

# Introduction to **Information Retrieval**

Lecture 8: Index Construction

# Index construction

---

- How do we construct an inverted index?
- What strategies can we use with limited main memory?

# Hardware basics

---

- Many design decisions in information retrieval are based on the characteristics of hardware
- We begin by reviewing hardware basics

# Hardware basics

---

- Access to data in memory is ***much*** faster than access to data on disk.
- Disk seeks:
  - No data is transferred from disk while the disk head is being positioned. → *takes a while for the disk head to move to the part of the disk where data is located.*
  - Therefore: Transferring one large chunk of data from disk to memory is faster than transferring many small chunks.
- Disk I/O is block-based: Reading and writing of entire blocks (as opposed to smaller chunks): 8 to 256 KB.

# Hardware basics

---

- Servers used in IR systems now typically have several GB of main memory, sometimes tens of GB.
- Available disk space is several orders of magnitude larger.

# Hardware assumptions for this lecture

---

symbol	statistic	value
■ s	average seek time	5 ms = $5 \times 10^{-3}$ s
■ b	transfer time per byte	0.02 $\mu$ s = $2 \times 10^{-8}$ s
■	processor's clock rate	$10^9$ s <sup>-1</sup>
■ p	low-level operation (e.g., compare & swap a word)	0.01 $\mu$ s = $10^{-8}$ s
■	size of main memory	several GB
■	size of disk space	1 TB or more

# RCV1: Our collection for this lecture

---

- As an example for applying scalable index construction algorithms, we will consider the Reuters RCV1 collection.
- This is one year of Reuters newswire (part of 1995 and 1996)

# A Reuters RCV1 document



You are here: [Home](#) > [News](#) > [Science](#) > [Article](#)

Go to a Section: [U.S.](#) [International](#) [Business](#) [Markets](#) [Politics](#) [Entertainment](#) [Technology](#) [Sports](#) [Oddly Enough](#)

## Extreme conditions create rare Antarctic clouds

Tue Aug 1, 2006 3:20am ET

[Email This Article](#) | [Print This Article](#) | [Reprints](#)

[\[-\]](#) Text [\[+\]](#)



SYDNEY (Reuters) - Rare, mother-of-pearl colored clouds caused by extreme weather conditions above Antarctica are a possible indication of global warming, Australian scientists said on Tuesday.

Known as nacreous clouds, the spectacular formations showing delicate wisps of colors were photographed in the sky over an Australian meteorological base at Mawson Station on July 25.



# Reuters RCV1 statistics

---

■ symbol	statistic	value
■ N	documents	800,000
■ M	terms (= word types)	400,000
■ non-positional postings	100,000,000	(100 Million)

# Recap: Lec 1 index construction

- Documents are parsed to extract words and these are saved with the Document ID.

Doc 1

I did enact Julius  
Caesar I was killed  
i' the Capitol;  
Brutus killed me.

Doc 2

So let it be with  
Caesar. The noble  
Brutus hath told you  
Caesar was ambitious



Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

# Key step

- After all documents have been parsed, the inverted file is sorted by terms.

We focus on this sort step.  
We have 100M items to sort.

Term	Doc #	Term	Doc #
I	1	ambitious	2
did	1	be	2
enact	1	brutus	1
julius	1	brutus	2
caesar	1	capitol	1
I	1	caesar	1
was	1	caesar	2
killed	1	caesar	2
i'	1	did	1
the	1	enact	1
capitol	1	hath	1
brutus	1	I	1
killed	1	I	1
me	1	i'	1
so	2	it	2
let	2	julius	1
it	2	killed	1
be	2	killed	1
with	2	let	2
caesar	2	me	1
the	2	noble	2
noble	2	so	2
brutus	2	the	1
hath	2	the	2
told	2	told	2
you	2	you	2
caesar	2	was	1
was	2	was	2
ambitious	2	with	2

# Scaling index construction

---

- In-memory index construction does not scale
  - Can't stuff entire collection into memory, sort, then write back
- How can we construct an index for very large collections?
- Taking into account the hardware constraints we just learned about . . . Memory, disk, speed, etc.

# Sort-based index construction

---

- As we build the index, we parse docs one at a time and update the postings lists of terms.
- The final postings for any term are incomplete until the end.
- At 12 bytes per non-positional postings entry (*term, doc, freq*), demands a lot of space for large collections.
- $T = 100,000,000$  in the case of RCV1
  - So ... we can do this in memory now, but typical collections are much larger. E.g., the *New York Times* provides an index of >150 years of newswire
- Thus: We need to store intermediate results on disk.

## Sort using disk as “memory”?

---

- Can we use the same index construction algorithm for larger collections, but by using disk instead of memory?
- No: Sorting  $T = 100,000,000$  records on disk is too slow – too many disk seeks.
- We need an *external sorting algorithm*.

# Bottleneck

---

- Parse and build postings entries one doc at a time
- Now sort postings entries by term (then by doc within each term)
- Doing this with random disk seeks would be too slow
  - must sort  $T=100M$  records

# BSBI: Blocked sort-based Indexing

## (Sorting with fewer disk seeks)

---

- 12-byte (4+4+4) records (*term*, *doc*, *freq*).
- These are generated as we parse docs.
- Must now sort 100M such 12-byte records by *term*.
- Define a Block ~ 10M such records
  - Can easily fit a couple into memory.
  - Will have 10 such blocks to start with.
- Basic idea of algorithm:
  - Accumulate postings for each block, sort, write to disk.
  - Then merge the blocks into one long sorted order.



## BSBINDEXCONSTRUCTION()

```
1   $n \leftarrow 0$ 
2  while (all documents have not been processed)
3  do  $n \leftarrow n + 1$ 
4       $block \leftarrow \text{PARSENEXTBLOCK}()$ 
5       $\text{BSBI-INVERT}(block)$ 
6       $\text{WRITEBLOCKTODISK}(block, f_n)$ 
7   $\text{MERGEBLOCKS}(f_1, \dots, f_n; f_{\text{merged}})$ 
```

Line 4: documents in this block (that we have just read) parsed into termID-docID pairs

## BSBINDEXCONSTRUCTION()

```
1   $n \leftarrow 0$ 
2  while (all documents have not been processed)
3  do  $n \leftarrow n + 1$ 
4       $block \leftarrow \text{PARSENEXTBLOCK}()$ 
5      BSBI-INVERT( $block$ )
6       $\text{WRITEBLOCKTODISK}(block, f_n)$ 
7   $\text{MERGEBLOCKS}(f_1, \dots, f_n; f_{\text{merged}})$ 
```

Line 5: BSBI-INVERT involves 2 steps:

- (1) sort the termID-docID pairs
- (2) collect all termID-docID pairs with the same termID into a postings list

Gives an inverted index for the block we have just read

# Sorting 10 blocks of 10M records

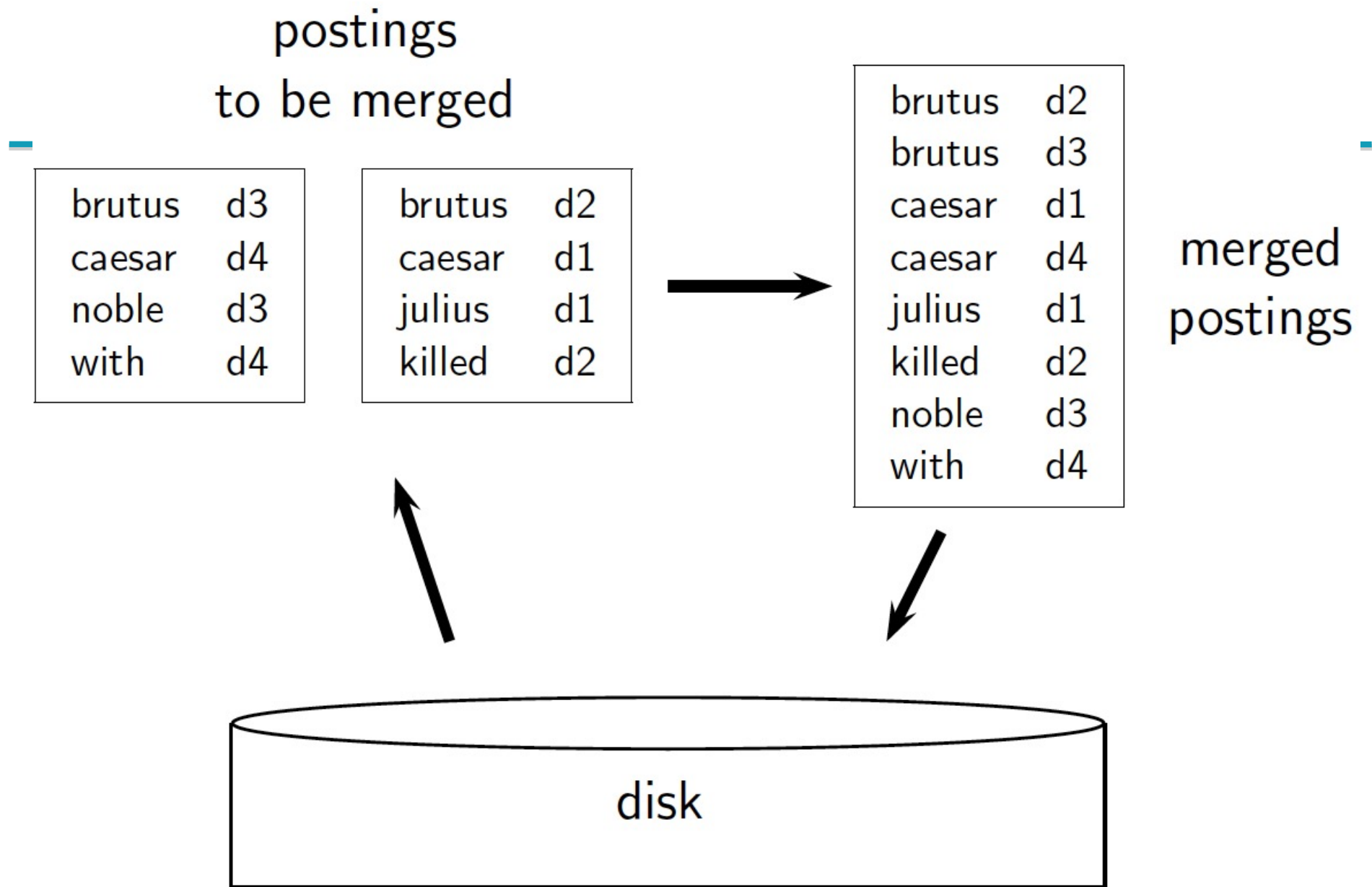
---

- First, read each block and sort within:
  - Quicksort takes  $2N \ln N$  expected steps
  - In our case  $2 \times (10M \ln 10M)$  steps
- *Exercise: estimate total time to read each block from disk and and quicksort it.*
- 10 times this estimate – gives us 10 sorted runs of 10M records each.

## BSBINDEXCONSTRUCTION()

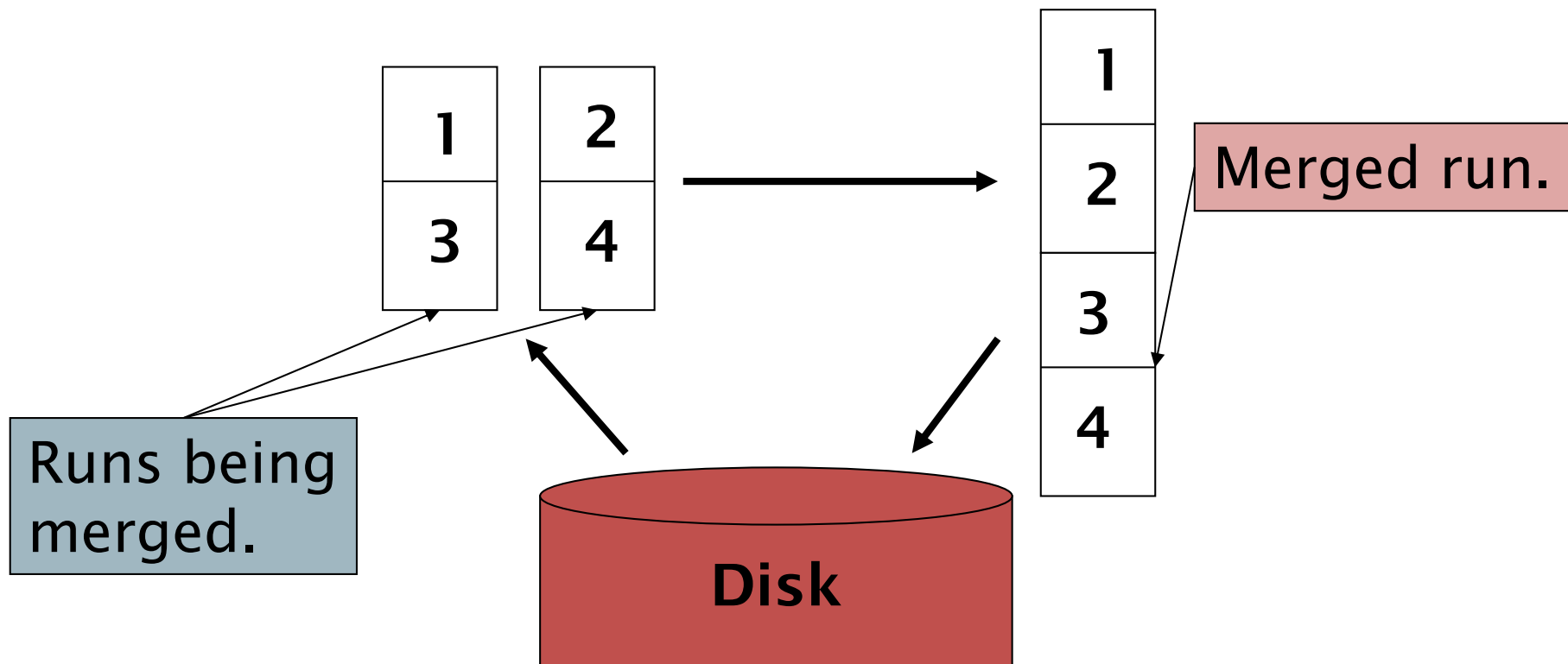
```
1   $n \leftarrow 0$ 
2  while (all documents have not been processed)
3  do  $n \leftarrow n + 1$ 
4       $block \leftarrow \text{PARSENEXTBLOCK}()$ 
5       $\text{BSBI-INVERT}(block)$ 
6       $\text{WRITEBLOCKTODISK}(block, f_n)$ 
7   $\text{MERGEBLOCKS}(f_1, \dots, f_n; f_{\text{merged}})$ 
```

Line 7: how to merge the blocks?



# How to merge the sorted runs?

- Can do binary merges, with a merge tree of  $\log_2 10 = 4$  layers.
- During each layer, read into memory runs in blocks of 10M, merge, write back.



# How to merge the sorted runs?

---

- But it is more efficient to do a **multi-way merge**, where you are reading from all blocks simultaneously
- Providing you read decent-sized chunks of each block into memory and then write out a decent-sized output chunk, then you're not killed by disk seeks

# Terms represented as termIDs?

---

- To make index construction more efficient, we can represent terms as termIDs
- Each *termID* is a unique serial number
- How to build the mapping from terms to termIDs?
  - On the fly while we are processing the collection; or,
  - A two-pass approach: compile the vocabulary in the first pass and construct the inverted index in the second pass
- Some indexing algorithms use termIDs, others directly use the term
  - Hybrid approaches possible: map frequently occurring terms to termIDs, rare terms directly handled



# Remaining problem with sort-based algorithm

---

- BSBI needs the dictionary (which grows dynamically) in order to implement a term to termID mapping.
- BSBI assumes that we can keep the dictionary in memory throughout.
- Actually, we could work with **term**,docID postings instead of **termID**,docID postings . . .
- . . . but then intermediate files become very large. (We would end up with a scalable, but very slow index construction method.)

# SPIMI:

## Single-pass in-memory indexing

---

- Key idea 1: Use terms instead of termIDs. Generate separate dictionaries for each block – no need to maintain term-termID mapping across blocks.
- Key idea 2: Don't sort. Accumulate postings in postings lists as they occur.
- With these two ideas, we can generate a complete inverted index for each block.
- These separate indexes can then be merged into one big index.

# SPIMI-Invert (to replace BSBI-Invert step on slide 17)

---

```
SPIMI-INVERT(token_stream)
1  output_file = NEWFILE()
2  dictionary = NEWHASH()
3  while (free memory available)
4  do token  $\leftarrow$  next(token_stream)
5      if term(token)  $\notin$  dictionary
6          then postings_list = ADDTODICTIONARY(dictionary, term(token))
7          else postings_list = GETPOSTINGSLIST(dictionary, term(token))
8          if full(postings_list)
9              then postings_list = DOUBLEPOSTINGSLIST(dictionary, term(token))
10         ADDTOPOSTINGSLIST(postings_list, docID(token))
11  sorted_terms  $\leftarrow$  SORTTERMS(dictionary)
12  WRITEBLOCKTODISK(sorted_terms, dictionary, output_file)
13  return output_file
```

Merging of blocks is analogous to BSBI (not shown here).

# SPIMI-Invert

---

```
SPIMI-INVERT(token_stream)
1  output_file = NEWFILE()
2  dictionary = NEWHASH()
3  while (free memory available)
4  do token ← next(token_stream)
5      if term(token) ∉ dictionary
6          then postings_list = ADDTODICTIONARY(dictionary, term(token))
7          else postings_list = GETPOSTINGSLIST(dictionary, term(token))
8          if full(postings_list)
9              then postings_list = DOUBLEPOSTINGSLIST(dictionary, term(token))
10         ADDTOPOSTINGSLIST(postings_list, docID(token))
11 sorted_terms ← SORTTERMS(dictionary)
12 WRITEBLOCKTODISK(sorted_terms, dictionary, output_file)
13 return output_file
```

Lines 8-9: We do not know how large the postings list of a term will be when we first encounter it, hence allocate space for a short postings list initially, double the space each time it is full

Merging of blocks is analogous to BSBI (not shown here).



# SPIMI-Invert

---

```
SPIMI-INVERT(token_stream)
1  output_file = NEWFILE()
2  dictionary = NEWHASH()
3  while (free memory available)
4  do token  $\leftarrow$  next(token_stream)
5      if term(token)  $\notin$  dictionary
6          then postings_list = ADDTOdictionary(dictionary, term(token))
7          else postings_list = GETPOSTINGSLIST(dictionary, term(token))
8          if full(postings_list)
9              then postings_list = DOUBLEPOSTINGSLIST(dictionary, term(token))
10         ADDTOPOSTINGSLIST(postings_list, docID(token))
11 sorted_terms  $\leftarrow$  SORTTERMS(dictionary)
12 WRITEBLOCKTODISK(sorted_terms, dictionary, output_file)
13 return output_file
```

Line 10: Instead of first collecting all termID-docID pairs and then sorting them (as we did in BSBI), SPIMI adds a posting directly to its postings list

Merging of blocks is analogous to BSBI (not shown here).

# SPIMI-Invert

---

```
SPIMI-INVERT(token_stream)
1  output_file = NEWFILE()
2  dictionary = NEWHASH()
3  while (free memory available)
4  do token  $\leftarrow$  next(token_stream)
5      if term(token)  $\notin$  dictionary
6          then postings_list = ADDTODICTIONARY(dictionary, term(token))
7          else postings_list = GETPOSTINGSLIST(dictionary, term(token))
8          if full(postings_list)
9              then postings_list = DOUBLEPOSTINGSLIST(dictionary, term(token))
10         ADDTODICTIONARY(postings_list, docID(token))
11 sorted_terms  $\leftarrow$  SORTTERMS(dictionary)
12 WRITEBLOCKTODISK(sorted_terms, dictionary, output_file)
13 return output_file
```

Lines 11: Sort the terms before writing to disk since we want to write postings lists in lexicographic order to facilitate the final merging step

Merging of blocks is analogous to BSBI (not shown here).

# SPIMI-Invert

---

```
SPIMI-INVERT(token_stream)
1  output_file = NEWFILE()
2  dictionary = NEWHASH()
3  while (free memory available)
4  do token  $\leftarrow$  next(token_stream)
5      if term(token)  $\notin$  dictionary
6          then postings_list = ADDTOdictionary(dictionary, term(token))
7          else postings_list = GETPOSTINGSLIST(dictionary, term(token))
8          if full(postings_list)
9              then postings_list = DOUBLEPOSTINGSLIST(dictionary, term(token))
10         ADDTOPOSTINGSLIST(postings_list, docID(token))
11 sorted_terms  $\leftarrow$  SORTTERMS(dictionary)
12 WRITEBLOCKTODISK(sorted_terms, dictionary, output_file)
13 return output_file
```

Lines 12: When memory exhausted, write the index of the block (which consists of the dictionary and the postings lists) to disk

Merging of blocks is analogous to BSBI (not shown here).

## SPIMI vs. BSBI

---

- Instead of first collecting all termID-docID pairs and then sorting them (as done in BSBI), SPIMI adds a post directly to its postings list (each postings list is dynamic)
- No sorting in SPIMI → SPIMI is faster and saves memory
- Individual calls of SPIMI-Invert can process much larger blocks than calls of BSBI-Invert



# Till now

---

- Sort-based indexing
  - Naïve in-memory or in-disk inversion
  - Blocked Sort-Based Indexing (BSBI)
    - Merge sort is effective for disk-based sorting (avoid seeks!)
- Single-Pass In-Memory Indexing (SPIMI)
  - No global dictionary
    - Generate separate dictionary for each block
  - Don't sort postings
    - Accumulate postings in postings lists as they occur
- Next
  - Distributed indexing using MapReduce
  - Dynamic indexing: Multiple indices, logarithmic merge

# Distributed Indexing

# Distributed indexing

---

- For web-scale indexing:
  - Must use a distributed computing cluster
- Individual machines are fault-prone
  - Can unpredictably slow down or fail
- How do we exploit such a pool of machines?

# Web search engine data centers

---

- Web search data centers (Google, Bing, Baidu) mainly contain commodity machines.
- Data centers are distributed around the world.
- Estimate: Google ~1 million servers, 3 million processors/cores (Gartner 2007)

# Distributed indexing

---

- **MapReduce**: a general architecture for distributed computing
- Maintain a **master** node (machine) directing the indexing job – considered “safe”.
- Many **worker** nodes, each of which can fail
- Break up indexing into sets of (parallel) tasks.
- Master assigns each task to an idle worker from a pool

# Distributed Indexing

---

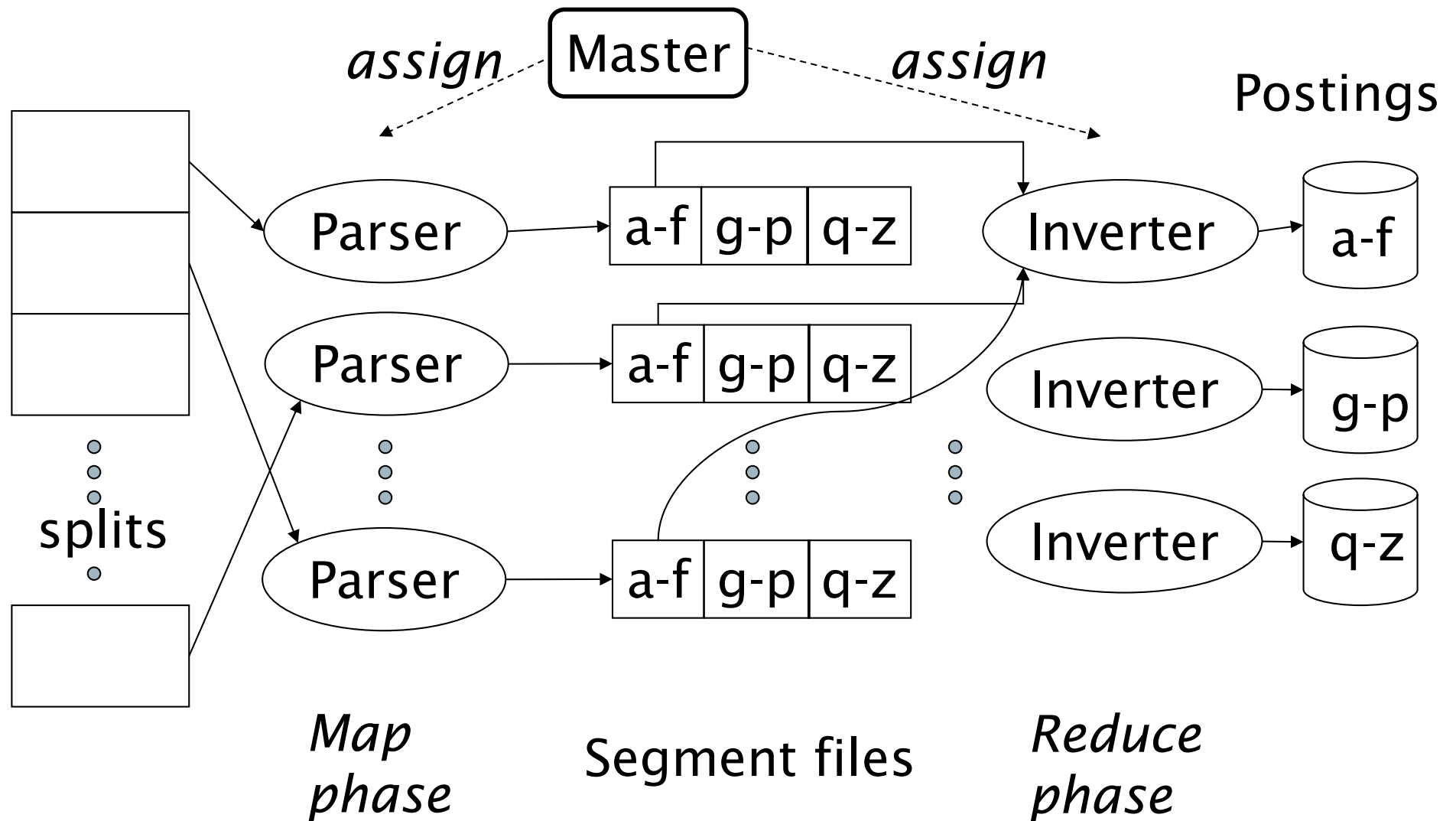
- In general, MapReduce breaks a large computing problem into smaller parts by recasting it in terms of manipulation of *key-value pairs*
- For the task of indexing, a key-value pair has the form (termID,docID)

# Parallel tasks

---

- We will use two sets of parallel tasks
  - Parsers
  - Inverters
- Break the input document collection into *splits*
- Each split is a subset of documents (corresponding to blocks in BSBI/SPIMI)
- *Map phase*: map splits of the input data to key-value pairs - done by parser machines
- *Reduce phase*: Collect all values (here: docIDs) for a given key (here: termID) – done by inverter machines

# Data flow

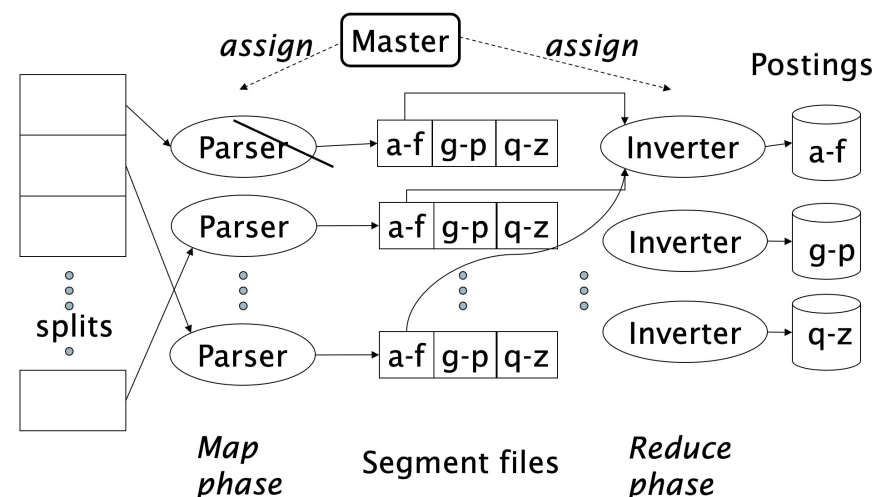




# Parsers (Map phase)

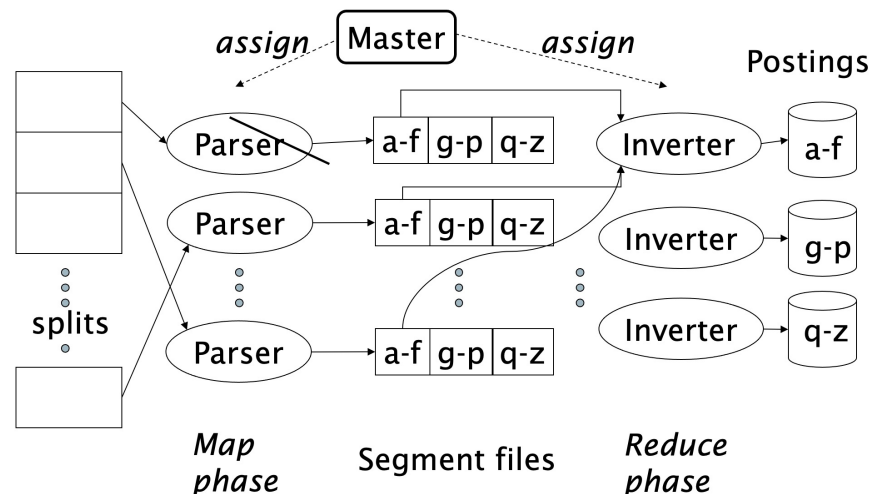
- Master assigns a split to an idle parser machine
- Parser reads a document at a time and emits (term, doc) pairs
- Parser writes pairs into  $j$  partitions
- Each partition is for a range of terms' first letters (called term-partition)

- (e.g., **a-f**, **g-p**, **q-z**) – here  $j = 3$ .



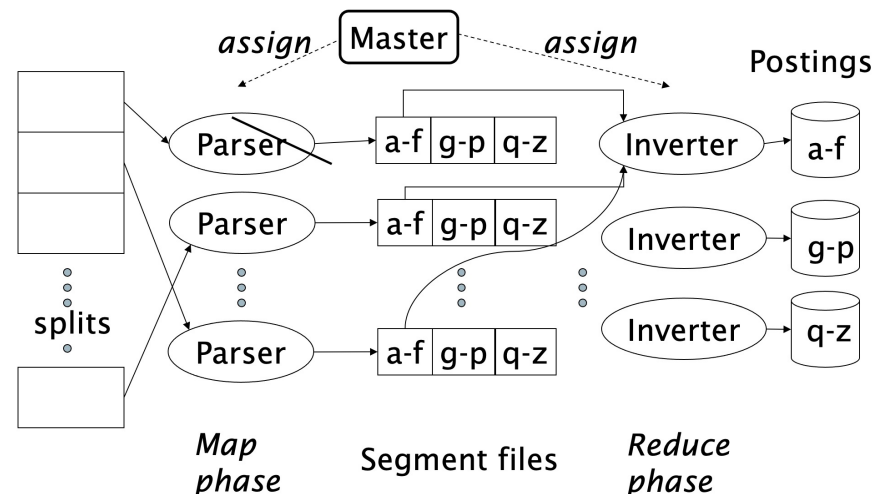
# Inverters (Reduce phase)

- An inverter collects all (term,doc) pairs (= postings) for one term-partition.
- Sorts and writes to postings lists
  - Finally, the list of values is sorted for each key and written to the final sorted postings list (“postings” in the figure).



# Map-reduce indexing

- Note:
  - Parsers and inverters are not separate sets of machines.
  - The same machine can be a parser in the map phase and an inverter in the reduce phase.



# Dynamic Indexing

# Dynamic indexing

---

- Up to now, we have assumed that collections are static.
- They rarely are:
  - Documents come in over time and need to be inserted.
  - Documents are deleted and modified.
- This means that the **dictionary and postings lists have to be modified**:
  - Postings updates for terms already in dictionary
  - New terms added to dictionary

# Simplest approach

---

- Maintain “big” main index
- New docs go into “small” auxiliary index
- Search across both, merge results
- Deletions
  - Invalidation bit-vector for deleted docs (in main index)
  - Filter docs output on a search result by this invalidation bit-vector
- Periodically, re-index into one main index

# Issues with main and auxiliary indexes

---

- Problem of frequent merges – you touch stuff a lot
- Poor performance during merge
- Actually:
  - Merging of the auxiliary index into the main index is efficient if we keep a separate file for each postings list.
  - Merge is the same as a simple append.
  - But then we would need a lot of files – inefficient for OS.
- Assumption for the rest of the lecture: The index is one big file.
- In reality: Use a scheme somewhere in between (e.g., split very large postings lists, collect postings lists of length 1 in one file, etc.)

# Logarithmic merge

---

- Maintain **a series of indexes**, each twice as large as the previous one
  - At any time, some of these powers of 2 are instantiated
- Keep smallest ( $Z_0$ ) in memory
- Larger ones ( $I_0, I_1, \dots$ ) on disk
- If  $Z_0$  gets too big ( $> n$ ), write to disk as  $I_0$
- or merge with  $I_0$  (if  $I_0$  already exists) as  $Z_1$
- Either write merge  $Z_1$  to disk as  $I_1$  (if no  $I_1$ )
- Or merge with  $I_1$  to form  $Z_2$



LMERGEADDTOKEN(*indexes*,  $Z_0$ , *token*)

```
1   $Z_0 \leftarrow \text{MERGE}(Z_0, \{\text{token}\})$ 
2  if  $|Z_0| = n$ 
3    then for  $i \leftarrow 0$  to  $\infty$ 
4      do if  $l_i \in \text{indexes}$ 
5        then  $Z_{i+1} \leftarrow \text{MERGE}(l_i, Z_i)$ 
6          ( $Z_{i+1}$  is a temporary index on disk.)
7           $\text{indexes} \leftarrow \text{indexes} - \{l_i\}$ 
8        else  $l_i \leftarrow Z_i$     ( $Z_i$  becomes the permanent index  $l_i$ .)
9           $\text{indexes} \leftarrow \text{indexes} \cup \{l_i\}$ 
10         BREAK
11      $Z_0 \leftarrow \emptyset$ 
```

Details in the Manning book, Section 4.5  
Also see example provided

LOGARITHMICMERGE()

```
1   $Z_0 \leftarrow \emptyset$     ( $Z_0$  is the in-memory index.)
2   $\text{indexes} \leftarrow \emptyset$ 
3  while true
4  do LMERGEADDTOKEN(indexes,  $Z_0$ , GETNEXTTOKEN())
```

# Further issues with multiple indexes

---

- Collection-wide statistics are hard to maintain
- E.g., when we spoke of spell-correction: which of several corrected alternatives (for a query) do we present to the user?
  - We said, pick the one with the most hits
- How do we maintain the top ones with multiple indexes and invalidation bit vectors?
  - One possibility: ignore everything but the main index for such ordering

# Dynamic indexing at search engines

---

- All the large search engines now do dynamic indexing
- Their indices have frequent incremental changes
  - News items, blogs, new topical web pages
- But (sometimes/typically) they also periodically reconstruct the index from scratch
  - Need more hardware resources
  - Query processing is then switched to the new index, and the old index is deleted

# Other sorts of indexes

---

- Positional indexes
  - Same sort of sorting problem ... just larger
- Building character  $n$ -gram indexes:
  - As text is parsed, enumerate  $n$ -grams
  - For each  $n$ -gram, need pointers to all dictionary terms containing it – the “postings”
  - Note that the same “postings entry” will arise repeatedly in parsing the docs – need efficient hashing to keep track of this.
    - E.g., that the trigram uou occurs in the term ***deciduous*** will be discovered on each text occurrence of ***deciduous***
    - Only need to process each term once