

Hands-on Apache Spark

Big Data Class

Spring 2017

DIAG, Sapienza University

Using Apache Spark

1) Cloud-based cluster

- Amazon, Azure, Google
- Databricks
- ...

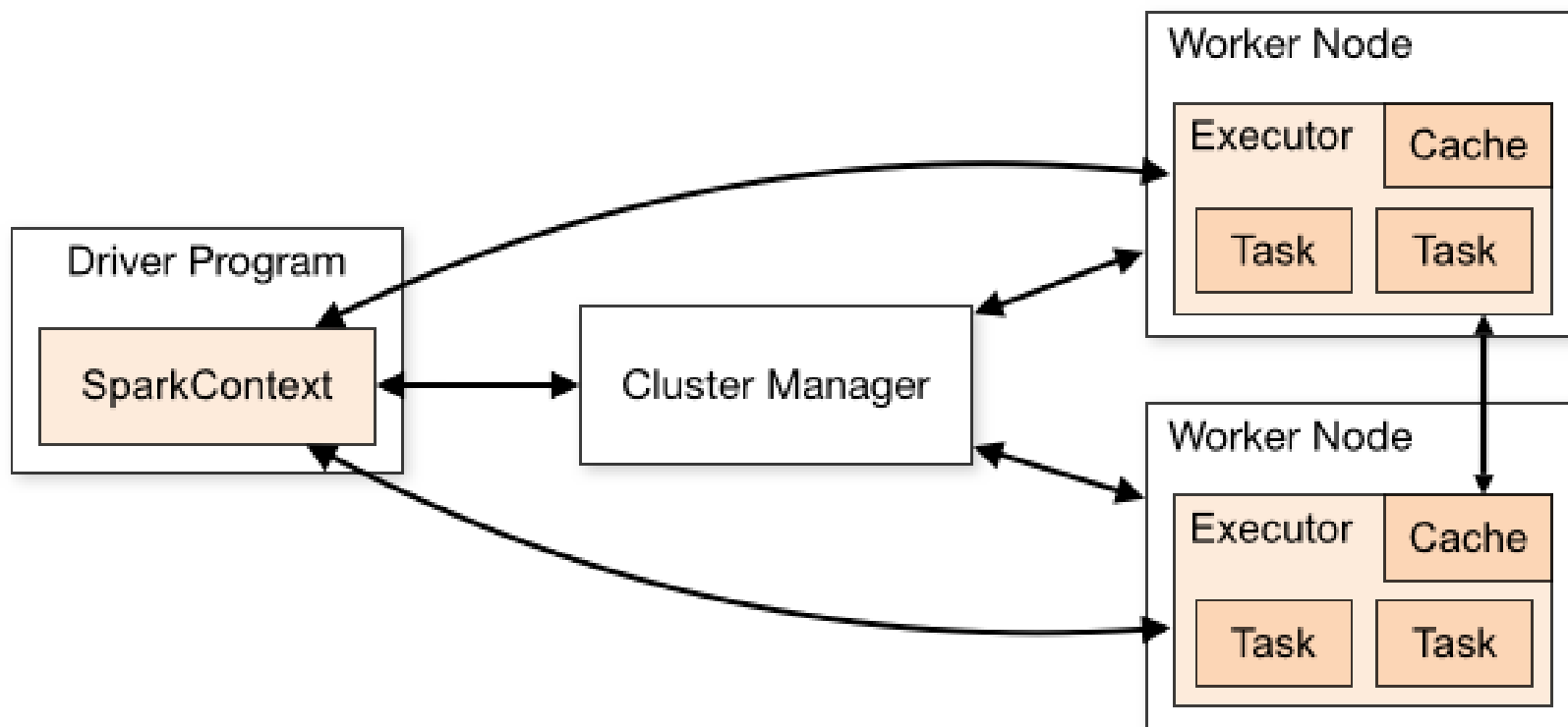
2) Standalone cluster

- Deploy spark on local machines

3) Single machine mode

- Good for testing
- Good for multi-core systems

Cluster Mode Components



Single Machine / Interactive Mode

- Download/Install Spark 2.1.0
- Invoke Python Shell
 - `./bin/pyspark`
- Enable an interactive shell (jupyter/ipython)
 - `PYSPARK_DRIVER_PYTHON="jupyter"`
 - `PYSPARK_DRIVER_PYTHON_OPTS="notebook"`

Standalone Cluster

- One machine assumes the role of the master
 - `./sbin/start-master.sh`
 - A simple web-ui console available at port 8080
- One or more machines become slaves
 - `./sbin/start-slave.sh spark://<ip>:7077`
- Declare the Master URL
 - `MASTER=spark://<ip>:7070 ./bin/pyspark`