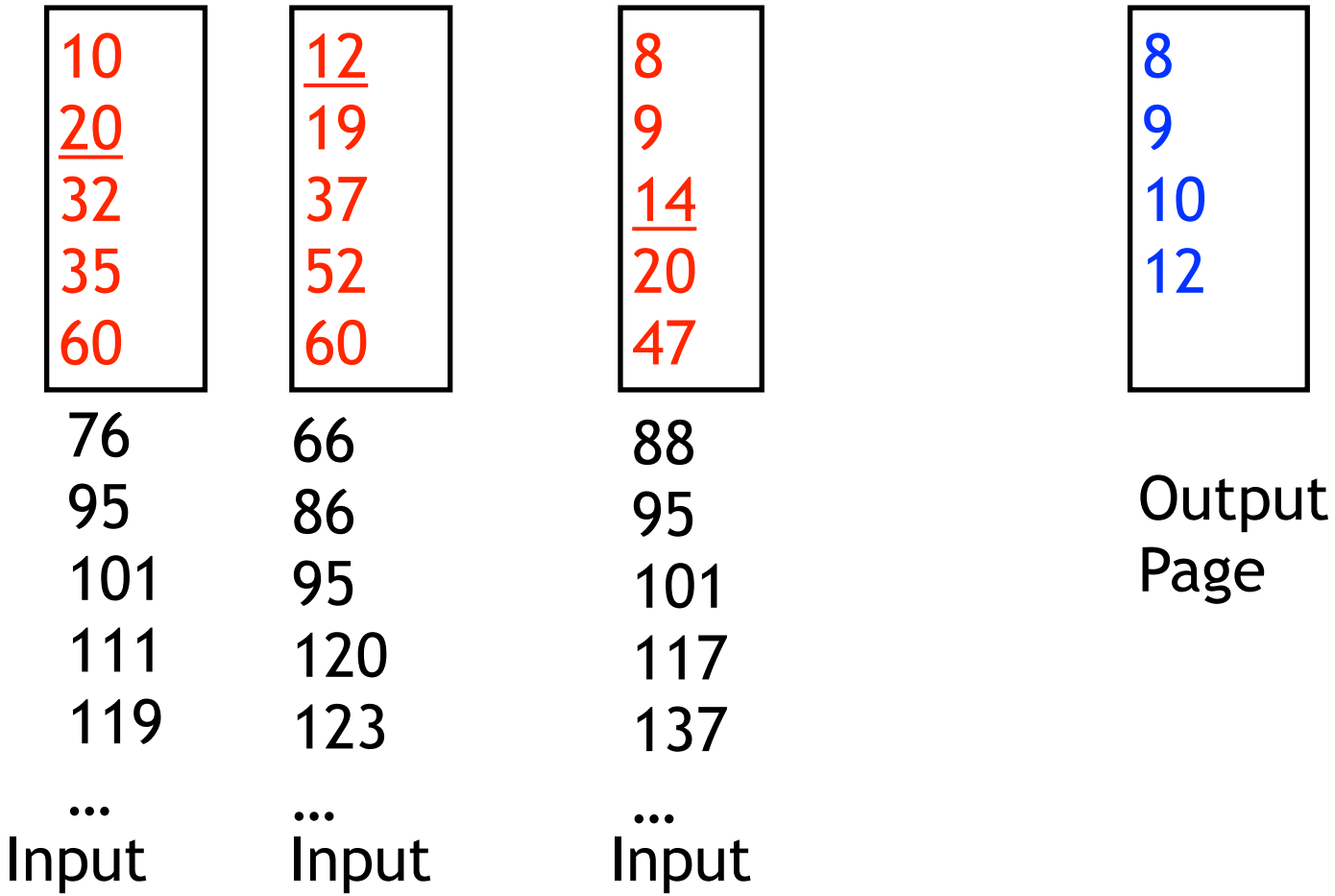
Merge Example



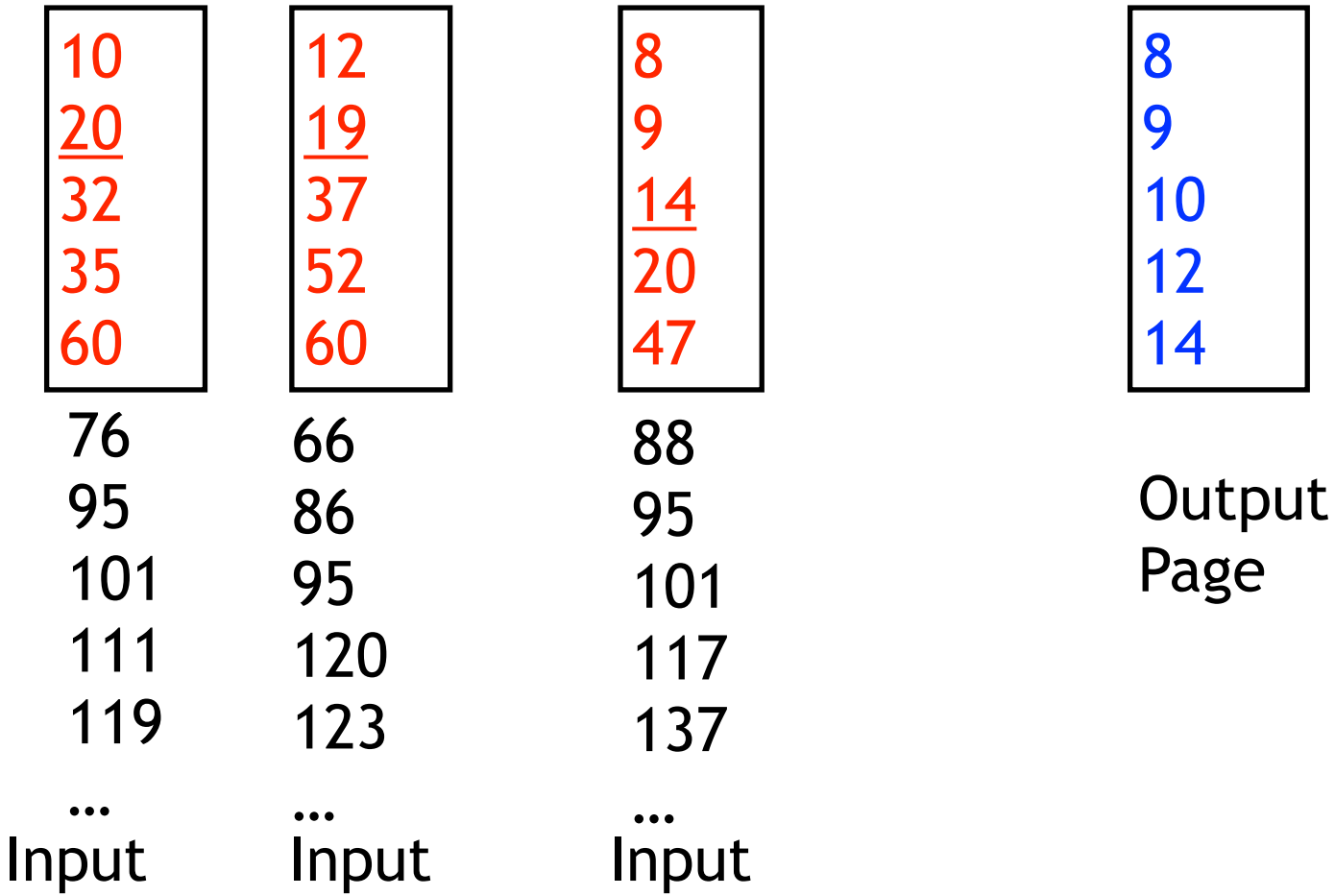
Merge Example



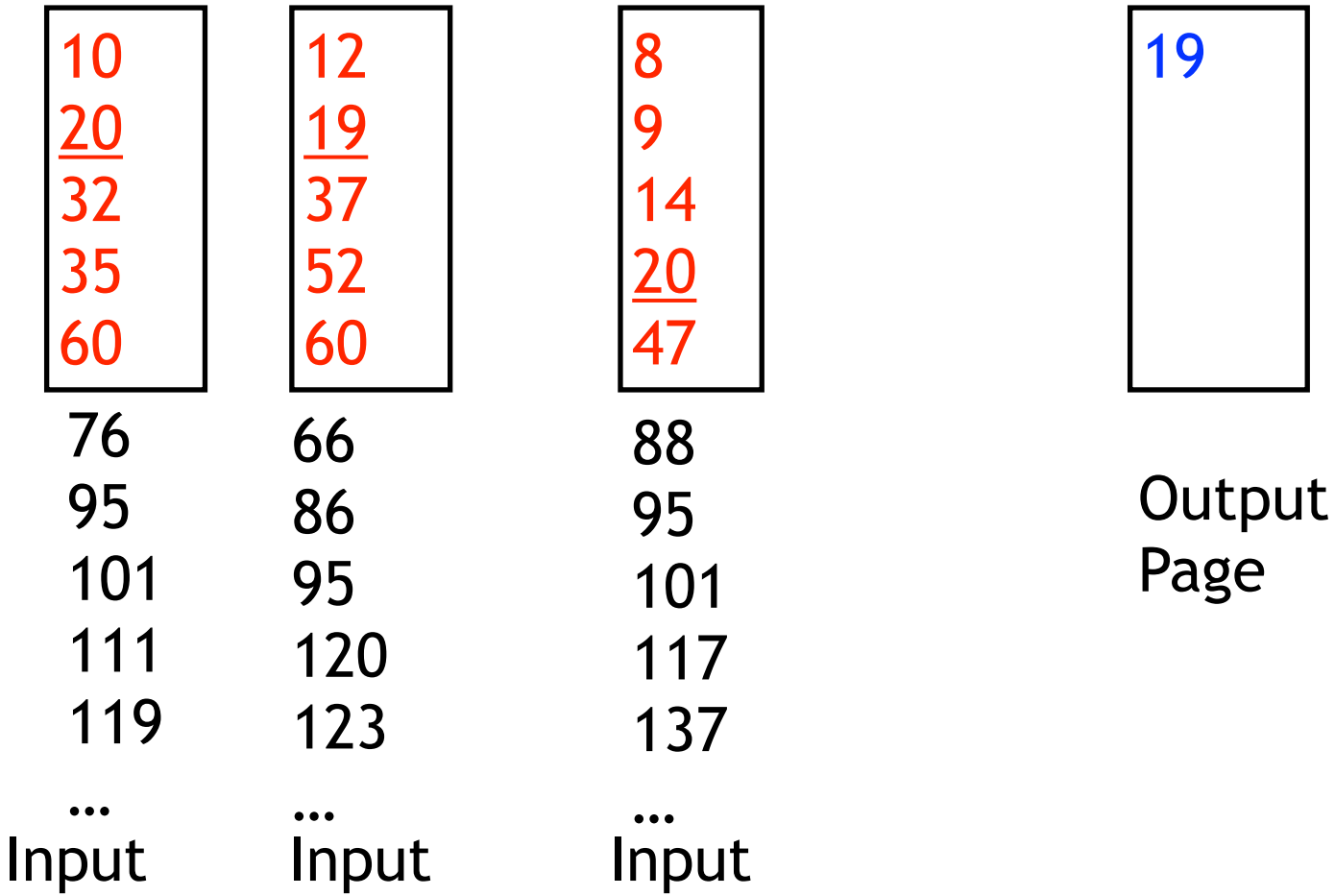
Merge Example



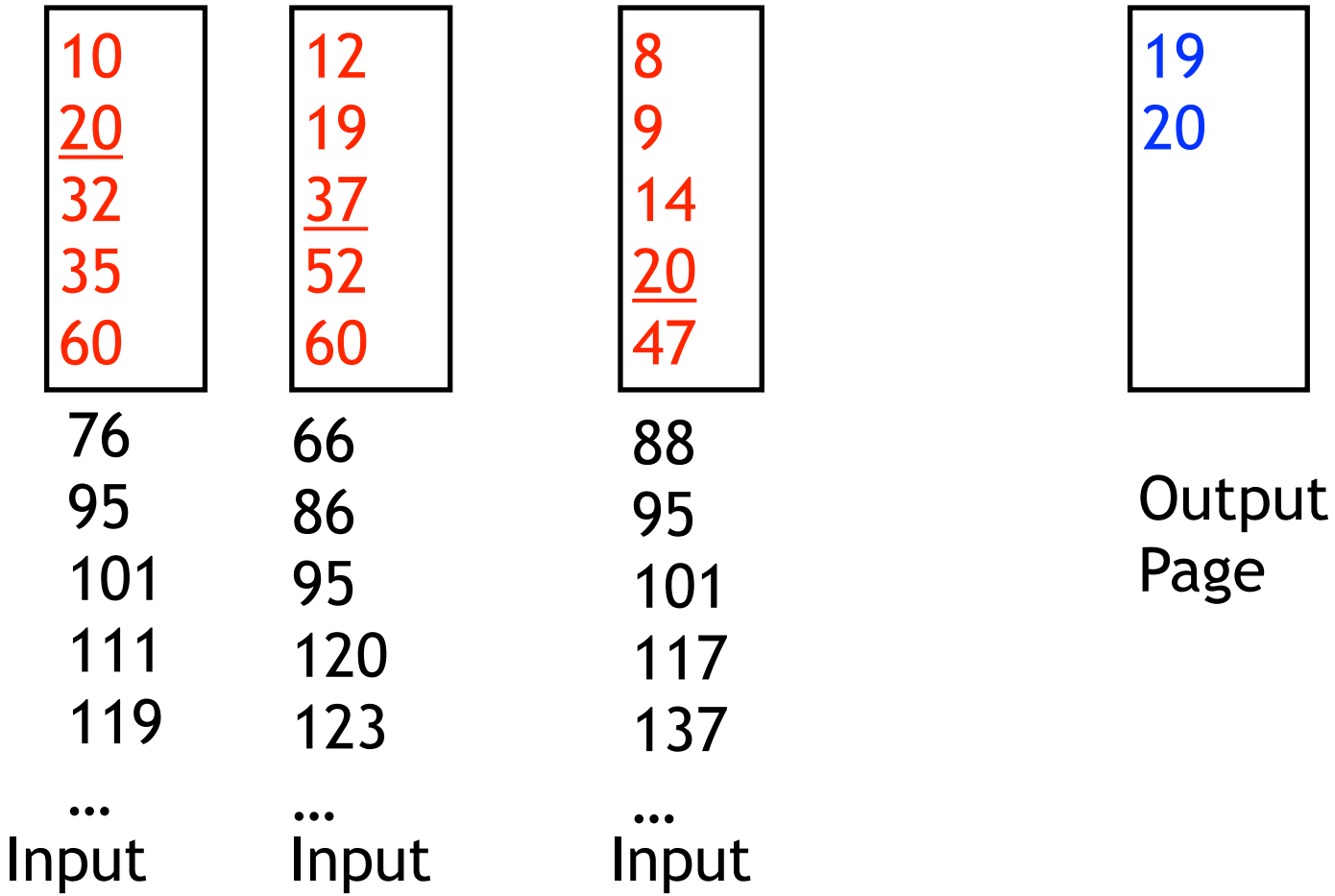
Merge Example



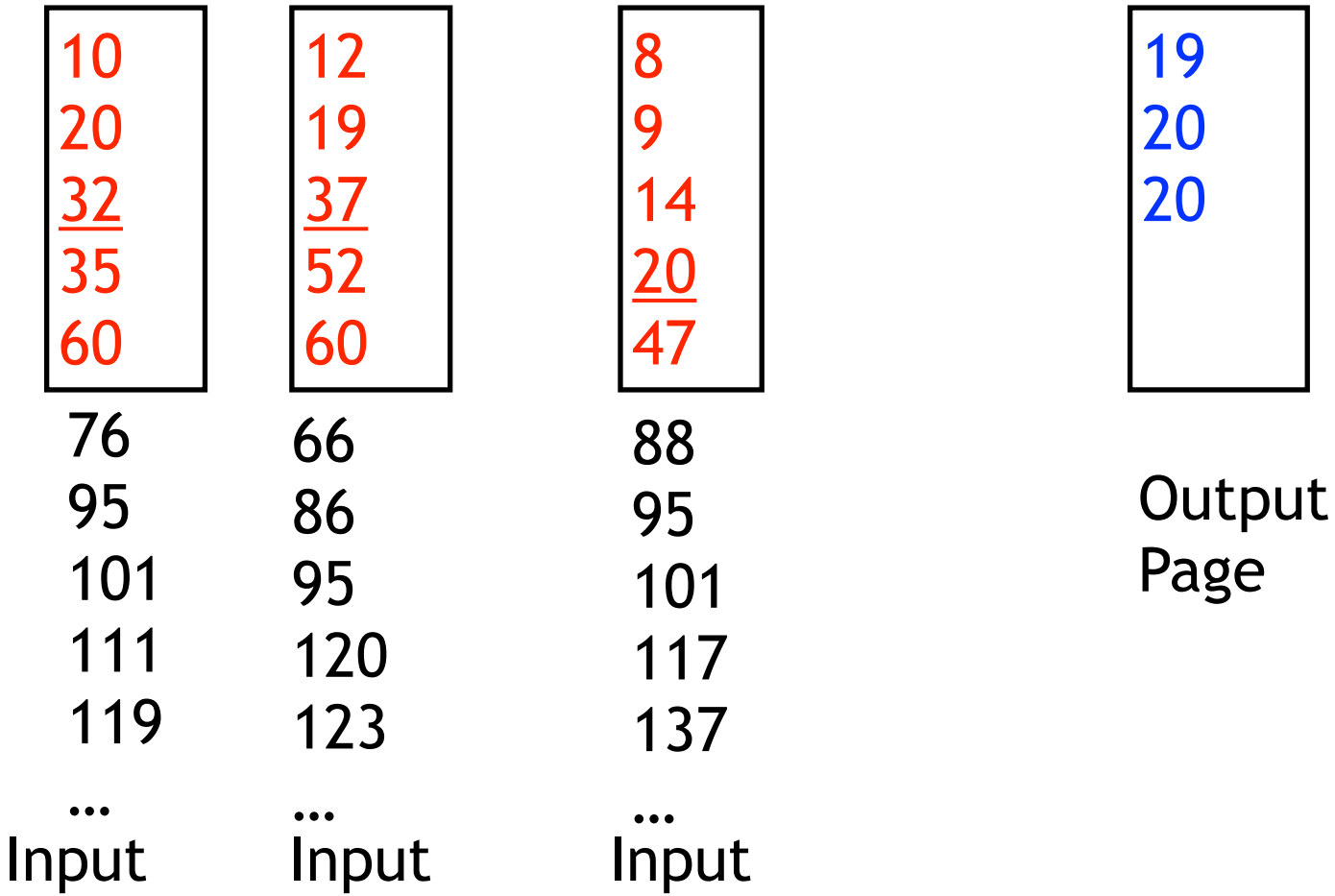
Merge Example



Merge Example



Merge Example



External Sorting

- Partial Sorting
- K-way merging
- Sorting cost
  - Dominated by I/O
  - Suppose a table with F pages and M in memory page buffers
  - Partial Sort Cost
    - 2F pages operations (F reads and F writes)
    - Produces ceiling(F/M) sorted sequences

External Sorting Cost

- K-way Merge
- ceiling(F/M) sorted sequences after partial sort
- Usually will require multiple passes
- Cost to Partial Sort and Merge into 1 sorted sequence
  - $2F * \text{ceiling}(\log_{(M-1)} F)$

External Sort Cost Example

- How many disk accesses (reads and writes) are needed to sort a relation with 10,000 pages and a 10 page in memory buffers
-

External Sort Cost Example

- Partial sort
  - 2*10,000 page accesses
  - 1000 sorted sequences
- First Merge
  - Merge 9 sequences at a time
  - ceiling(1000/9) sequences

External Sort Cost Example

- Second Merge Phase
  - ceiling(112/9) sequences
  - 13 sequences
- Third Merge Phase
  - ceiling(13/9) sequences
  - 2 sequences
- Fourth Merge Phase
  - ceiling(2/9) = 1

External Sort Cost Example

- Total costs
  - Each merge phase costs 2F
  - Partial sort costs + Merge costs
  - 2F + 4*2F = 10F
  - 10*10000 pages accesses
- Formula estimate
  - 2*10000* ceiling($\log_9 10000$)
  - 10*10000 pages accesses