Introduction to

Presented by Dara Adib
Realtime System Operations & Realtime Data
WTF is Mesos

- Resource manager and scheduler for shared, distributed applications.
- Fault-tolerant and scalable.
- Supports Linux containers and Docker.
Frameworks

Frameworks are distributed applications which run on the cluster. Frameworks share resources in a cluster.

- **Scheduler**
  - Connects to Mesos master.
  - Accepts or declines resources.
  - Contains delay scheduling logic for rack locality, etc.

- **Executor**
  - Connects to local Mesos slave.
  - Runs framework tasks.

- **Mesos handles scheduler-executor communication.**

- **Metaframeworks**
  - Aurora, Marathon, Chronos

- **Data processing**
  - Spark, Storm, Myriad (Hadoop)

- **Many others**
  - HDFS, Cassandra, Kafka, MySQL
Masters and Slaves

- **Master**
  - Shares resources between frameworks.
  - Keeps state (frameworks, slaves, tasks, etc.) in memory.
  - **HA:**
    - 1 master elected leader.
    - Majority of masters needed for ZooKeeper quorum.
    - Paxos-based replicated log.

- **Slave**
  - Runs on each cluster node.
  - Specifies resources and attributes.
  - Starts executors.
  - Communicates with master and executors to run tasks.
1. Slave reports available resources to the master.
2. Master sends a resource offer to the framework scheduler.
3. Framework scheduler requests two tasks on the slave.
4. Master sends the tasks to the slave which allocates resources to the framework’s executor, which in turn launches the two tasks.
Resources

Types

- cpu: CPU share
  - optional CFS for fixed
- mem: memory limit
- disk space: disk limit
- ports: integer port range
- bandwidth

Custom resources: k,v pairs

Isolation

- Linux container
  - control groups (cgroup)
  - namespaces
- Docker container
- External container

Other features

- Reserved resources by role
- Oversubscription
- Persistent volumes
Failure Modes

In High-Availability mode with ZooKeeper

● Elected master fails.
  ○ New master elected.
  ○ Slaves and frameworks connect to new Mesos master via ZooKeeper.

● Majority of masters fail (quorum).
  ○ Executors continue to run tasks (framework dependent).

● Slaves fail.
  ○ Executors continue to run tasks (framework dependent).
  ○ If checkpointing (slave recovery) is enabled, slave will connect to running executors on start.
  ○ If health check on master fails (~75 seconds), master will treat tasks as failed and kill tasks on slave start.

● Fail-fast approach with rate limits. Use a process supervisor.
In Theory

From Static Partitioning to Elastic Sharing

Static Partitioning
- WASTED
  - WEB
  - CACHE
  - HADOOP

Elastic Sharing
- HADOOP
  - WEB
  - CACHE
  - FREE
  - FREE
MESOS
ALL THE THINGS
In Practice

<redacted slide>
Additional Facts

- Written in C++.
- Originally developed at UC Berkeley AMP Lab in 10,000 lines of code.
  - Apache Spark was written as an “example framework”.
- Prominently used by Twitter with Apache Aurora.
- Commercially supported by Mesosphere with Marathon.
  - Provides Debian packages.
Additional Resources

- **Official website**
  - mesos.apache.org

- **Mesosphere**
  - mesosphere.com

- **Papers**

- **Books**