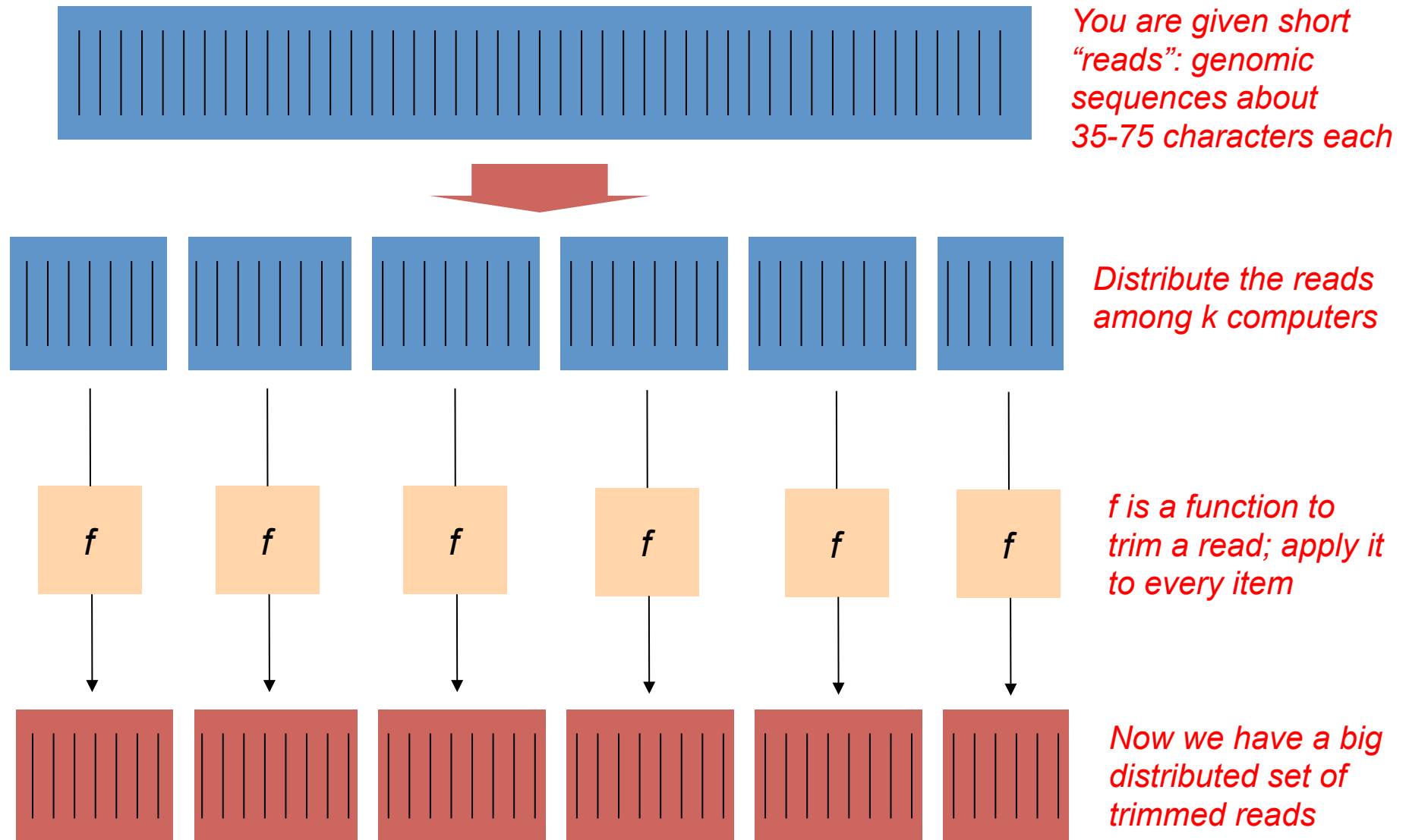


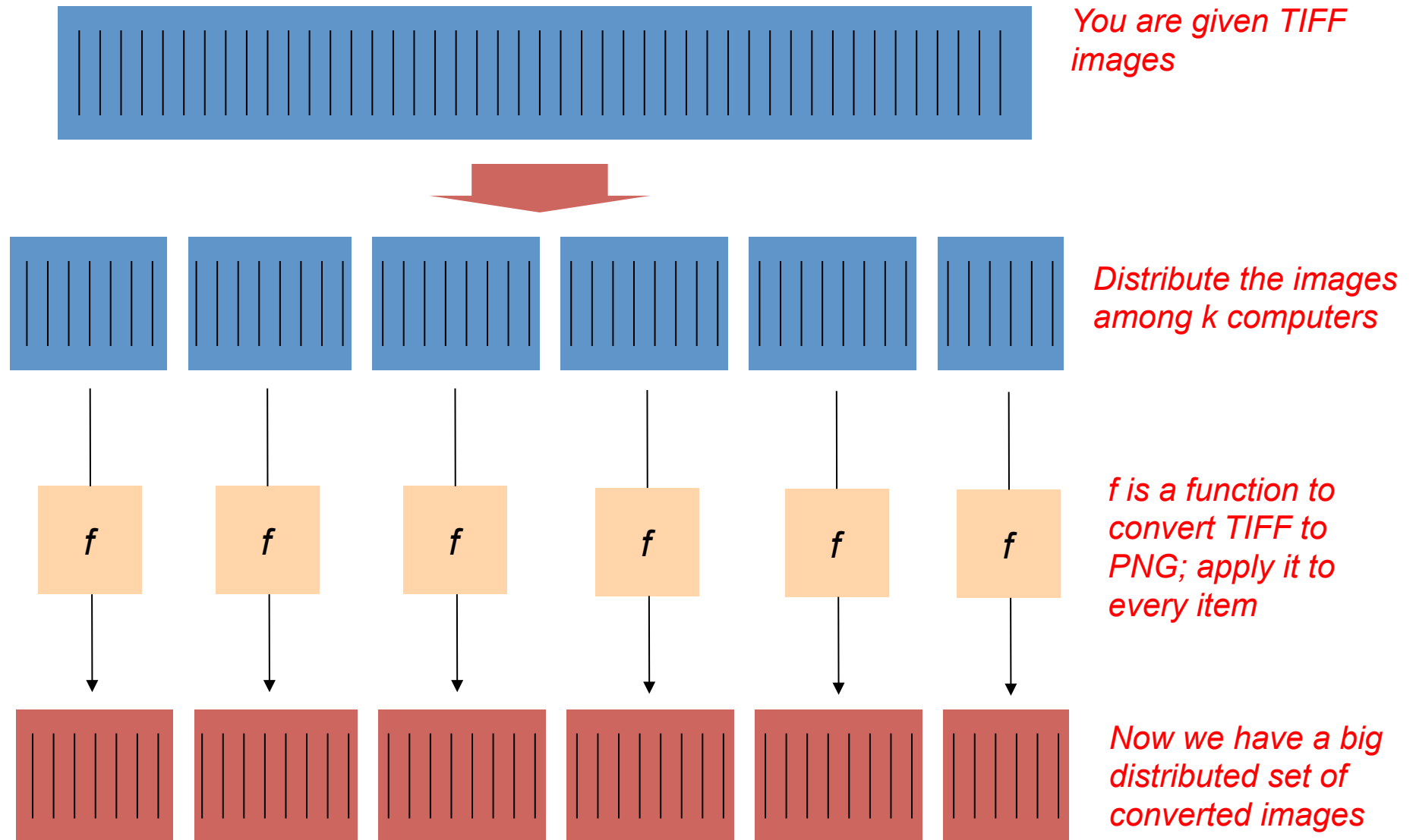
# W

## Schematic of a Parallel “Read Trimming” Task

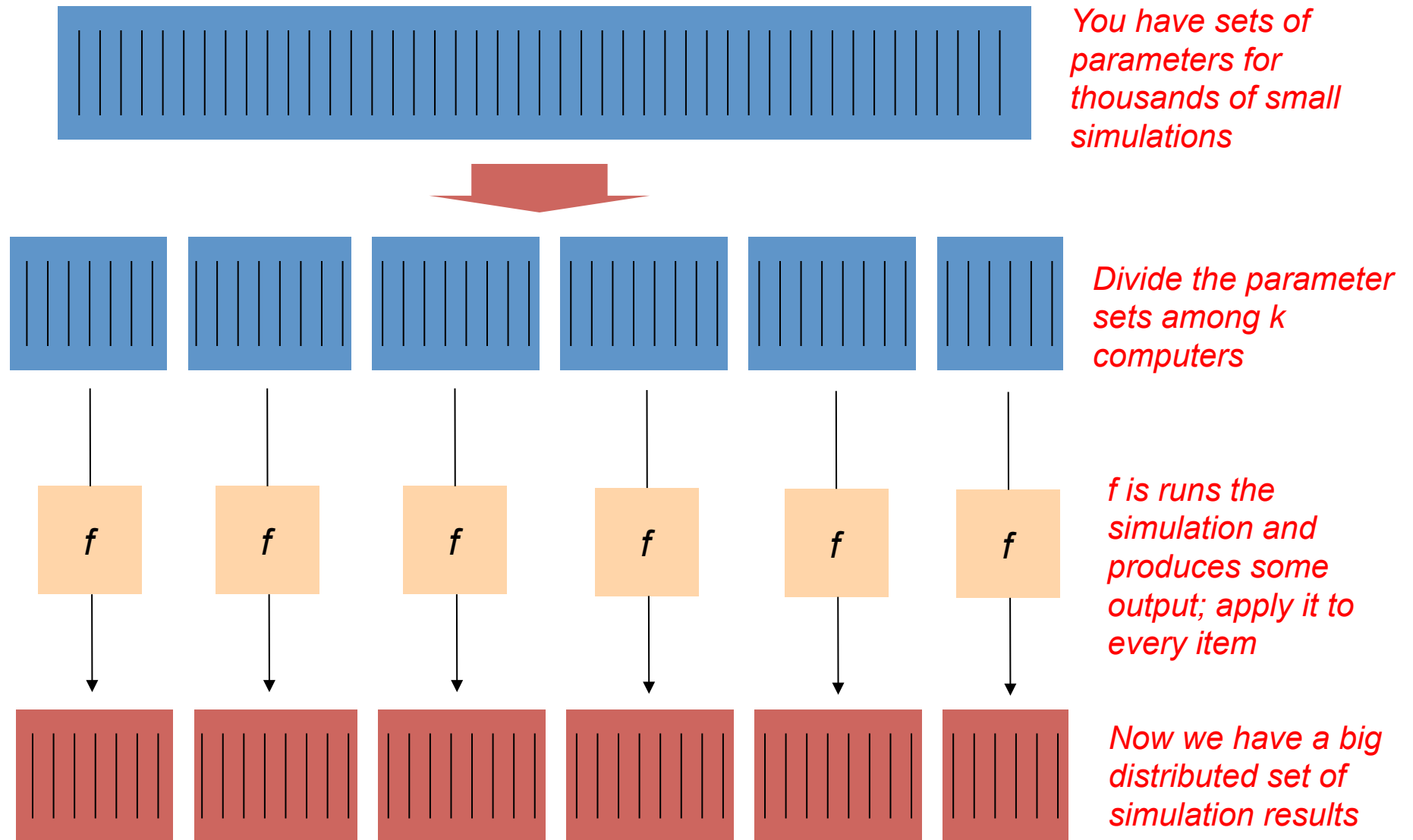


W

## New Task: Convert 405k TIFF images to PNG

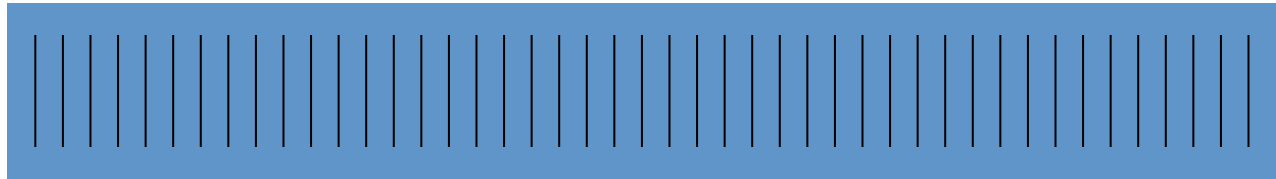


# New Task: Run thousands of simulations

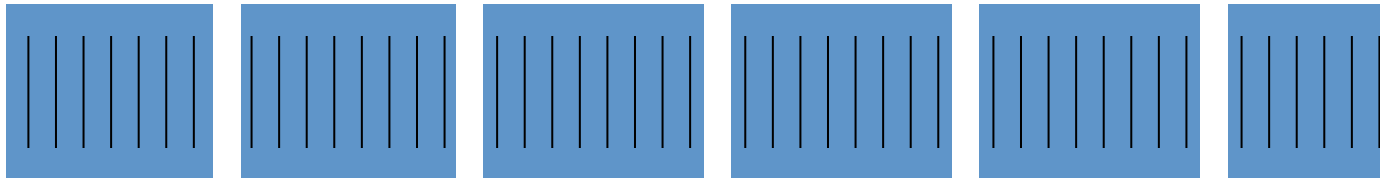


W

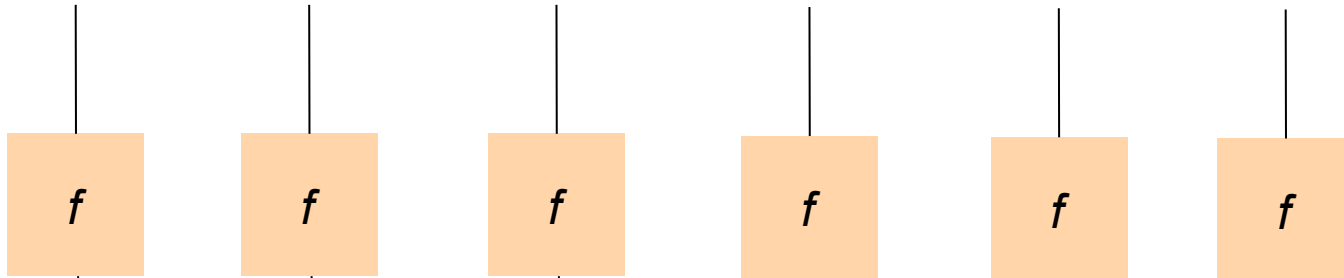
Find the most common word in each document



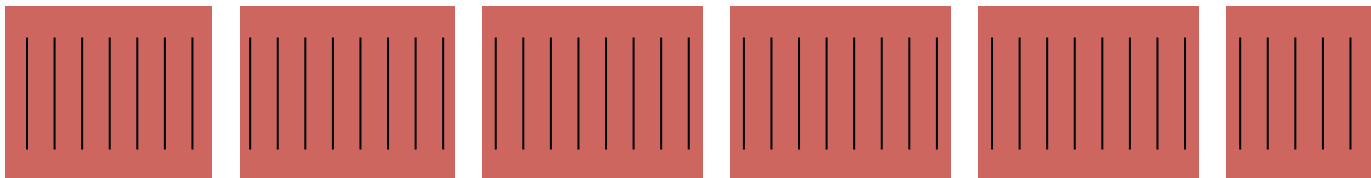
*You have millions of documents*



*Distribute the documents among  $k$  computers*



*$f$  finds the most common word in a single document*



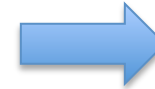
*Now we have a big distributed list of (doc\_id, word) pairs*

## Consider a slightly more general program to compute the word frequency of every word in a single document

### Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled. When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

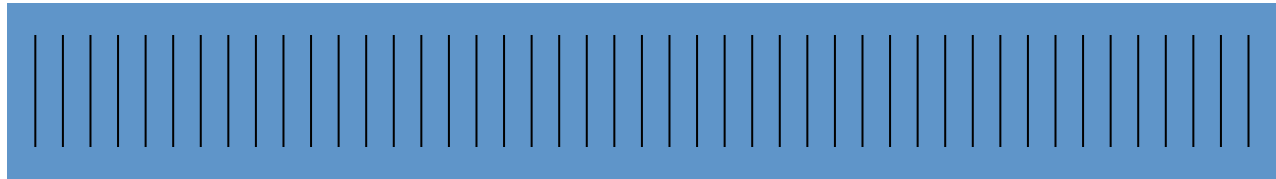


(people, 2)  
(government, 6)  
(assume, 1)  
(history, 2)

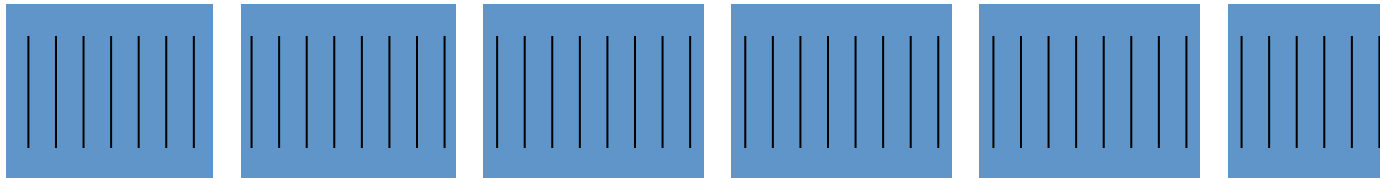
...

W

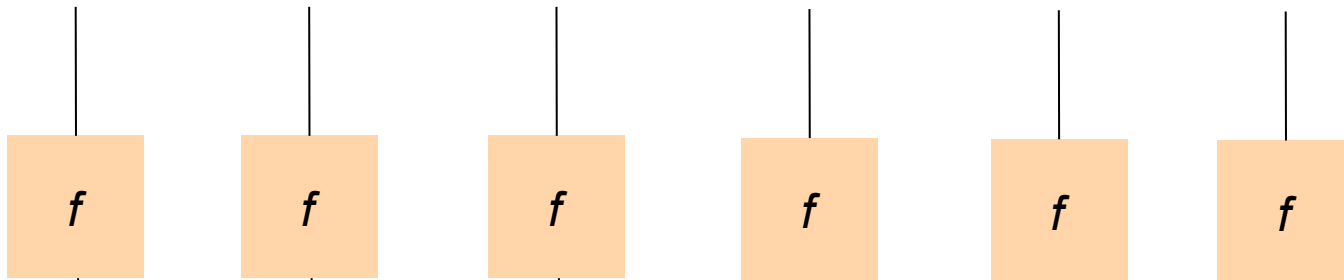
# Compute the word frequency of 5M documents



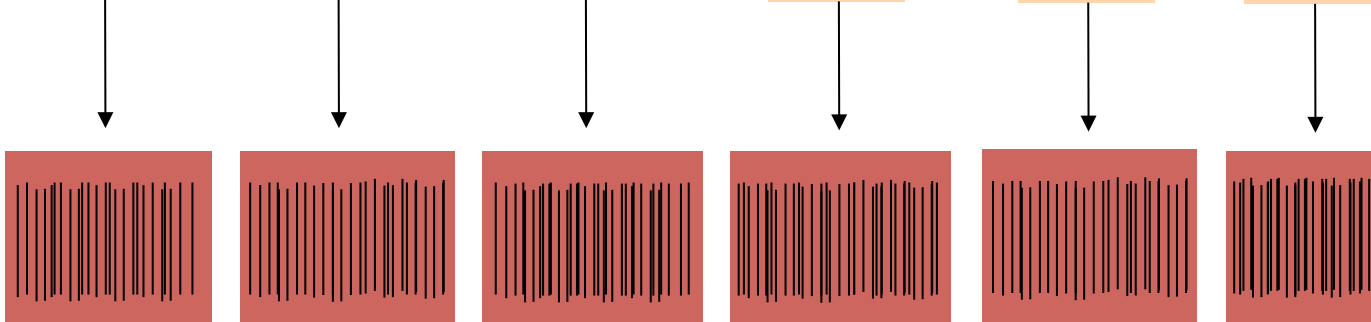
*You have millions of documents*



*Distribute the documents among  $k$  computers*



*For each document  $f$  returns a set of (word, freq) pairs*



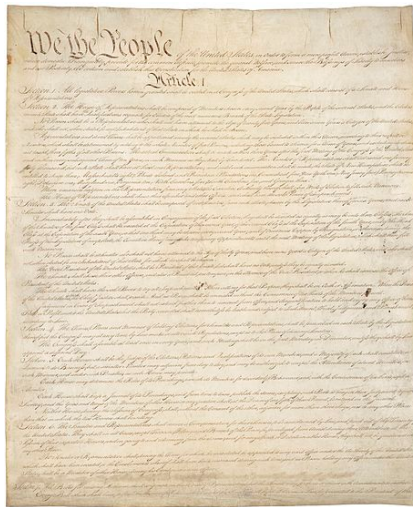
*Now we have a big distributed list of sets of word freqs.*

## There's a pattern here....

- A function that [maps](#) a read to a trimmed read
- A function that [maps](#) a TIFF image to a PNG image
- A function that [maps](#) a set of parameters to a simulation result
- A function that [maps](#) a document to its most common word
- A function that [maps](#) a document to a histogram of word frequencies

W

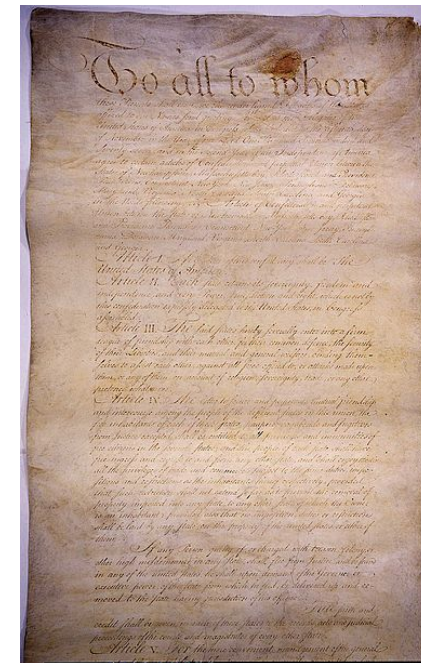
# What if we want to compute the word frequency across *all* documents?



US Constitution



Declaration of Independence

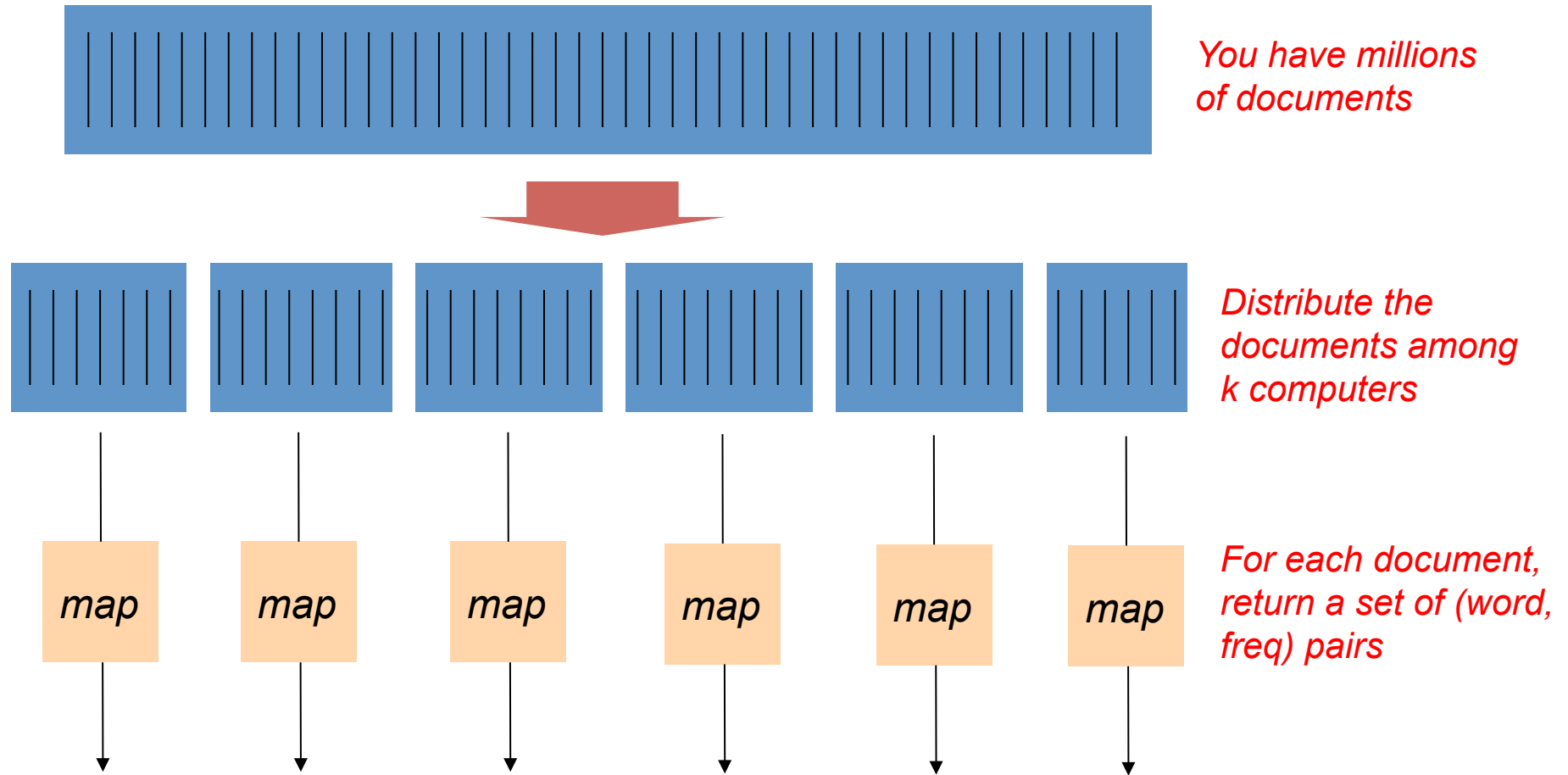


Articles of Confederation

(people, 78)  
(government, 123)  
(assume, 23)  
(history, 38)  
...



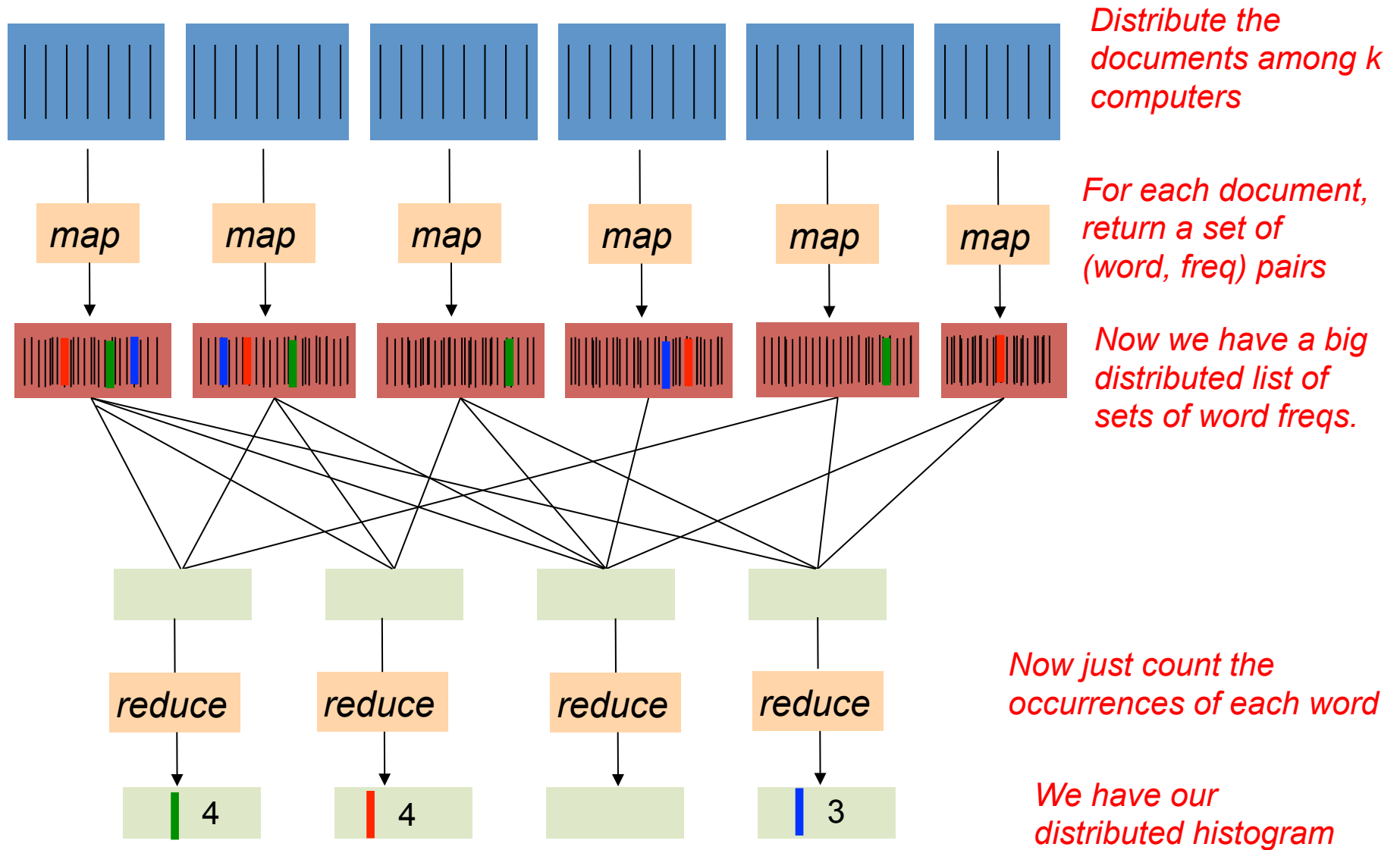
# W Compute the word frequency **across** 5M documents



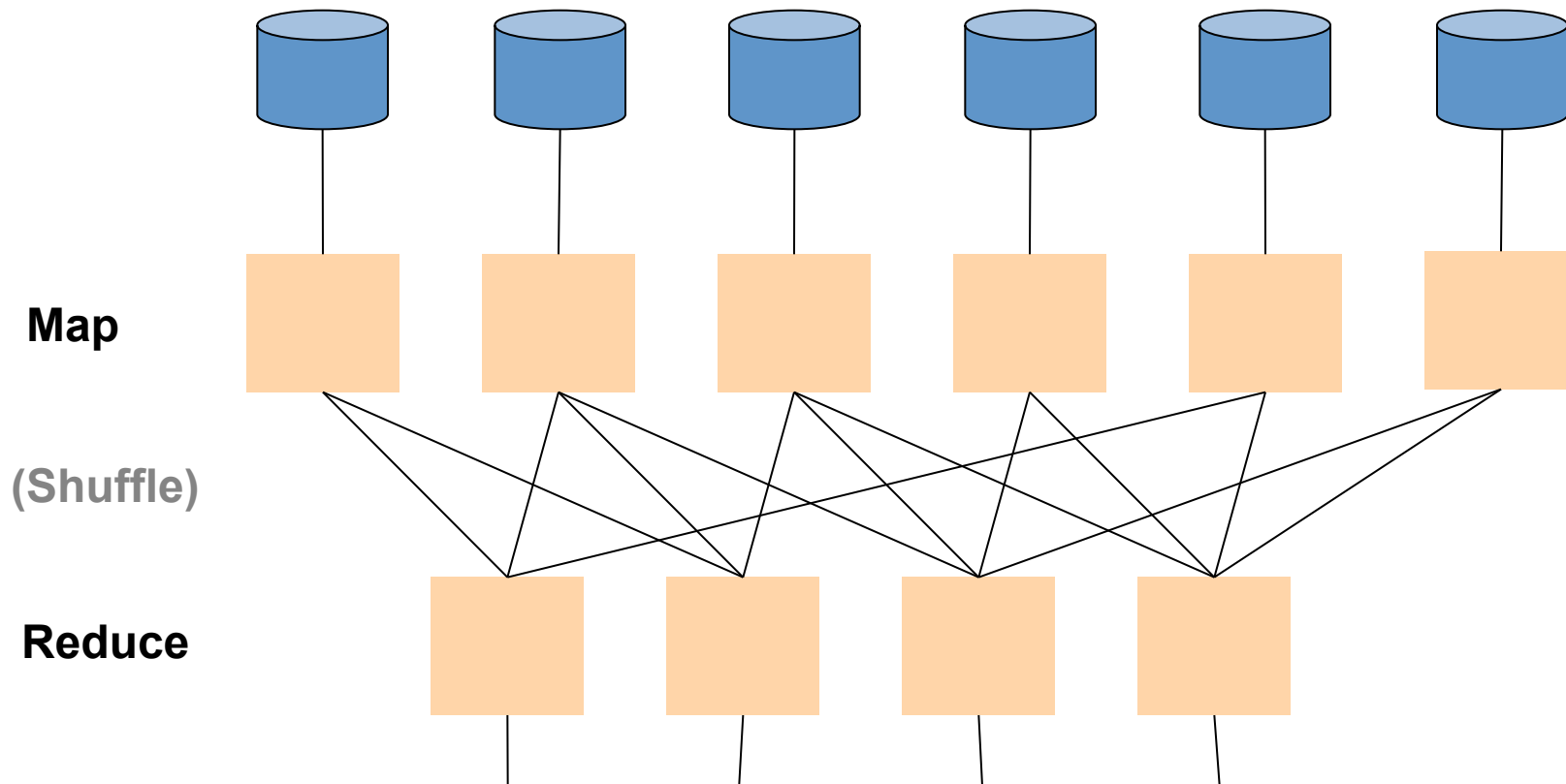
*Now what?*

- *But we don't want a bunch of little histograms – we want one big histogram.*
- *How can we make sure that a single computer has access to every occurrence of a given word regardless of which document it appeared in?*

# W Compute the word frequency **across** 5M documents



## Some distributed algorithm...



# MapReduce Programming Model

- Input & Output: each a set of key/value pairs
- Programmer specifies two functions:

**map (in\_key, in\_value) → list(out\_key, intermediate\_value)**

- Processes input key/value pair
- Produces set of intermediate pairs

**reduce (out\_key, list(intermediate\_value)) → list(out\_value)**

- Combines all intermediate values for a particular key
- Produces a set of merged output values (usually just one)

*Inspired by primitives from functional programming languages such as Lisp, Scheme, and Haskell*

## Example: What does this do?

```
map(String input_key, String input_value):
```

```
// input_key: document name
```

```
// input_value: document contents
```

```
for each word w in input_value:
```

```
EmitIntermediate(w, 1);
```

```
reduce(String output_key, Iterator intermediate_values):
```

```
// output_key: word
```

```
// output_values: ????
```

```
int result = 0;
```

```
for each v in intermediate_values:
```

```
result += v;
```

```
Emit(result);
```

Poll here:

<https://pollev.com/billhowe>

slide source: Google, Inc.

# Example: Document Processing

## Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled. When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying its foundation on such principles and organizing its power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

# Example: Word length histogram

## Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled. When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying its foundation on such principles and organizing its power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

How many “big”, “medium”, and “small” words are used?

# Example: Word length

Abridged Declaration of Independence

## histogram

Big = Yellow = 10+ letters

Medium = Red = 5..9 letters

Small = Blue = 2..4 letters

Tiny = Pink = 1 letter

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.



# Example: Word length histogram

Split the document into chunks and process each chunk on a different computer

Chunk 1

## Abridged Declaration of Independence

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

Chunk 2

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

# Example: Word length histogram

## Abridged Declaration of Independence

Map Task 1  
(204 words)

A Declaration By the Representatives of the United States of America, in General Congress Assembled.  
When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.  
We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying it's foundation on such principles and organizing it's power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

(key, value)

(yellow, 17)  
(red, 77)  
(blue, 107)  
(pink, 3)

Map Task 2  
(190 words)

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government. the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

(yellow, 20)  
(red, 71)  
(blue, 93)  
(pink, 6 )

# Example: Word length histogram

“Shuffle step”

## Map task 1

A Declaration By the Representatives of the United States of America, in General Congress Assembled.

When in the course of human events it becomes necessary for a people to advance from that subordination in which they have hitherto remained, and to assume among powers of the earth the equal and independent station to which the laws of nature and of nature's god entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the change.

We hold these truths to be self-evident; that all men are created equal and independent; that from that equal creation they derive rights inherent and inalienable, among which are the preservation of life, and liberty, and the pursuit of happiness; that to secure these ends, governments are instituted among men, deriving their just power from the consent of the governed; that whenever any form of government shall become destructive of these ends, it is the right of the people to alter or to abolish it, and to institute new government, laying its foundation on such principles and organizing its power in such form, as to them shall seem most likely to effect their safety and happiness. Prudence indeed will

(yellow, 17)

(red, 77)

(blue, 107)

(pink, 3)

## Map task 2

dictate that governments long established should not be changed for light and transient causes: and accordingly all experience hath shewn that mankind are more disposed to suffer while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, begun at a distinguished period, and pursuing invariably the same object, evinces a design to reduce them to arbitrary power, it is their right, it is their duty, to throw off such government and to provide new guards for future security. Such has been the patient sufferings of the colonies; and such is now the necessity which constrains them to expunge their former systems of government, the history of his present majesty is a history of unremitting injuries and usurpations, among which no one fact stands single or solitary to contradict the uniform tenor of the rest, all of which have in direct object the establishment of an absolute tyranny over these states. To prove this, let facts be submitted to a candid world, for the truth of which we pledge a faith yet unsullied by falsehood.

(yellow, 20)

(red, 71)

(blue, 93)

(pink, 6)

## Reduce tasks

(yellow, 17)

(yellow, 20)

(yellow, 37)

(red, 77)

(red, 71)

(red, 148)

(blue, 93)

(blue, 107)

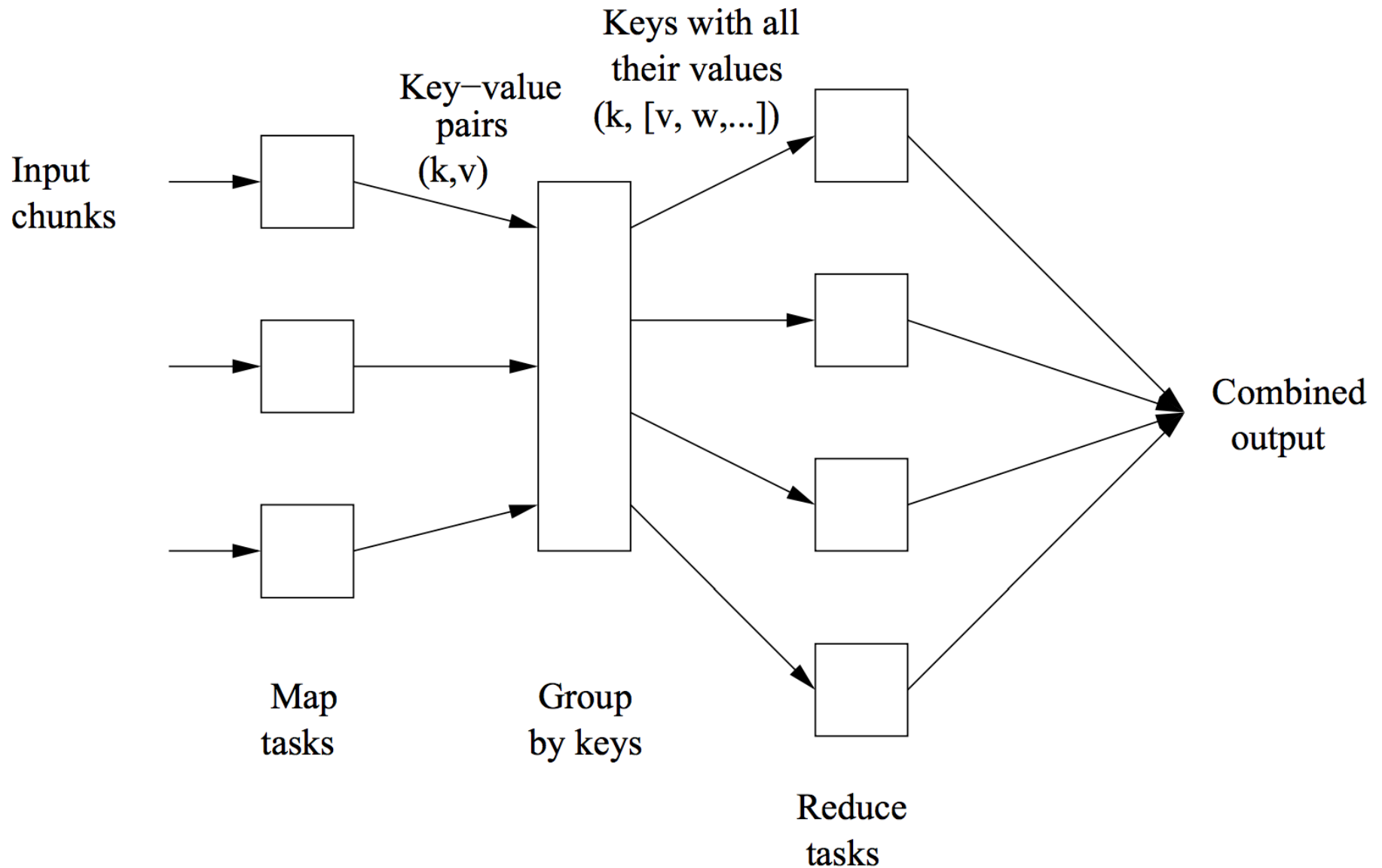
(blue, 200)

(pink, 6)

(pink, 3)

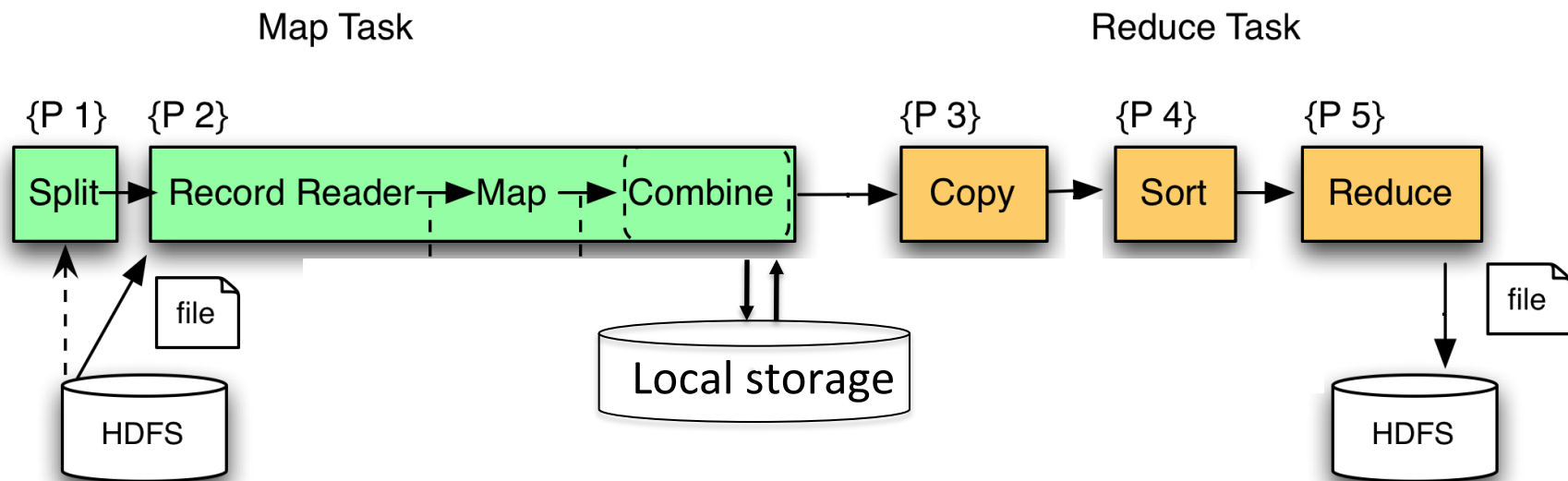
(pink, 9)

## MapReduce computation



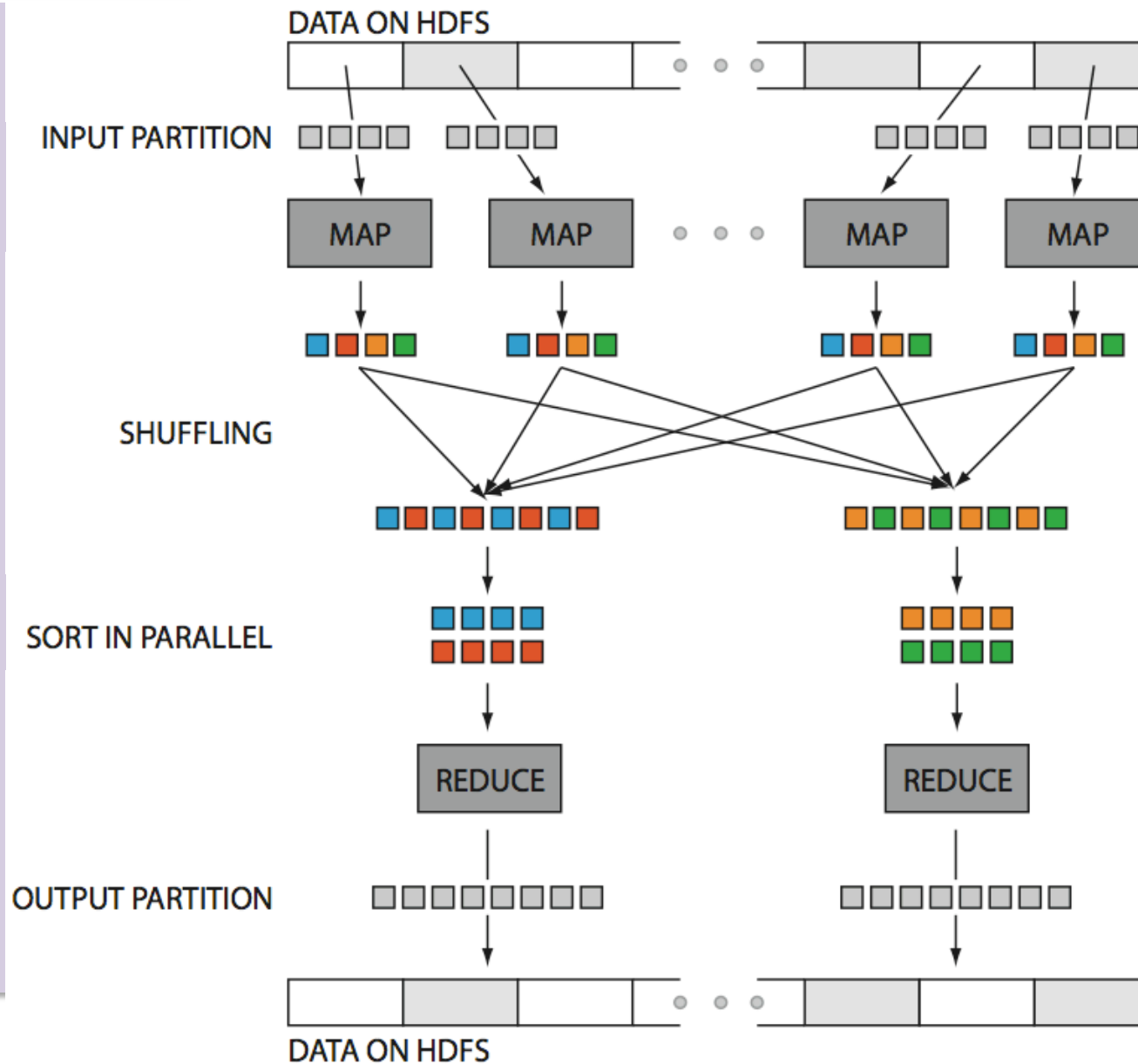
# MR Phases

- Each Map and Reduce task has multiple phases:



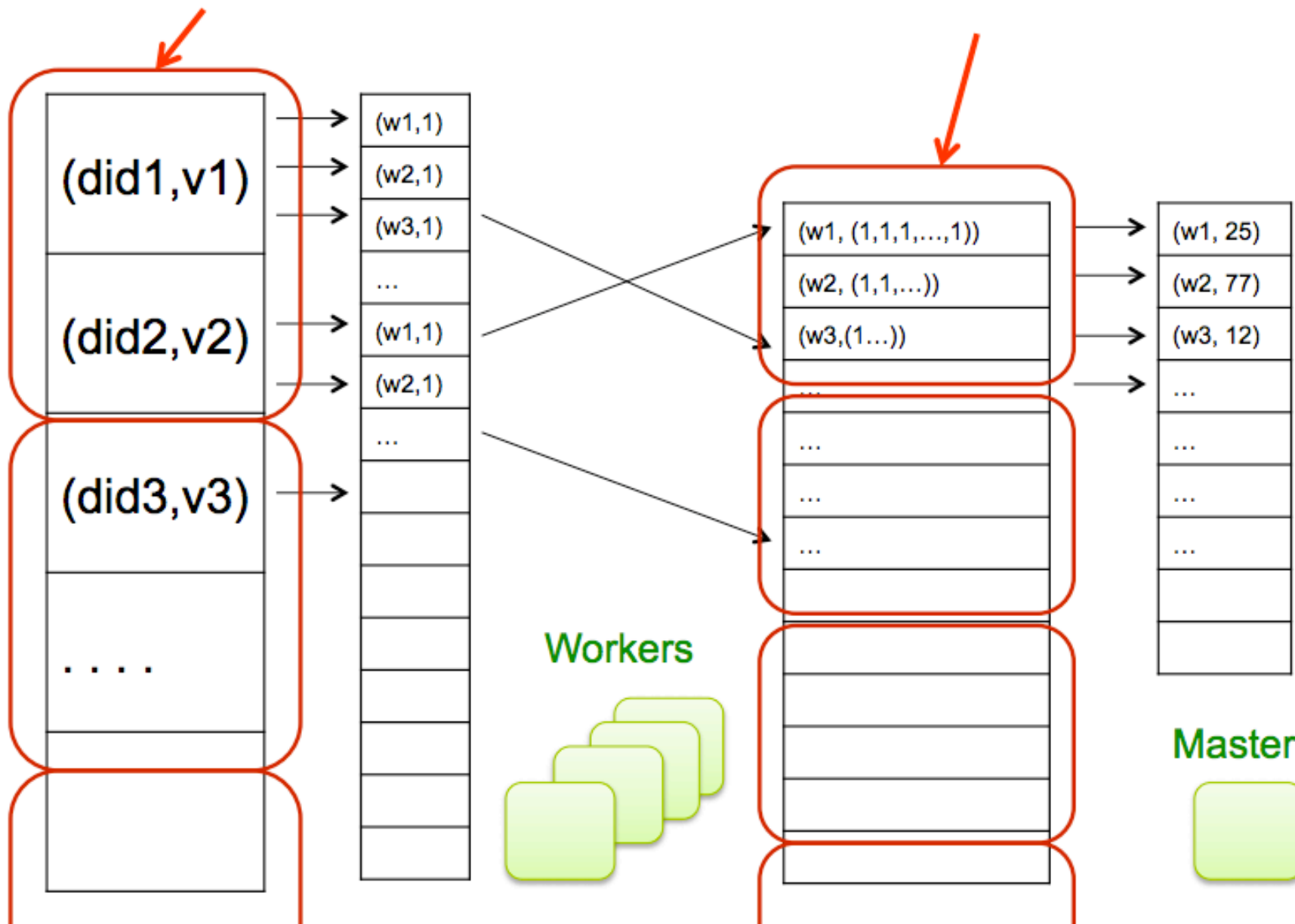
# Hadoop in One Slide

src: Huy Vo,  
NYU Poly



## MAP Tasks

## REDUCE Tasks



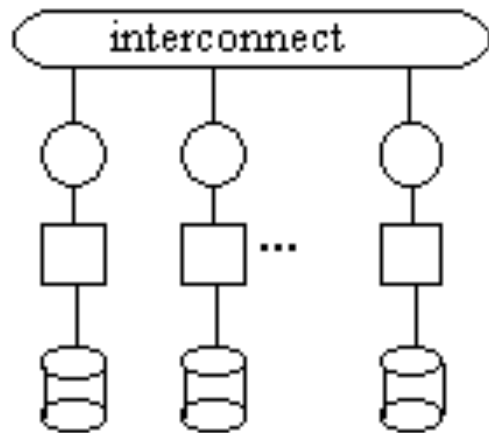
# Implementation

- There is one master node
- Master partitions input file into *M splits*, by key
- Master assigns *workers* (=servers) to the *M map tasks*, keeps track of their progress
- Workers write their output to local disk, partition into *R regions*
- Master assigns workers to the *R reduce tasks*
- Reduce workers read regions from the map workers' local disks

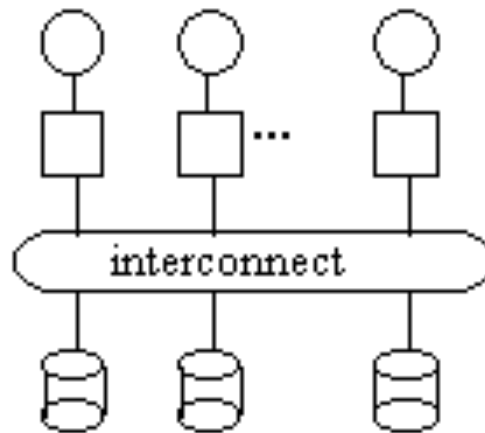


# Taxonomy of Parallel Architectures

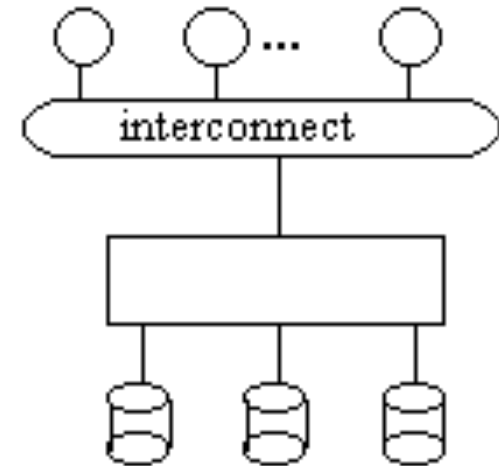
Scales to 1000s of computers






a) shared nothing



b) shared disc



c) shared memory

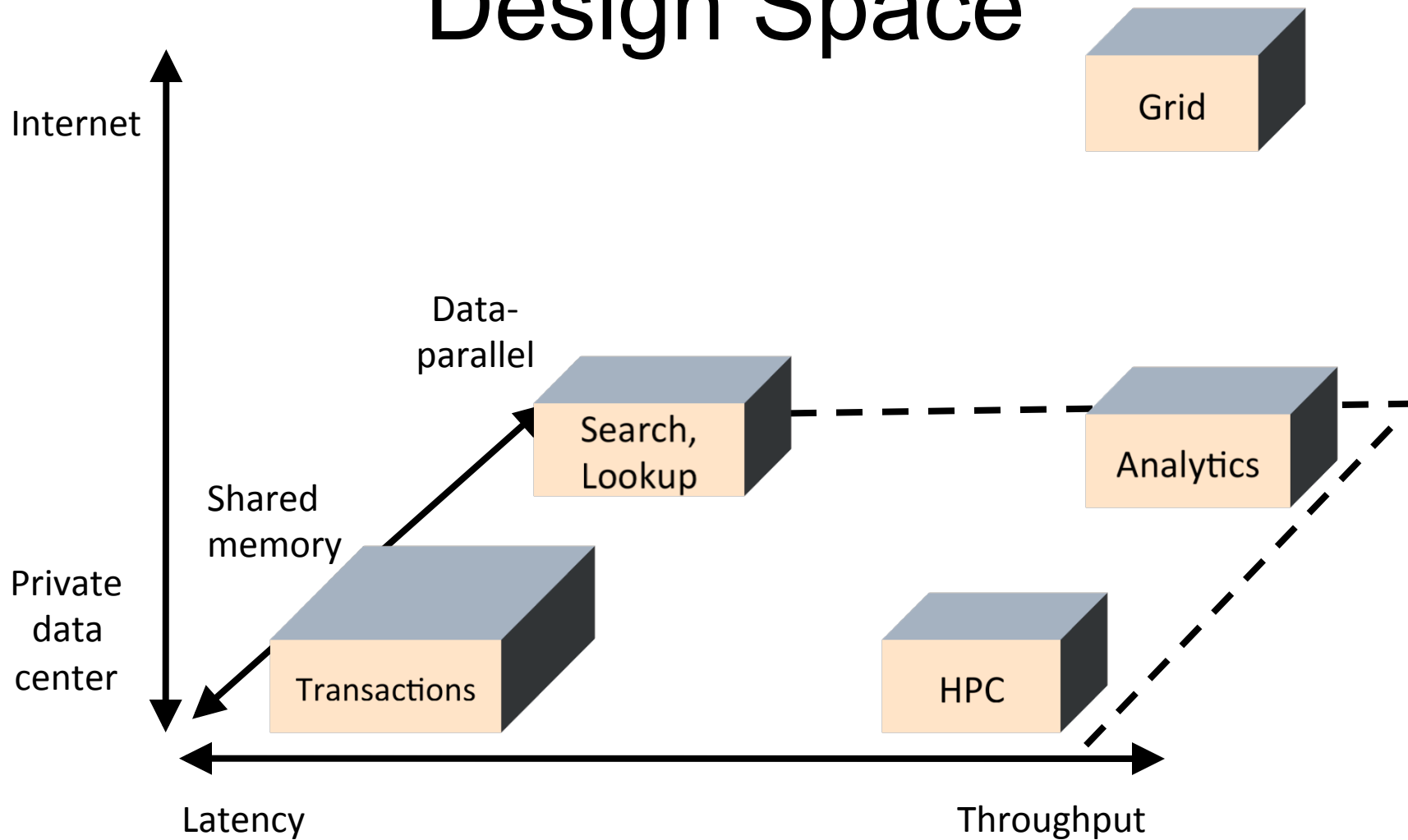
 = disc       = memory       = processor

Easiest to program,

but \$\$\$\$

Fig. 3.1 Logical multi-processor database designs (diagram after [DEWI92])

# Design Space



## Large-Scale Data Processing

- Many tasks process big data, produce big data
- Want to use hundreds or thousands of CPUs
  - ... but this needs to be easy
  - **Parallel databases** exist, but they are expensive, difficult to set up, and do not necessarily scale to hundreds of nodes.
- MapReduce is a *lightweight* framework, providing:
  - **Automatic parallelization and distribution**
  - **Fault-tolerance**
  - **I/O scheduling**
  - **Status and monitoring**