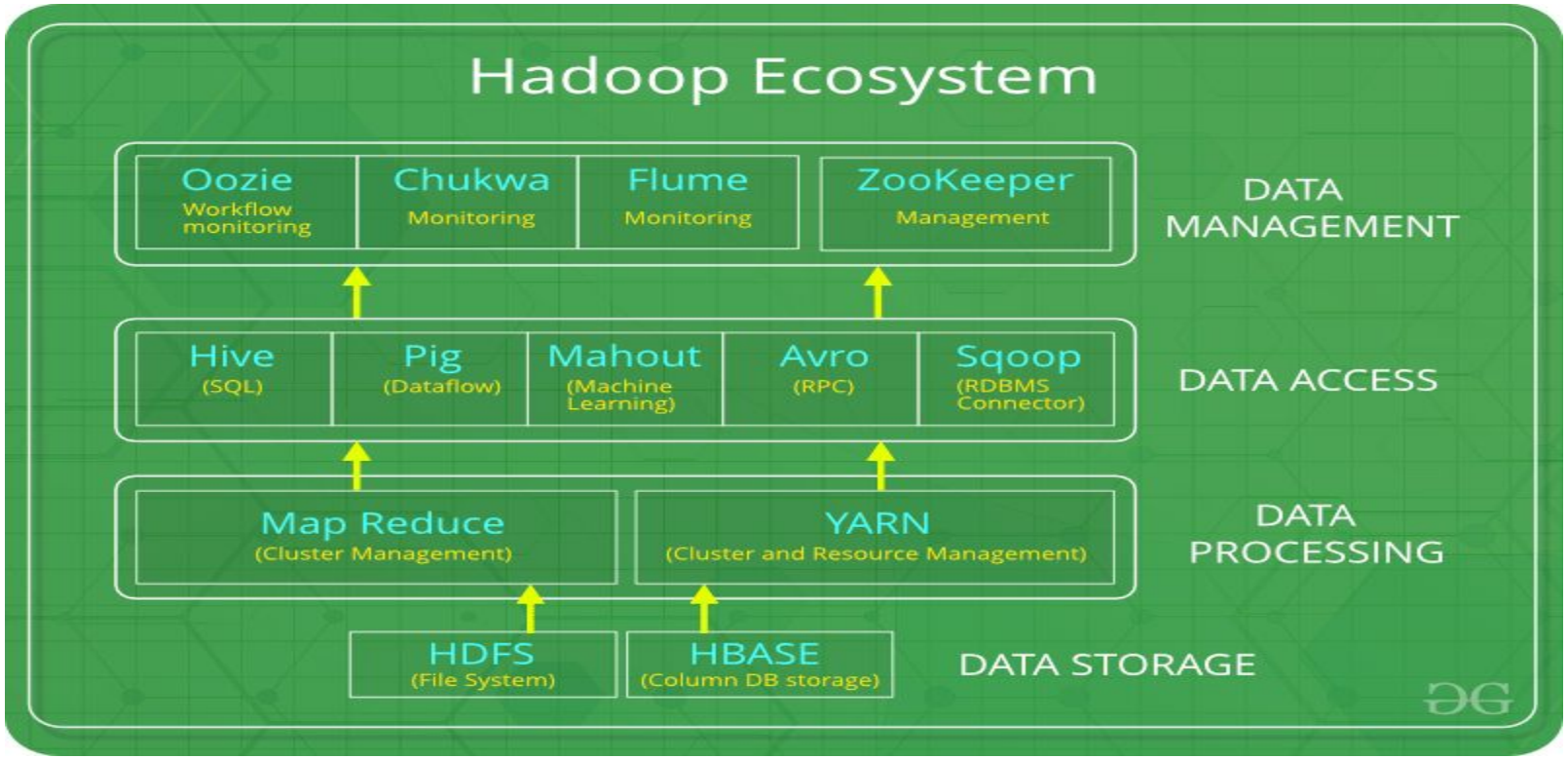
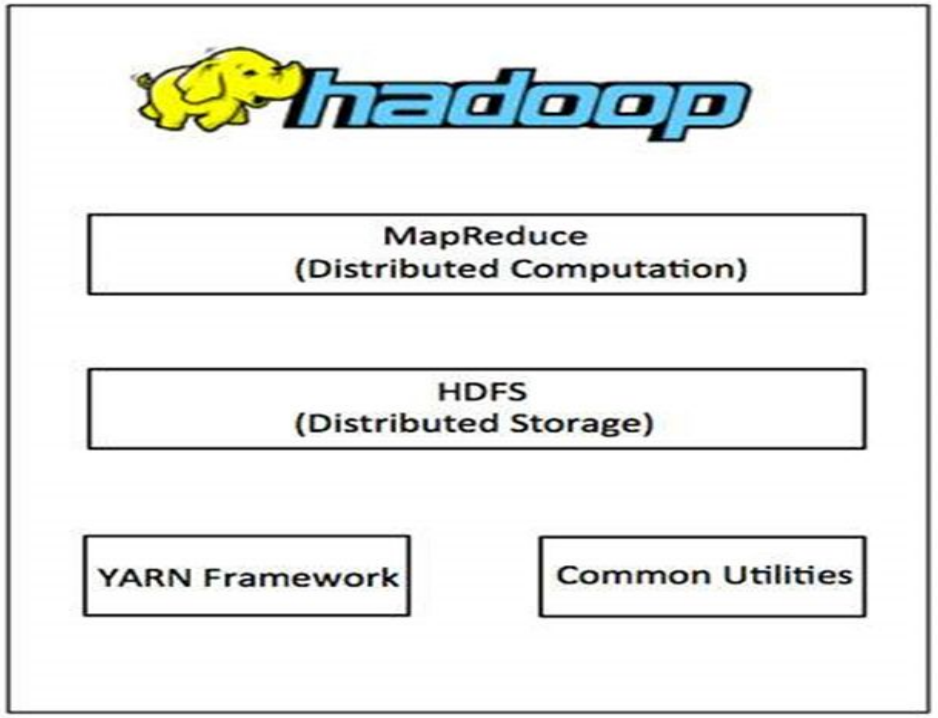
YARN


Hadoop Ecosystem
Oozie
Workflow monitoring
Chukwa
Monitoring
Flume
Monitoring
ZooKeeper
Management
DATA MANAGEMENT
Hive
(SQL)
Pig
(Dataflow)
Mahout
(Machine Learning)
Avro
(RPC)
Sqoop
(RDBMS Connector)
DATA ACCESS
Map Reduce
(Cluster Management)
YARN
(Cluster and Resource Management)
DATA PROCESSING
HDFS
(File System)
HBASE
(Column DB storage)
DATA STORAGE

YARN Architecture

- YARN (**Y**et **A**nother **R**esource **N**egotiator) is a new component added in Hadoop 2.0
- In Hadoop 2.0 a new layer has been introduced between HDFS and MapReduce.
- This is YARN framework which is responsible for doing Cluster Resource Management.

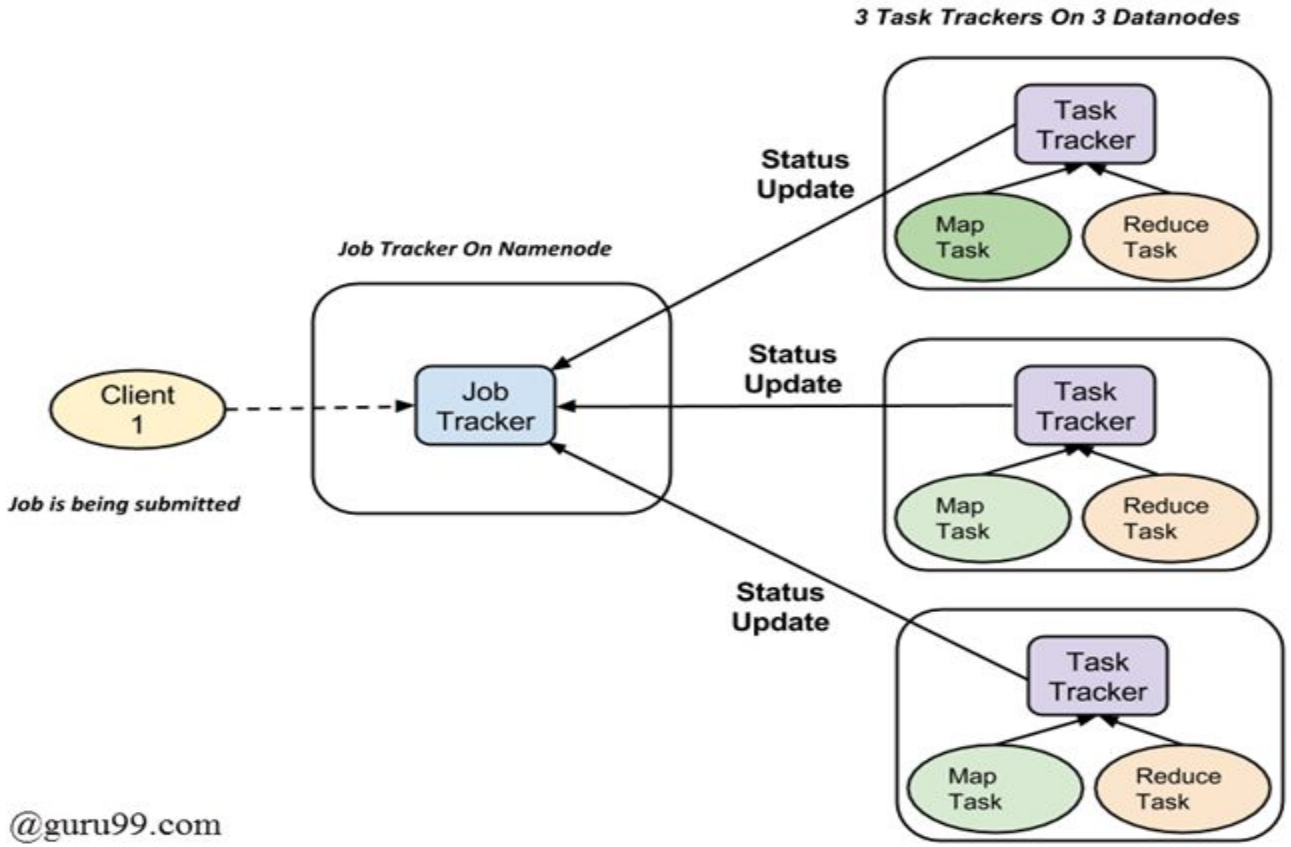


Cluster Resource Management

Cluster resource management means managing the resources of the Hadoop Clusters. And by resources we mean Memory, CPU etc.

- YARN took over this task of cluster management from Map Reduce and Map Reduce is streamlined to perform Data Processing only in which it is best.
- We have learned Map reduce is a programming model used for processing large amounts data.
- It is part of Apache Hadoop project that has been widely used for handling Big data.

- The Map reduce model is based on Map reduce programming paradigm that divides large datasets into multiple small datasets distributed over a number of mapper and reducer nodes used for processing data.



Limitations of Map Reduce

- Scalability:The Hadoop1 architecture provides restricted level of scaling because of the single machine that runs the job tracker to handle all task tracker nodes.
- Job tracker has already too much responsibility and addition of more nodes to cluster increases the work of job tracker.
- Availability: The scope of availability is limited. Job tracker is the only point of availability in case of failure all the jobs need to restart
- Resource Utilization: Hadoop 1 defines the number of nodes separately to handle the mapper and reducer tasks in advance. When all the nodes defined for handling mapper tasks are full you cant assign a mapper task to reducer
- Non-Map reduce applications: Hadoop 1 is designed to do only map reduce programming jobs which means you cannot handle non map reduce applications with it

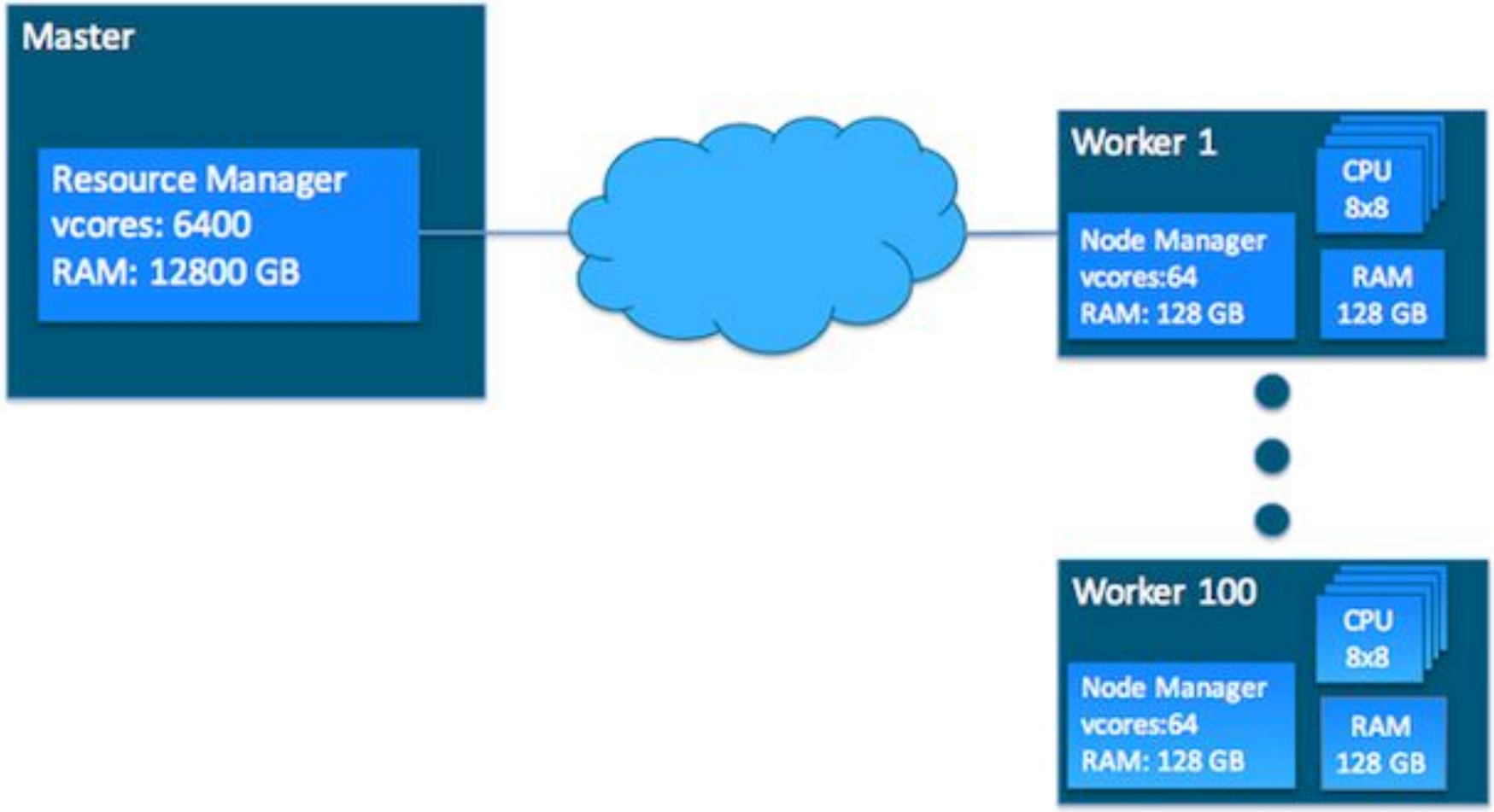
Advantages of YARN

- Efficient Resource management
- Multiple applications can run simultaneously by sharing common resources
- YARN allows Hadoop ecosystem to even run those applications in which the Map reduce model is not followed
- YARN provides backward compatibility that means applications developed on Hadoop 1 can run on Hadoop 2 directly
- YARN doesnt have jobtracker and tasktracker nodes. Instead two main functionalities of jobtracker are assigned to different components

Two primary components

- Resource manager
- Application Manager

Yarn Resource Monitoring (ii)



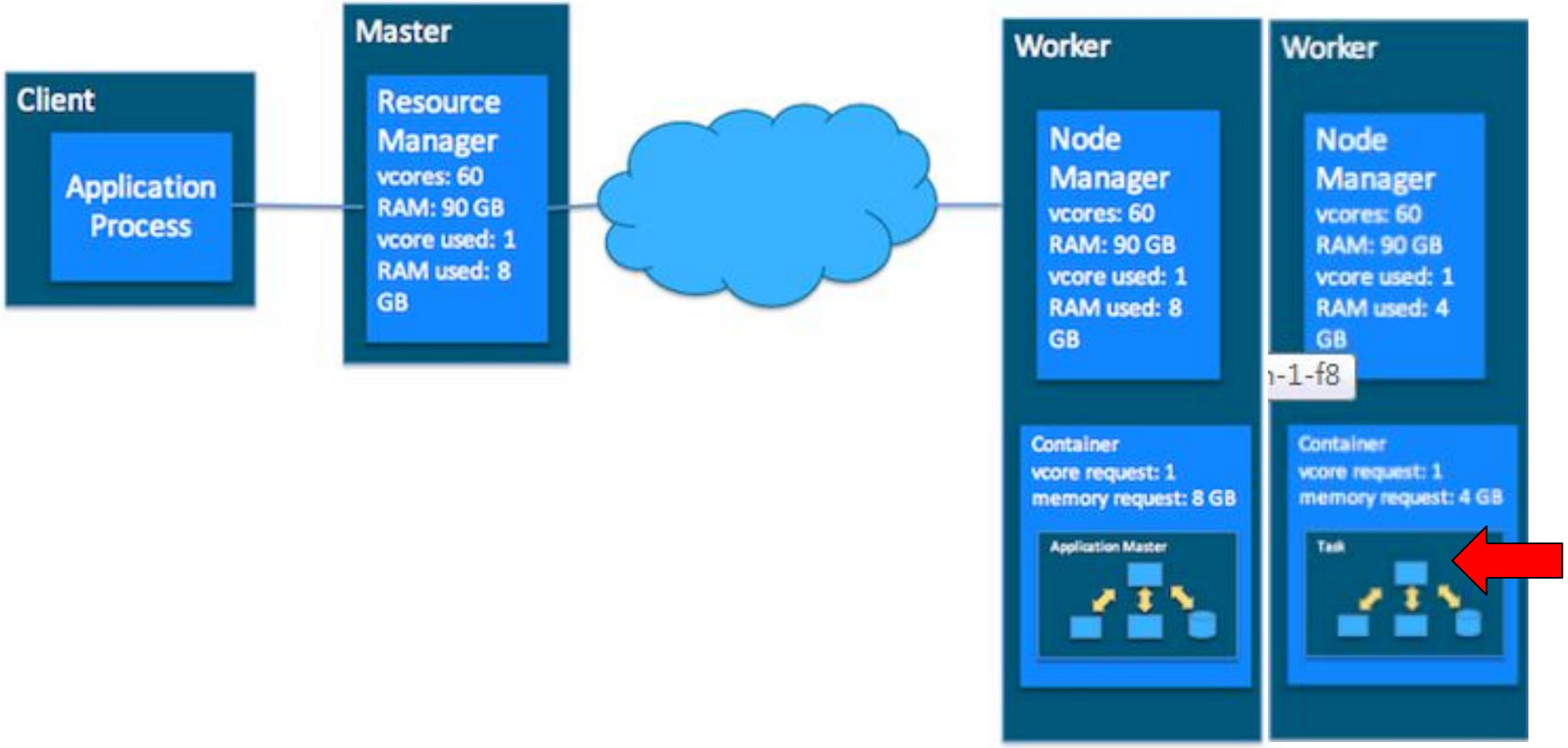
100 workers of same resources

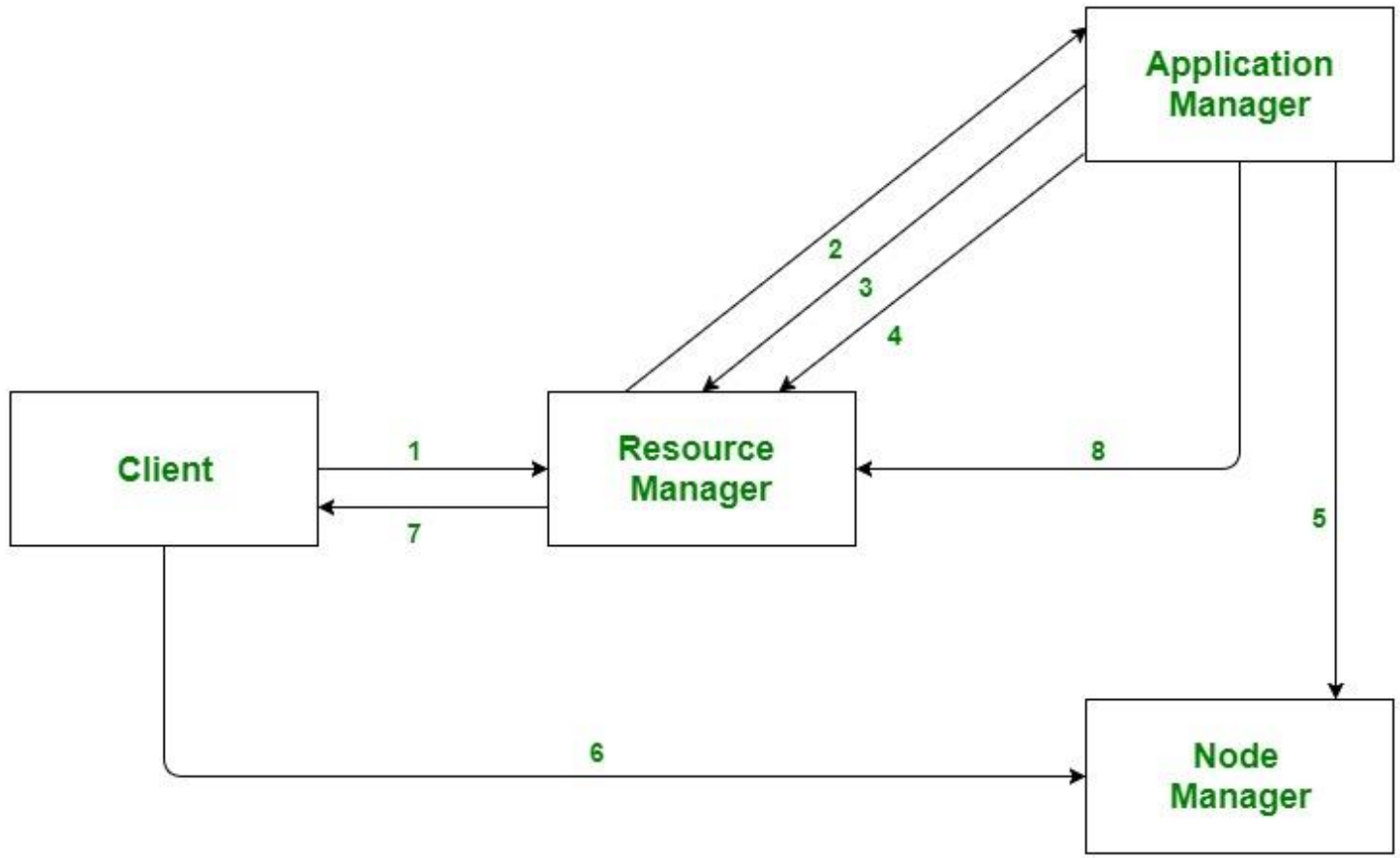
Yarn Application and ApplicationMaster

- Yarn application
  - It is a YARN client program that is made up of one or more tasks
  - Example: MapReduce Application
- ApplicationMaster
  - It helps coordinate tasks on the YARN cluster for each running application
  - It is the first process run after the application starts.

Interactions among Yarn Components (iv)

4. The ApplicationMaster requests subsequent containers from the ResourceManager that are allocated to run tasks for the application. Those tasks do most of the status communication with the ApplicationMaster allocated




Application Manager
2
3
4
Client
1
Resource Manager
8
7
5
6
Node Manager

- Client submits an application
- The Resource Manager allocates a container to start the Application Manager
- The Application Manager registers itself with the Resource Manager
- The Application Manager negotiates containers from the Resource Manager
- The Application Manager notifies the Node Manager to launch containers
- Application code is executed in the container
- Client contacts Resource Manager/Application Manager to monitor application’s status
- Once the processing is complete, the Application Manager un-registers with the Resource Manager