Hadoop Internals

YongChul Kwon

At Stratosphere

INPUT

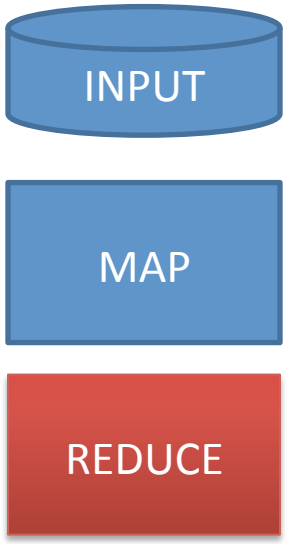
MAP

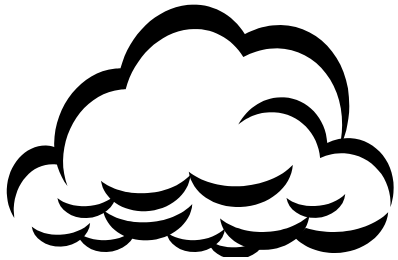
REDUCE

Hadoop

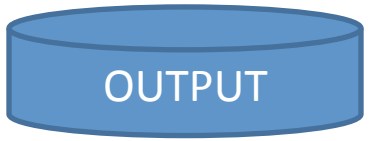
OUTPUT

At Troposphere





Hadoop



At Data Center X

INPUT

MAP

REDUCE

Hadoop

OUTPUT

From Open Compute Project V1.0

Oops. Wrong Plane?

INPUT

MAP

REDUCE

Hadoop

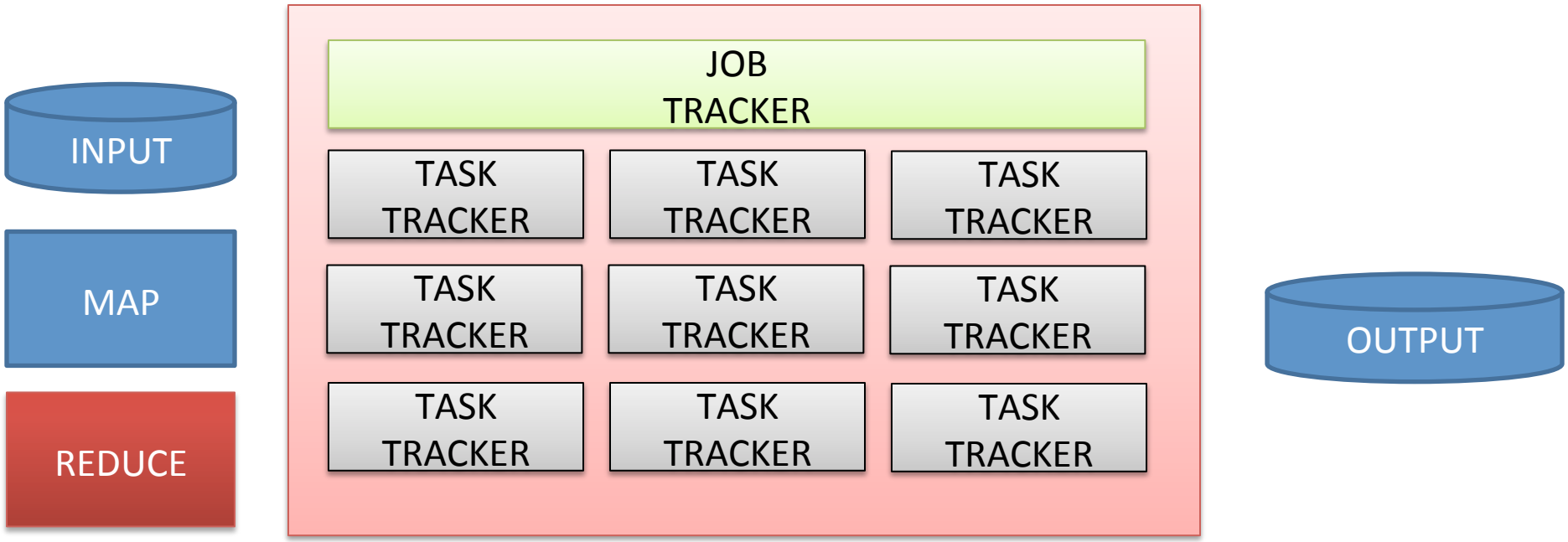
OUTPUT

So,

- Your job is running on many many servers
  - What's going on?

- MapReduce "Job"
  - Execution of your MapReduce program
- MapReduce "Task"
  - Map Task and Reduce Task
  - Distributed execution of your Map and Reduce

At Software Plane



In reality,

INPUT

MAP

REDUCE

Hadoop

OUTPUT

Where's Waldo?

How my job is executed?

JOB TRACKER

TASK TRACKER

Run my Job A!

I'm done with task X
Here's progress of task Y, Z
Anything to run?
BTW, I'm still alive!

X completed. Y, Z are in progress…
Hmm… which task should I assign…
Good. Task 1 of Job A

Dear TaskTracker,
Thank you for your hard working.
Please run Task 1 of Job A

Job Complete!

Job Tracker and Task Tracker

- Job Tracker
  - Governs execution of jobs
  - Task scheduling decision
  - Respond to heartbeat message from task trackers

- Trask Tracker
  - Governs execution of tasks
  - Periodically report the progress of tasks via heartbeat message

How my job is executed?

JOB TRACKER

TASK TRACKER

Heartbeat

I'm done with task X
Here's progress of task Y, Z
Anything to run?
BTW, I'm still alive!

Run my Job A!

Book keeping
Scheduling decision

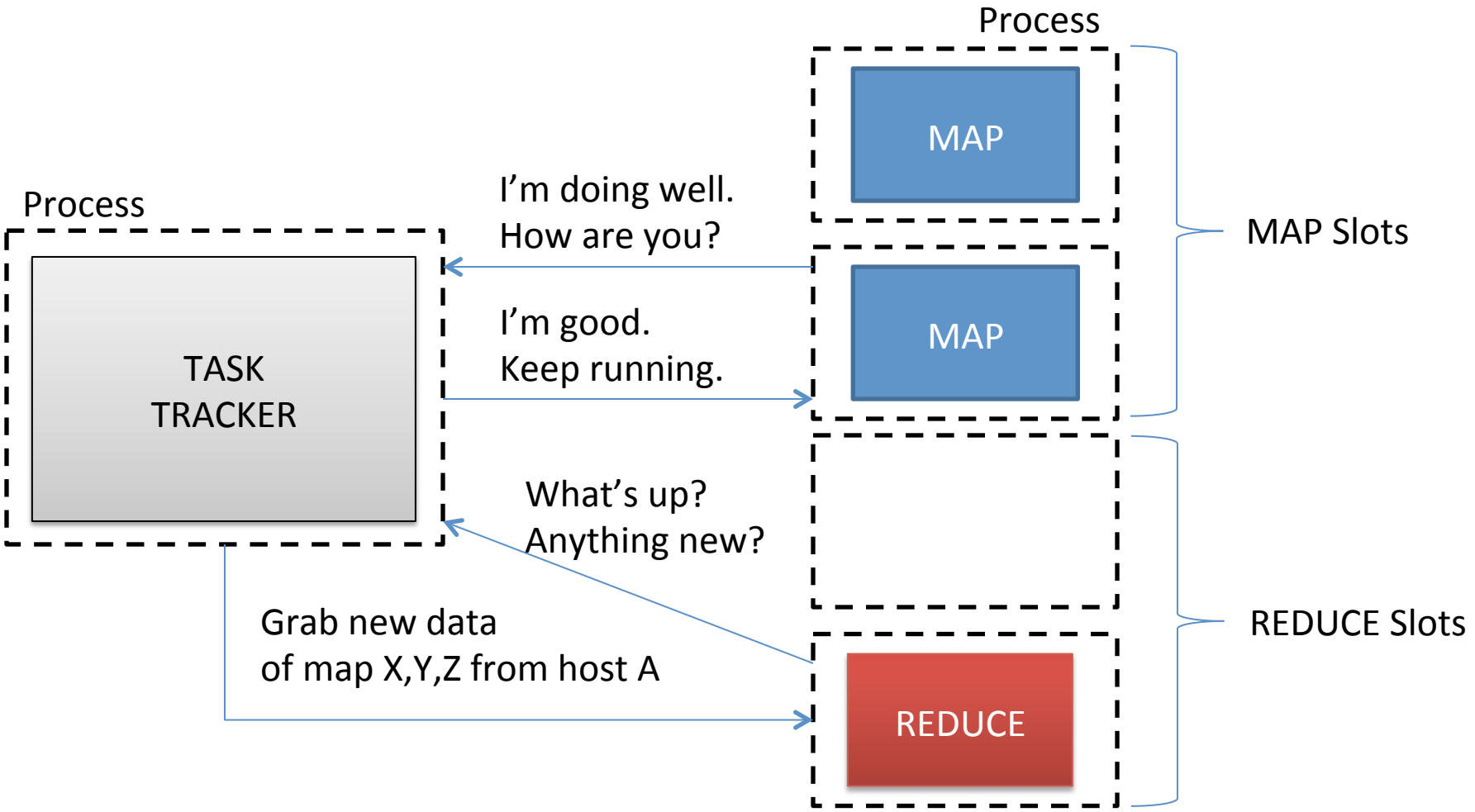
X completed. Y, Z are in progress…
Hmm… which task should I assign…
Good. Task 1 of Job A

Dear TaskTracker,
Thank you for your hard working.
Please run Task 1 of Job A

Heartbeat Response

Job Complete!

Task Tracker



Map Task

- Action
  - While hasMore():
  - read a record
  - process – map()

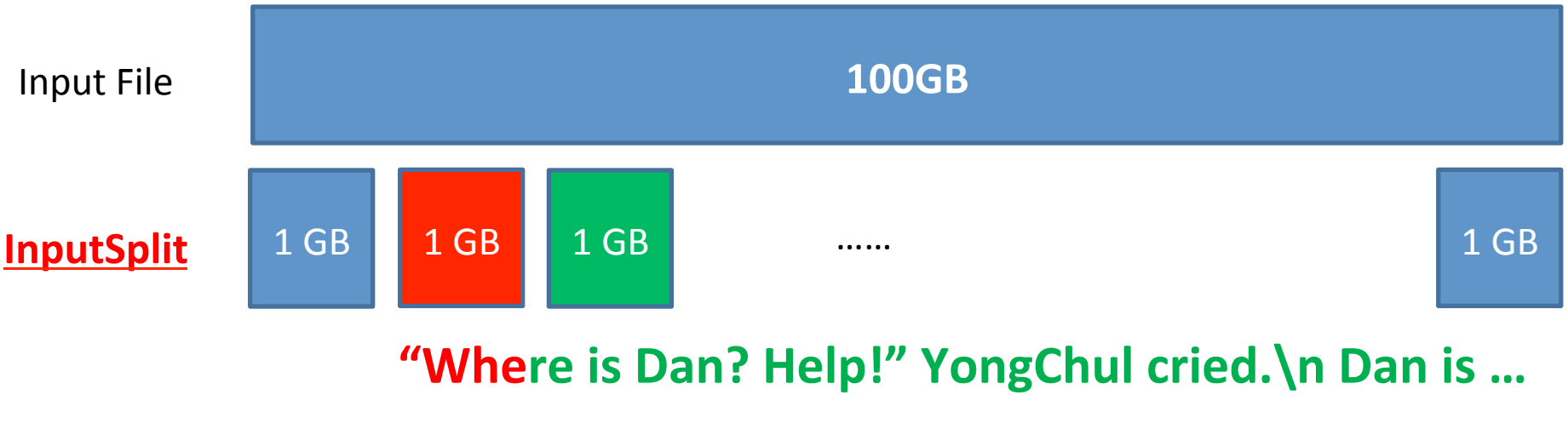
- How map() is distributed?
  - How many records per task?
- How a record is parsed?

How map is distributed?

- InputFormat
  - How to split the input data
  - How to read the input data
- InputSplit
  - One InputSplit = One Map Task
  - Split tells how the input data is partitioned
- RecordReader
  - Parse the data, offer one “complete” record per call

Example: File

- 100 GB text file to 100 Map tasks
  - How best to split the input data?
  - Which information is kept in InputSplit?
  - How to handle boundary case?



Reduce Task

- Input is already prepared by Hadoop
  - No InputFormat, RecordReader, …
- How data is distributed?
  - Typically, hash partition on key
  - You can specify your own logic for this
- Is it easy to assign the same number of inputs?
  - No. 1) the output key is generated by map(), 2) partition logic may not guarantee even distribution

Straggler Problem

- Your neighbor competes for resources
  - CPU/Memory/Network/Disk/…
- It is possible that one of your tasks becomes unfortunate
- Reaction: reschedule the straggler on different machines

Job Configuration

- There are many configurations
  - ~ 110 configurations as of Hadoop 0.21 (only MapReduce)

- # of Maps?
  - Determined by # of input splits
  - For files in HDFS, typically one split per block
- # of Reduces?
  - Currently, manually specified ☹
  - Rule of thumb: (0.95 or 1.75) * (# of nodes * # of reduce slots per task tracker)

Why Pig/SQL is good?

- Why? Any thoughts?

Further Reference

- Hadoop Documentation
  - http://hadoop.apache.org/mapreduce/docs/r0.21.0/mapred_tutorial.html
  - http://hadoop.apache.org/mapreduce/docs/r0.21.0/api/index.html

- Hadoop: The Definitive Guide 2nd Ed.