

# Distributed Systems

CS425/ECE428

*Instructor: Radhika Mittal*

*Acknowledgements for some of the materials: Indy Gupta, Nikita Borisov*

# Our agenda for the next 2-3 classes

- Brief overview of key-value stores
- Distributed Hash Tables
  - Peer-to-peer protocol for efficient insertion and retrieval of key-value pairs.
- Key-value stores in the cloud
  - How to run large-scale distributed computations over key-value stores?
    - Map-Reduce Programming Abstraction
    - Job scheduling
  - How to design a large-scale distributed key-value store?
    - Case-study: Facebook's Cassandra

# Many Cloud Providers

- AWS: Amazon Web Services
  - EC2: Elastic Compute Cloud
  - S3: Simple Storage Service
- Microsoft Azure
- Google Cloud/Compute Engine/AppEngine
- Rightscale, Salesforce, EMC, Gigaspaces, I0gen, Datastax, Oracle, VMWare, Yahoo, Cloudera
- And many many more!

# What is a cloud?

- Cloud = Lots of storage + compute cycles nearby



- Cloud services provide:
  - managed *clusters* for distributed computing.
  - managed *distributed datastores*.

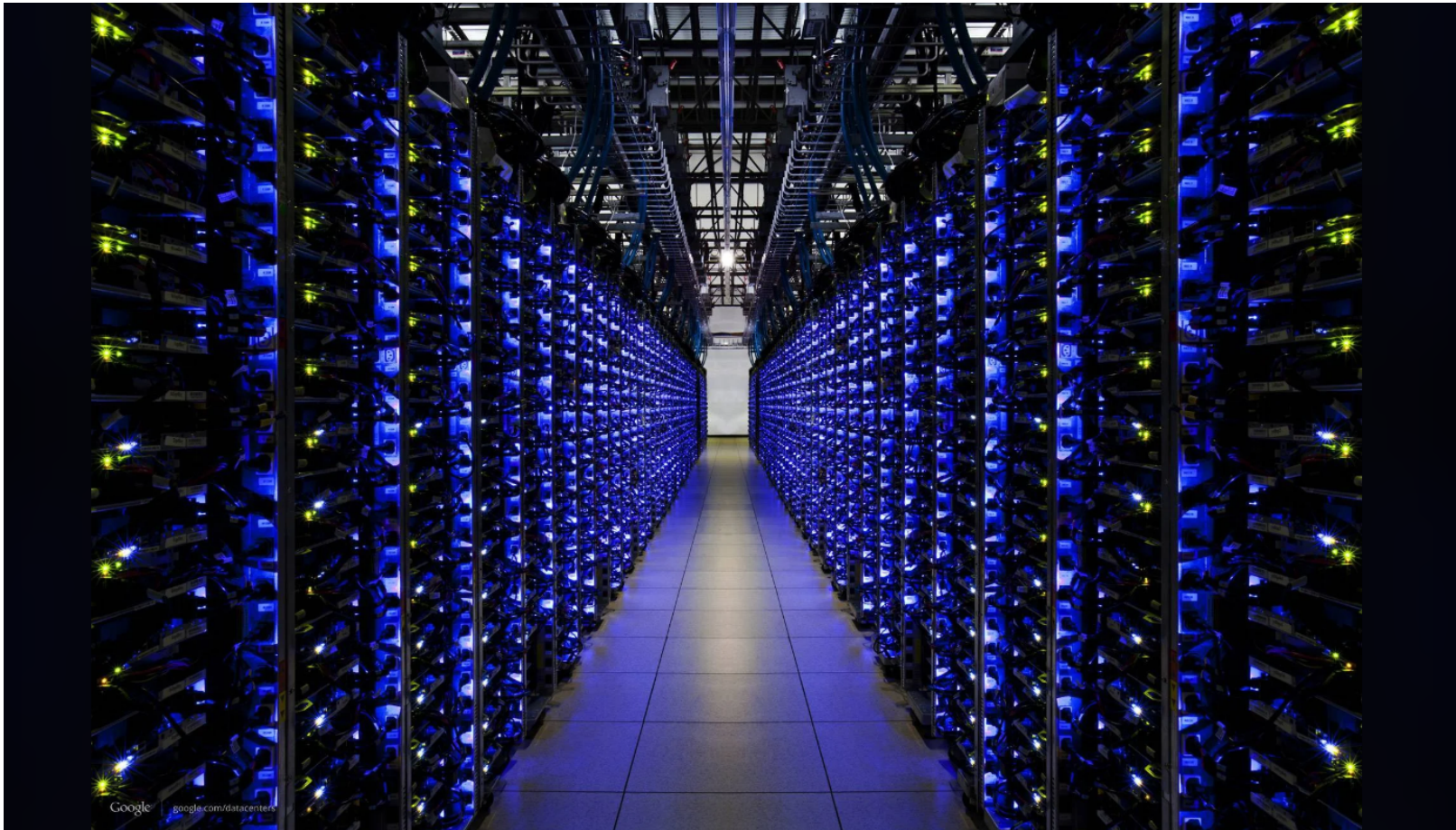
# What is a cloud?

- A single cloud-site (aka “Datacenter”) consists of
  - Compute nodes (grouped into racks) (2)
  - Switches, connecting the racks in a hierarchical network topology.
  - Storage (backend) nodes connected to the network (3)
  - Front-end for submitting jobs and receiving client requests (1)
  - (1-3: Often called “three-tier architecture”)
- A geographically distributed cloud consists of
  - Multiple such sites
  - Each site perhaps with a different structure and services

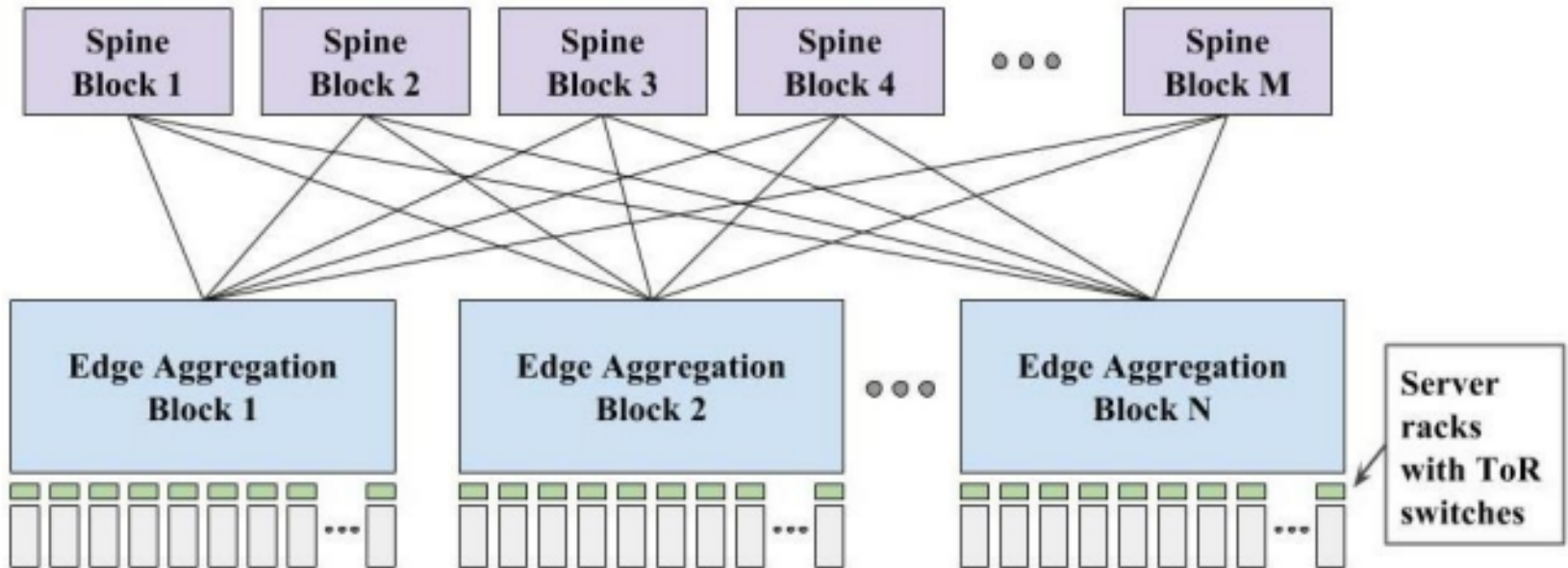
# Picture of Google's Datacenter



# Picture of Google's Datacenter



# Typical Datacenter Topology



# Features of cloud

## I. Massive scale.

- Tens of thousands of servers and cloud tenants, and hundreds of thousands of VMs.

## II. On-demand access:

- Pay-as-you-go, no upfront commitment, access to anyone.

## III. Data-intensive nature:

- What was MBs has now become TBs, PBs and XBs.
  - Daily logs, forensics, Web data, distributed training, etc.

# Must deal with immense complexity!

- Fault-tolerance and failure-handling
- Replication and consensus
- Cluster scheduling
  
- How would a cloud user deal with such complexity?
  - **Powerful abstractions and frameworks**
  - Provide **easy-to-use** API to users.
  - Deal with the complexity of distributed computing under the hood.

**MapReduce**  
is one such powerful  
abstraction.

# MapReduce Architecture

- *MapReduce programming abstraction:*
  - Easy to program distributed computing tasks.
- MapReduce programming abstraction offered by multiple open-source *application frameworks*:
  - Handle creation of “map” and “reduce” tasks.
  - e.g. *Hadoop: one of the earliest map-reduce frameworks.*
  - e.g. *Spark: easier API and performance optimizations.*
- Application frameworks use *resource managers*.
  - Deal with the hassle of distributed cluster management.
  - e.g. *Kubernetes, YARN, Mesos, etc.*

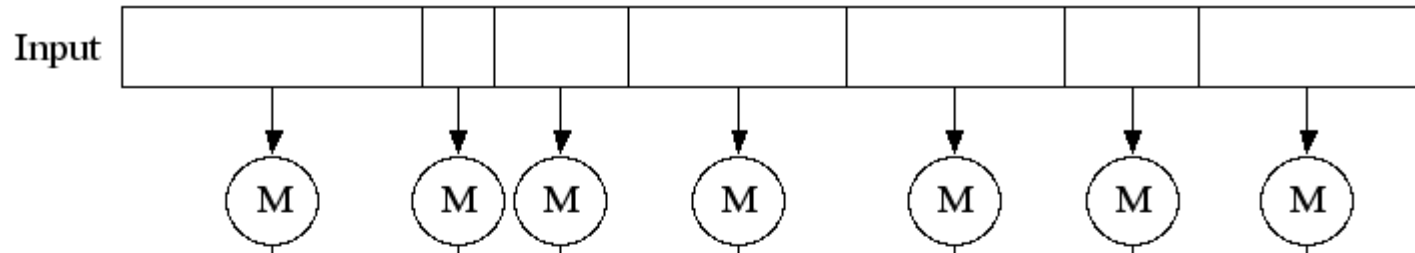
# MapReduce Architecture

- *Map/Reduce abstraction:*
    - Easy to program distributed computing tasks.
  - MapReduce open-source
    - Cre
    - e.g.
    - e.g.
  - Application frameworks use *resource managers*.
    - Deal with the hassle of distributed cluster management.
    - e.g. *Kubernetes, YARN, Mesos, etc.*
- Automatic parallelization & distribution
  - Fault tolerance
  - Scheduling
  - Monitoring & status updates

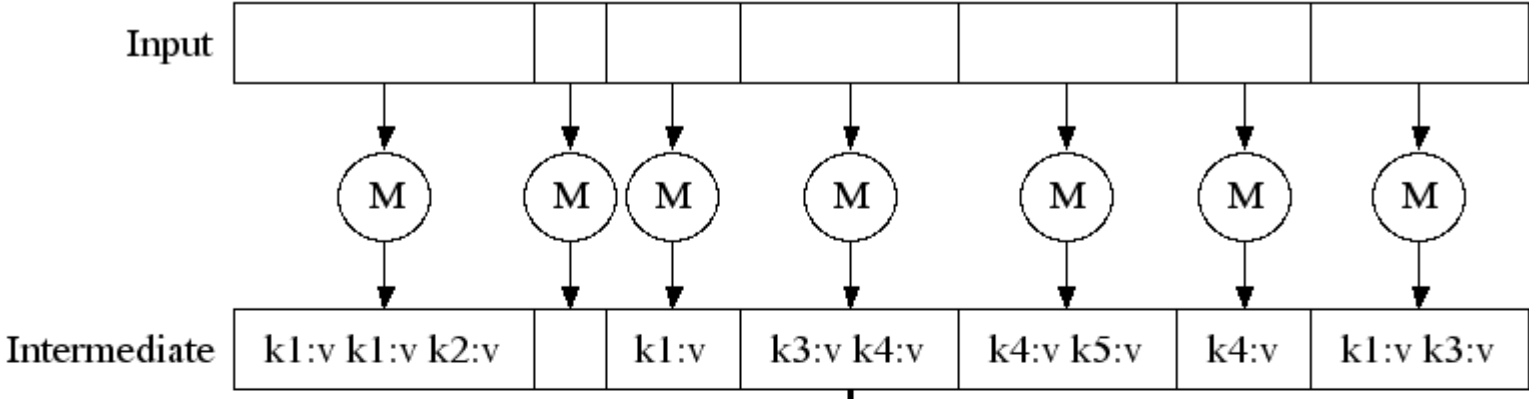
# MapReduce Overview

- Input: a set of key/value pairs  $(k,v)$
- User supplies two functions:
  - $\text{map}(k,v) \rightarrow \text{list}(k1,v1)$
  - $\text{reduce}(k1, \text{list}(v1)) \rightarrow v2$
- $(k1,v1)$  is an intermediate key/value pair.
- Output is the set of  $(k1,v2)$  pairs.

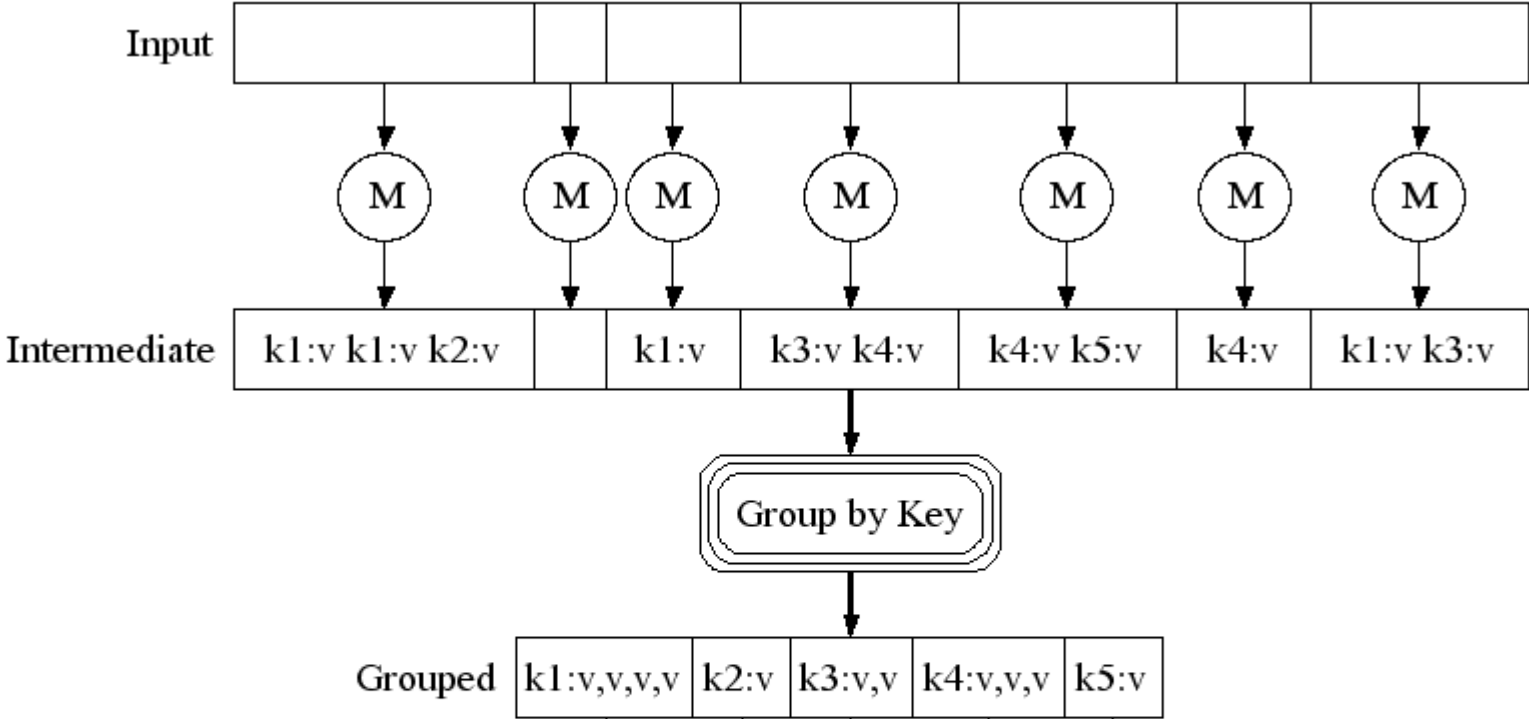
# MapReduce Overview



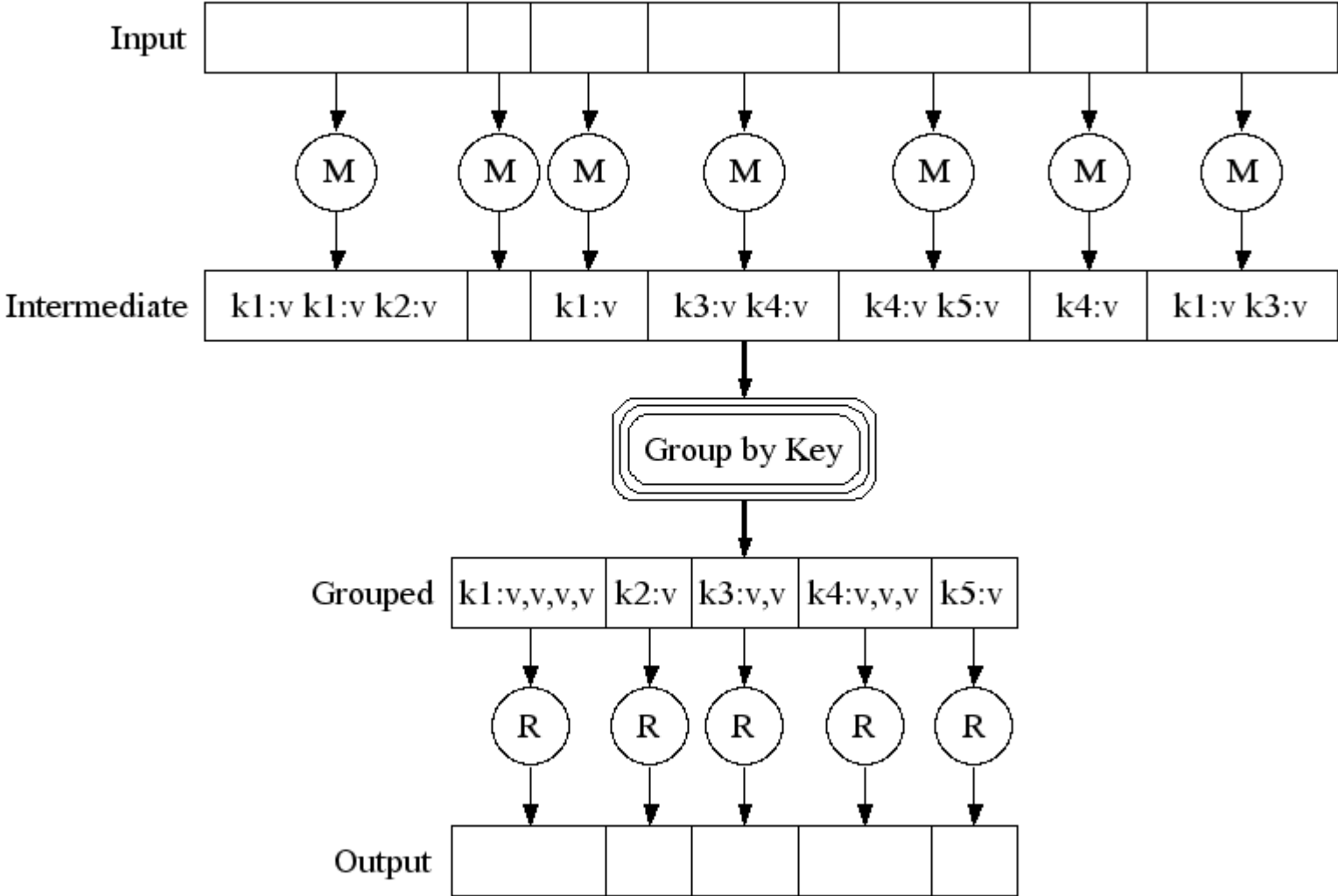
# MapReduce Overview



# MapReduce Overview



# MapReduce Overview

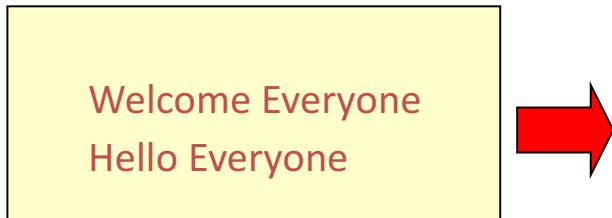


# Typical Example: Word Count

- We have a large file of words containing multiple lines (or records).
- Count the number of times each distinct word appears in the file.
- *Sample application:* analyze web server logs to find popular URLs.

# Map

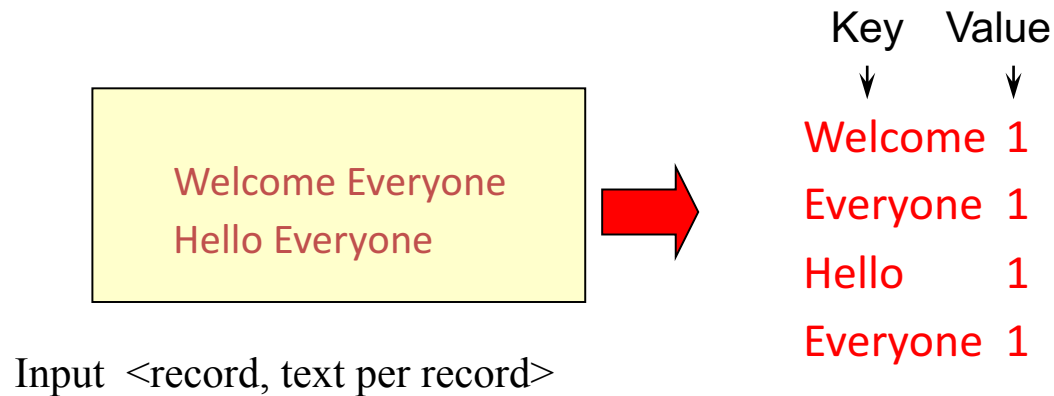
- Process individual records to generate *intermediate key/value pairs*.



Input <record, text per record>

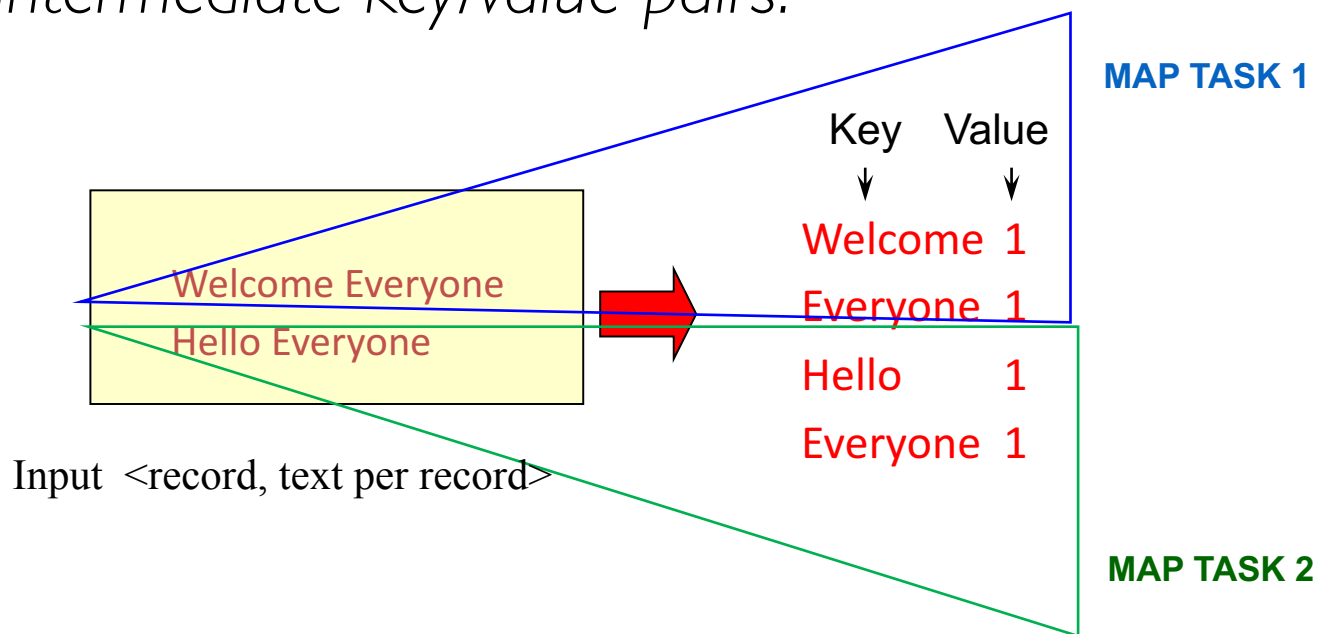
# Map

- Process individual records to generate *intermediate key/value pairs*.



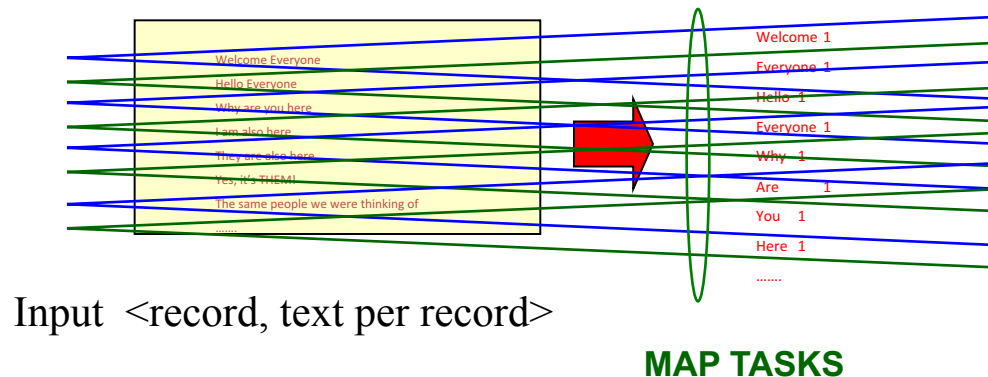
# Map

- **Parallely** process individual records to generate *intermediate key/value pairs*.



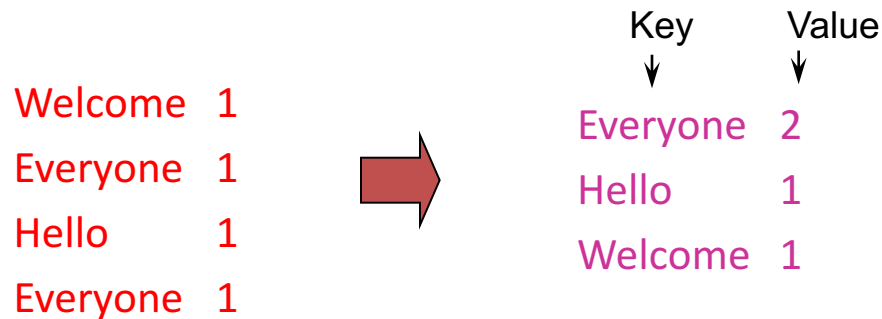
# Map

- **Parallely** process a **large number** of individual records to generate intermediate key/value pairs.



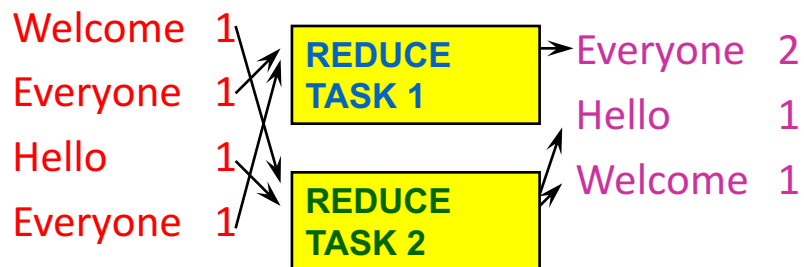
# Reduce

- Processes and merges all intermediate values associated per key.



# Reduce

- Each key assigned to one Reduce task.
- **Parallely** processes and merges all intermediate values partitioned per key.



- Popular: *Hash partitioning*, i.e., key is assigned to
  - $\text{reduce \#} = \text{hash}(\text{key}) \% \text{number of reduce tasks}$

# MapReduce Overview

- Input: a set of key/value pairs
- User supplies two functions:
  - $\text{map}(k,v) \rightarrow \text{list}(k1,v1)$
  - $\text{reduce}(k1, \text{list}(v1)) \rightarrow v2$
- $(k1,v1)$  is an intermediate key/value pair.
- Output is the set of  $(k1,v2)$  pairs.

# MapReduce Overview

- Input: a set of key/value pairs (record, list of words)
- User supplies two functions:
  - $\text{map}(k, v) \rightarrow \text{list}(k1, v1)$
  - $\text{reduce}(k1, \text{list}(v1)) \rightarrow v2$
- $(k1, v1)$  is an intermediate key/value pair. (word, 1)
- Output is the set of  $(k1, v2)$  pairs. (word, count)

# Word Count using MapReduce

```
map(key, value):
```

```
// key: record (line no.); value: list of words in the record
```

```
  for each word w in value:
```

```
    emit(w, 1)
```

```
reduce(key, values):
```

```
// key: a word; values: an iterator over counts
```

```
  result = 0
```

```
  for each count v in values:
```

```
    result += v
```

```
  emit(key, result)
```

# Hadoop Code - Map

```
public static class MapClass extends MapReduceBase

    implements Mapper<LongWritable, Text, Text, IntWritable> {

    private final static IntWritable one =

        new IntWritable(1);

    private Text word = new Text();

    public void map( LongWritable key, Text value,

        OutputCollector<Text, IntWritable> output, Reporter reporter) // key is empty, value is the line

        throws IOException {

        String line = value.toString();

        StringTokenizer itr = new StringTokenizer(line);

        while (itr.hasMoreTokens()) {

            word.set(itr.nextToken());

            output.collect(word, one);

        }

    }

} // Source: http://developer.yahoo.com/hadoop/tutorial/module4.html#wordcount
```

# Hadoop Code - Reduce

```
public static class ReduceClass extends MapReduceBase           implements
Reducer<Text, IntWritable, Text, IntWritable> {

    public void reduce(

        Text key,

        Iterator<IntWritable> values,

        OutputCollector<Text, IntWritable> output,

        Reporter reporter)

            throws IOException {

        // key is word, values is a list of 1's

        int sum = 0;

        while (values.hasNext()) {

            sum += values.next().get();

        }

        output.collect(key, new IntWritable(sum));

    }

} // Source: http://developer.yahoo.com/hadoop/tutorial/module4.html#wordcount
```

# Hadoop Code - Driver

```
// Tells Hadoop how to run your Map-Reduce job

public void run (String inputPath, String outputPath)
    throws Exception {

    // The job. WordCount contains MapClass and Reduce.
    JobConf conf = new JobConf(WordCount.class);

    conf.setJobName("mywordcount");

    // The keys are words
    (strings) conf.setOutputKeyClass(Text.class);

    // The values are counts (ints)
    conf.setOutputValueClass(IntWritable.class);

    conf.setMapperClass (MapClass.class);

    conf.setReducerClass (ReduceClass.class);

    FileInputFormat.addInputPath(
        conf, newPath(inputPath));

    FileOutputFormat.setOutputPath(
        conf, new Path(outputPath));

    JobClient.runJob(conf);
} // Source: http://developer.yahoo.com/hadoop/tutorial/module4.html#wordcount
```

# Spark Code

Python:

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap \
    (lambda line: line.split(" ")) \
    .map(lambda word: (word, 1)) \
    .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```

// Source: <http://spark.apache.org/examples.html>

# More examples: Host size

- Suppose we have a large web corpus
- Metadata file
  - Lines of the form (URL, size, date, ...)
- For each host, find the total number of bytes
  - i.e., the sum of the sizes for all pages from a given host/URL

```
map(key, value):  
// key: metadata record#;  
//value: (URL, size, ...) :  
    for each (URL, size) in value:  
        emit(URL, size)
```

```
reduce(key, values):  
// key: URL, values: iterator over sizes:  
result = 0  
for each size s in values:  
    result += s  
emit(key, result)
```

# More examples: Distributed Grep

- Input: large set of files
- Output: unique lines that match pattern

```
map(key, value):
```

```
// key: file, value: list of lines
```

```
  for each line in value:
```

```
    if "pattern" in line:
```

```
      emit(line, 1)
```

```
reduce(key, values):
```

```
// key: line that matches pattern; values: 1's
```

```
  emit(key, 1)
```

# More examples: Graph reversal

- Input: Web graph: tuples (a, b) where (page a → page b)
- Output: For each page, list of pages that link to it

map(key, value):

// key: source page,

//value: target page

emit(value, key)

reduce(key, values):

// key: target; values: list of pages that link to it.

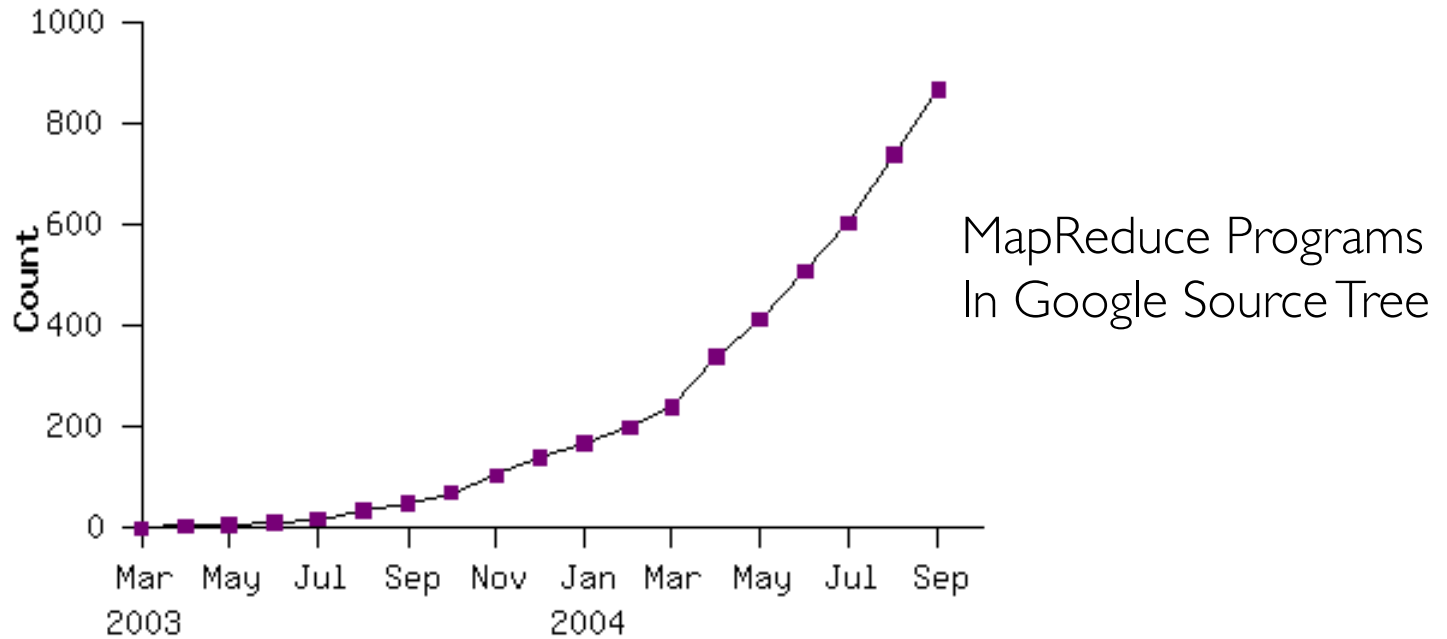
result = concatenate(values)

emit(key, result)

# MapReduce Chains

- map1 -> reduce1 -> map2 -> reduce2
- E.g., output words grouped by their frequency
  - Map1: emit (“word”, 1)
  - Reduce1: emit (“word”, count)
  - Map2: emit (count, “word”)
  - Reduce2: identity, i.e. emit(count, list of words)

# MapReduce is popular and widely applicable



## Example uses:

distributed grep

term-vector / host

document clustering

...

distributed sort

web access log stats

machine learning

...

web link-graph reversal

inverted index construction

statistical machine  
translation

...

# MapReduce Execution

Externally: For user

1. Write a Map program (short), write a Reduce program (short)
2. Specify number of Maps and Reduces (parallelism level)
3. Submit job; wait for result
4. Need to know very little about parallel/distributed programming!

# MapReduce Execution

Internally: For the framework and resource manager in the cloud

1. Parallelize Map
2. Transfer data from Map to Reduce (**shuffle data**)
3. Parallelize Reduce
4. Implement Storage for Map input, Map output, Reduce input, and Reduce output

(Ensure that no Reduce starts before all Maps are finished. That is, ensure the **barrier** between the Map phase and Reduce phase)

# MapReduce Execution

Internally: For the framework and resource manager in the cloud

1. Parallelize Map (*easy!*)
  - Each map task is independent of the other!
2. Transfer data from Map to Reduce (*shuffle data*)
3. Parallelize Reduce
4. Implement Storage for Map input, Map output, Reduce input, and Reduce output

(Ensure that no Reduce starts before all Maps are finished. That is, ensure the *barrier* between the Map phase and Reduce phase)

# MapReduce Execution

Internally: For the framework and resource manager in the cloud

1. Parallelize Map (*easy!*)
2. Transfer data from Map to Reduce (**shuffle data**)
  - All Map output records with same key assigned to same Reduce
  - Use *partitioning function, e.g.,  $\text{hash}(\text{key})\% \text{number of reducers}$*
3. Parallelize Reduce
4. Implement Storage for Map input, Map output, Reduce input, and Reduce output

(Ensure that no Reduce starts before all Maps are finished. That is, ensure the **barrier** between the Map phase and Reduce phase)

# MapReduce Execution

Internally: For the framework and resource manager in the cloud

1. Parallelize Map (*easy!*)
2. Transfer data from Map to Reduce (**shuffle data**)
3. Parallelize Reduce (*easy!*)
  - Each reduce task is independent of the other!
4. Implement Storage for Map input, Map output, Reduce input, and Reduce output

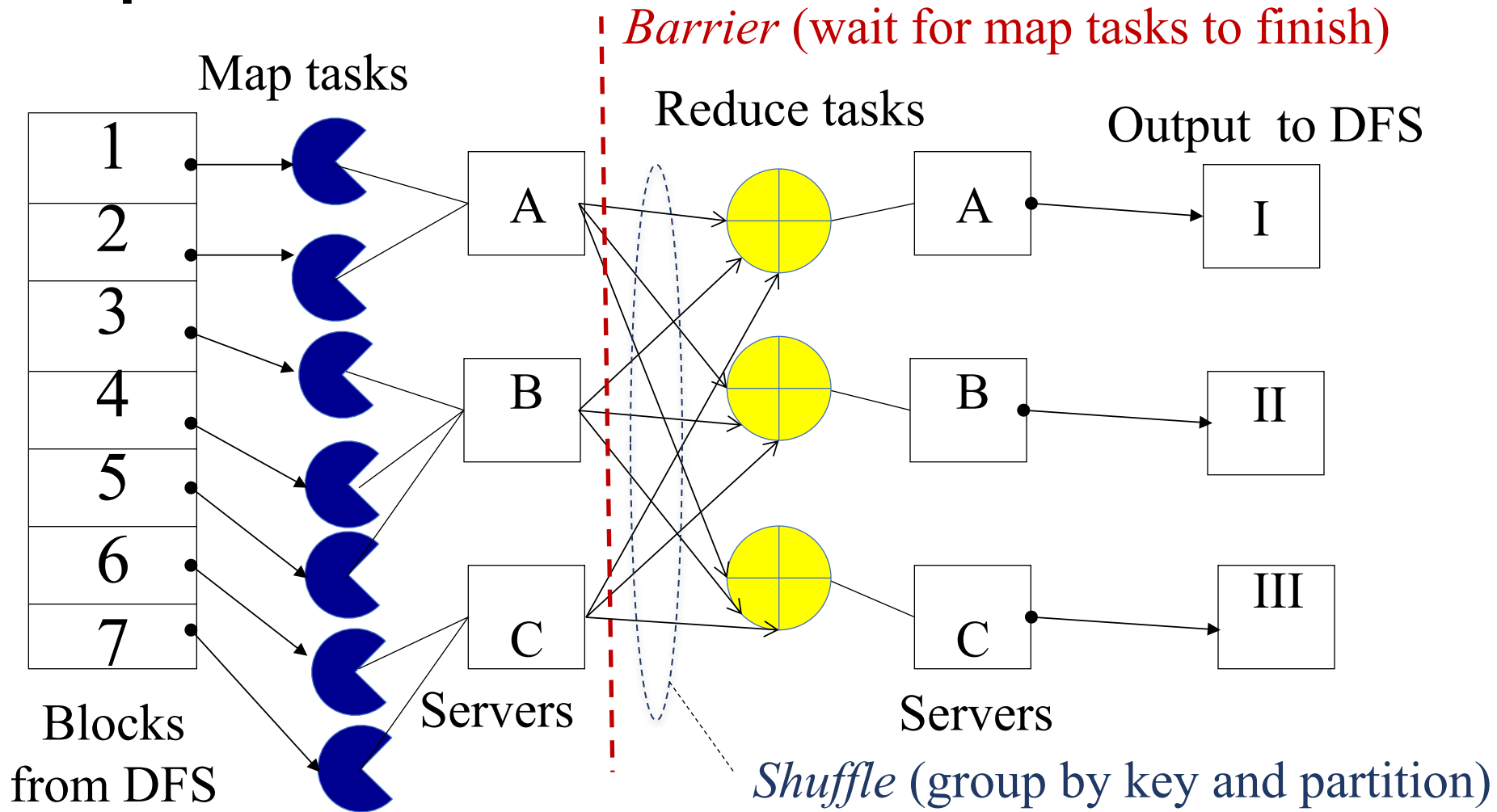
(Ensure that no Reduce starts before all Maps are finished. That is, ensure the **barrier** between the Map phase and Reduce phase)

# MapReduce Execution

Internally: For the framework and resource manager in the cloud

1. Parallelize Map
2. Transfer data from Map to Reduce (**shuffle data**)
3. Parallelize Reduce
4. Implement Storage for Map input, Map output, Reduce input, and Reduce output
  - Map input: from **distributed file system/data store**
  - Map output: to local disk (at Map node); uses **local file system**
  - Reduce input: from (multiple) remote disks; uses **local file systems**
  - Reduce output: to **distributed file system/data store**
    - local file system** (e.g. Linux FS)
    - distributed file system** (e.g. Google File System, Hadoop Distributed File System)
    - distributed data store** (e.g. Cassandra, BigTable, Spanner, DynamoDB)

# MapReduce Execution



Resource Manager (assigns map and reduce tasks to servers)

# Resource Manager

- Examples:
  - *YARN* (Yet Another Resource Negotiator), used underneath Hadoop 2.x +
  - *Kubernetes, Borg, Mesos*, etc.
- Treats each server as a collection of *containers*
  - Container = fixed CPU + fixed memory (e.g. *Docker*)
  - Each task runs in a container.
- Has 3 main components
  - *Global Resource Manager (RM)*: Cluster Scheduling
  - *Per-server Node Manager (NM)*: Daemon and server-specific functions
  - *Per-application (job) Application Master (AM)*
    - Container negotiation with RM and NMs.
    - Handling task failures of that job.

# Fault Tolerance

- NM heartbeats to RM
  - If server fails: RM times out waiting for next heartbeat, RM lets all affected AMs know, and AMs take appropriate action.
- NM keeps track of each task running at its server
  - If task fails while in-progress, mark the task as idle and restart it.
- AM heartbeats to RM
  - On failure, RM restarts AM, which then syncs it up with its running tasks.
- RM Failure
  - Use old checkpoints and bring up secondary RM.

# Slow Servers

Slow tasks are called **Stragglers**.

- The slowest task slows the entire job down (why?)
- Due to bad disk, network bandwidth, CPU, or memory
- Keep track of “progress” of each task (% done)
- Perform proactive backup (replicated) execution of some straggler tasks
  - A task considered done when its first replica complete (other replicas can then be killed).
  - Approach called **Speculative Execution**.
- Straggler mitigation has been a very active area of research.

Barrier at the end  
of Map phase!

# Task Scheduling

- Cluster managers schedule nodes based on multiple criteria (scored by importance).
- E.g. favour data locality:
  - attempts to schedule a map task on a machine that contains a replica of corresponding input data.
  - *if that's not possible*, on the same rack as a machine containing the input.
  - *if that's not possible*, anywhere.
- What does “*if that's not possible*” mean?
  - No more resources available on the machine.
  - Might be worth waiting a while for resources to become available.
    - Delay scheduling in Spark!
- Cluster scheduling is also a very active area of research.

# Summary

- Cloud provides distributed computing infrastructure as a service.
- Running a distributed job on the cloud cluster can be very complex:
  - Must deal with parallelization, scheduling, fault-tolerance, etc.
- MapReduce is a powerful abstraction to hide this complexity.
  - User programming via easy-to-use API.
  - Distributed computing complexity handled by underlying frameworks and resource managers
- Plenty of research work in scheduling, fault-tolerance, and straggler mitigation for MapReduce, especially in late 2010s.

# More general “microservices”

Many modern applications deployed as modular microservices that communicate through RPCs.

Each microservice can have multiple replicas (auto-scaled).

Each replica deployed as a pod in a container (analogous to a task).

Managed using systems like Kubernetes and service meshes (Istio/Envoy).

Active area of research today!

