MESOS

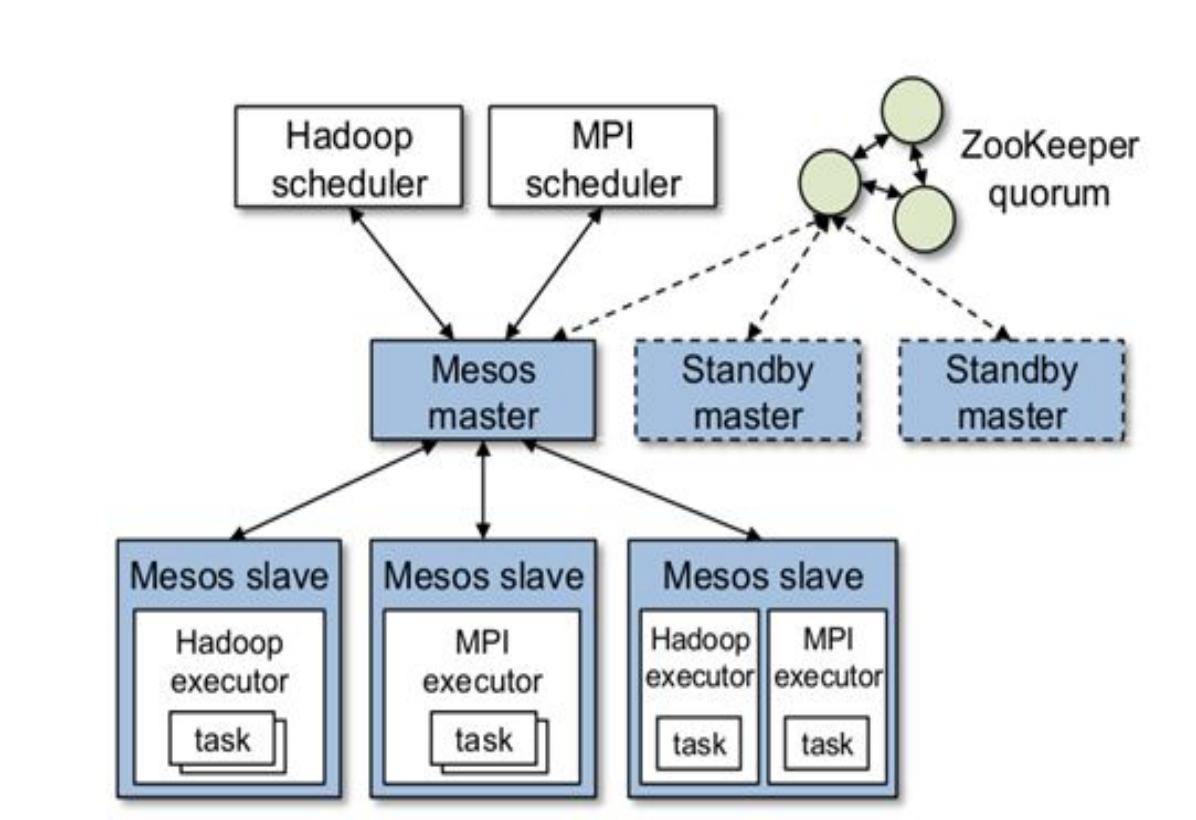
What is Mesos

- Resource scheduler
- Cluster manager
  - Manage shared pool of servers
  - Simplifies the complexity of running applications
- Data center kernel
- Provides node abstraction
- Runs on Linux, Solaris and OS X, and supports frameworks written in C++, Java, and Python.

Why Mesos

- Abstracts CPU, Memory, storage and other compute resources away from machines
  - Enables fault-tolerant and elastic distibutes systems to be
    - Built and run effectively
- Fine-Grained resource sharing in a data center
- Easily scale up to 10,000s of nodes
- Fault tolerant
- Provides HTTP APIs for developing new distributed applications, for operating the cluster, and for monitoring
- Mesos achieves higher utilization than static partitioning, and that jobs finish at least as fast in the shared cluster as they do in their static partition, and possibly faster due to gaps in the demand of other frameworks.

Architecture:Overview



- The architecture make Mesos simple and minimizes the rate of change required of the system, which makes it easier to keep Mesos scalable and robust.
- Figure shows two running frameworks on Mesos(Hadoop and MPI) [1]

Architecture:Overview

- Mesos consists of a
  - Master deamon
    - Manages agents
  - Agent deamons
    - Runs on each cluster node
  - Mesos frameworks
    - Runs tasks on agents
- Master enables fine-grained sharing of resources (CPU, RAM, …) across frameworks by making them *resource offers*.
- Resource offer contains a list of
  - <agent ID, resource1: amount1, resource2: amount2, ...>

Architecture:Framework

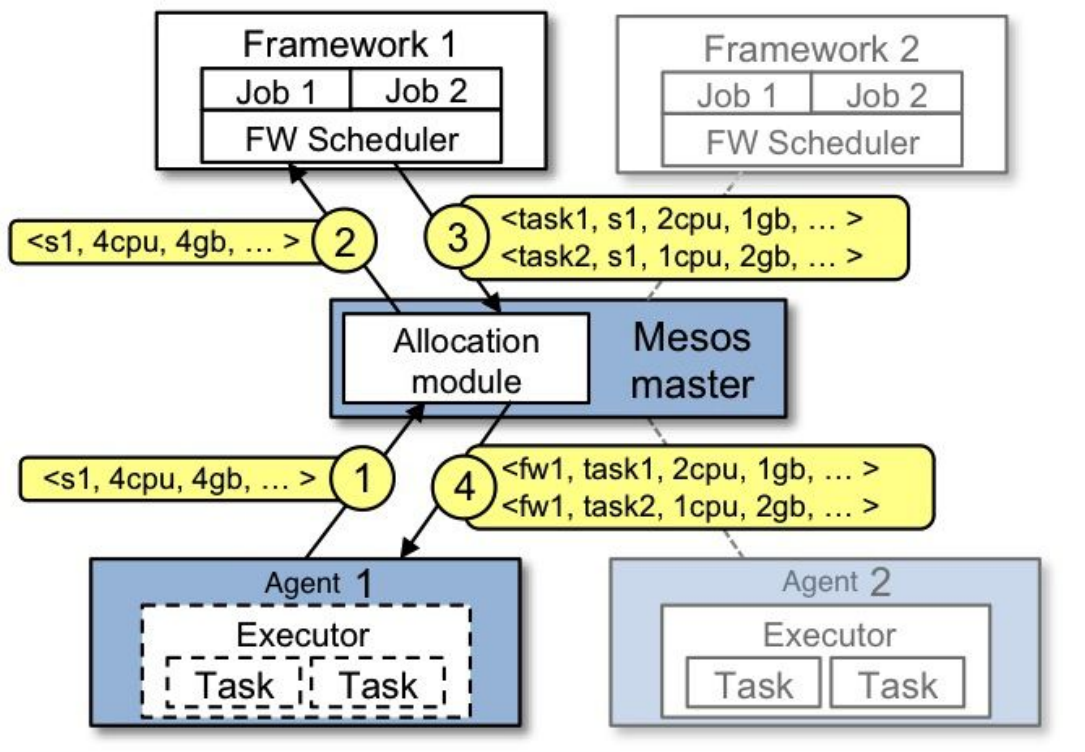
Mesos framework consists of scheduler and executor

- Scheduler
  - Registers with the master
  - Offered resources by master
    - Selects among offered resources
- Executor
  - Launched on agent nodes
  - Runs the framework’s tasks

Architecture: Resource Allocation

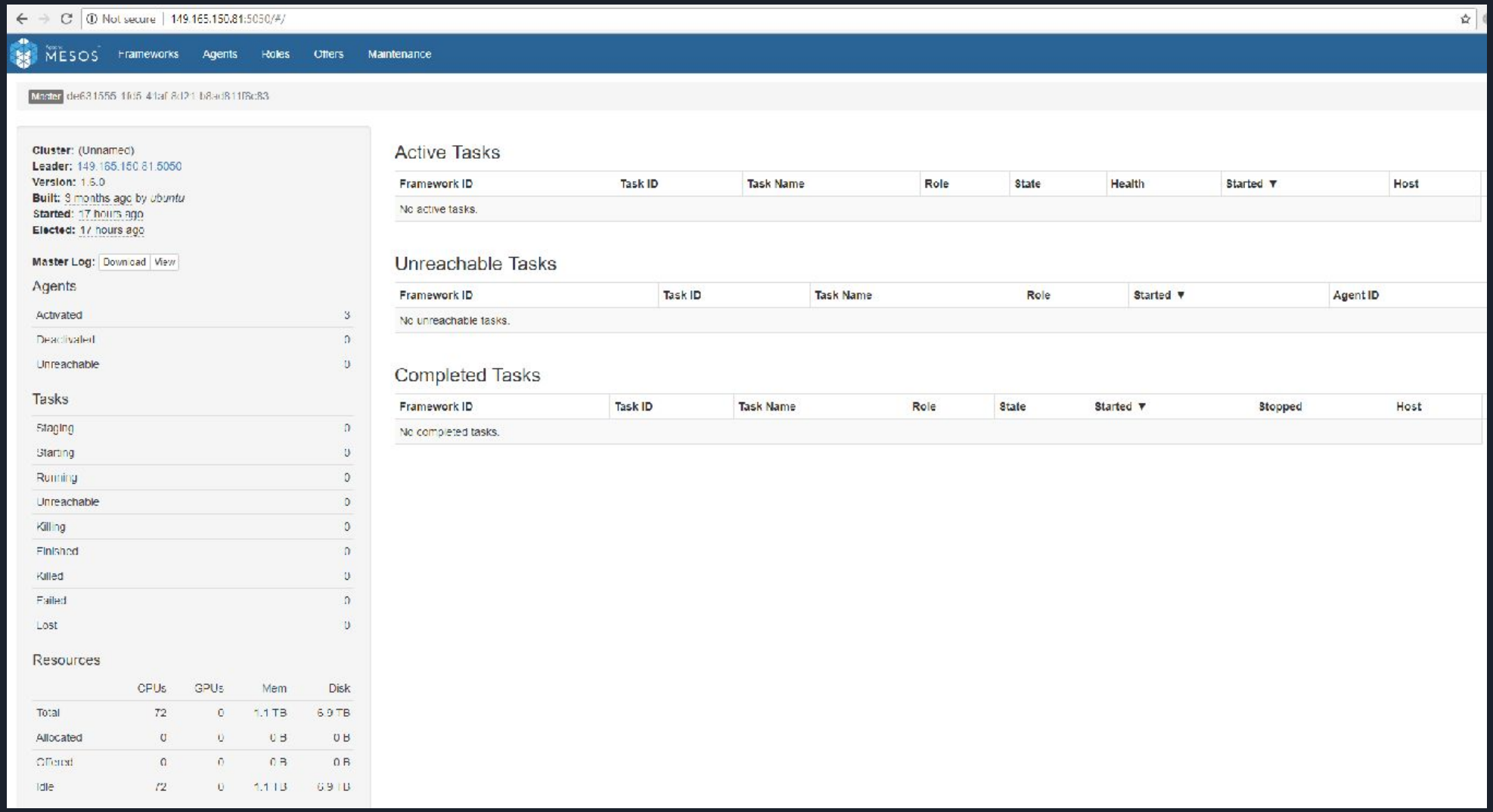
- The master decides how many resources to offer to each framework according to an organizational policy, such as fair sharing or priority.
- Mesos lets organizations define their own policies via a pluggable allocation module.
- Mesos takes advantage of the fact that most tasks are short, and only reallocates resources when tasks finish.
- Due to a buggy job or a greedy framework, the allocation module can also revoke (kill) tasks.

Example Of Resource Offer



- Resource Allocation Procedure Example:
- (1) Agent 1 reports to the master that it has 4 CPUs and 4 GB of memory free. The master then invokes the allocation policy module, which tells it that framework 1 should be offered all available resources.
- (2) The master sends a resource offer describing what is available on agent 1 to Framework 1.
- (3) The framework's scheduler replies to the master with information about two tasks to run on the agent, using <2 CPUs, 1 GB RAM> for the task1, and <1 CPU, 2 GB RAM> for the task2.
- (4) Finally, the master sends the tasks to the agent, which allocates appropriate resources to the framework's executor, which in turn launches the two tasks (depicted with dotted-line borders in the figure). Because 1 CPU and 1 GB of RAM are still unallocated, the allocation module may now offer them to framework 2.

Mesos:Web UI


Not secure | 149.165.150.81:5050/#/
MESOS
Frameworks
Agents
Roles
Offers
Maintenance
Master de631555-1b5-41af-8d21-b8ed811f6c83
Cluster: (Unnamed)
Leader: 149.165.150.81:5050
Version: 1.6.0
Built: 9 months ago by ubuntu
Started: 17 hours ago
Elected: 17 hours ago
Master Log: Download View
Agents
Activated 3
Deactivated 0
Unreachable 0
Tasks
Staging 0
Starting 0
Running 0
Unreachable 0
Killing 0
Finished 0
Killed 0
Failed 0
Lost 0
Resources
CPUs GPUs Mem Disk
Total 72 0 1.1 TB 6.9 TB
Allocated 0 0 0 B 0 B
Offered 0 0 0 B 0 B
Idle 72 0 1.1 TB 6.9 TB
Active Tasks
Framework ID Task ID Task Name Role State Health Started ▼ Host
No active tasks.
Unreachable Tasks
Framework ID Task ID Task Name Role Started ▼ Agent ID
No unreachable tasks.
Completed Tasks
Framework ID Task ID Task Name Role State Started ▼ Stopped Host
No completed tasks.

Fault Tolerance

- All the frameworks depend on the Mesos master, it is critical to make the master fault-tolerant.
- Master’s only state is the list of active slaves, active frameworks, and running tasks.
- Run multiple masters in a hot-standby configuration using ZooKeeper for leader election. When the active master fails, the slaves and schedulers connect to the next elected master and re-populate its state.
- Scheduler failures, Mesos allows a framework to register multiple schedulers such that when one fails, another one is notified by the Mesos master to take over.

Containerization

- Containerizers are used to run task in containers
  - Isolate tasks from other running tasks
  - Contain tasks to run in limited resource runtime env.
  - Control a task’s resource usafe programmatically
  - Run a software in a pre-packaged file system image, allowing it to run in different environments.
- Mesos implements the following containerizers
  - Composing
  - Docker
  - Mesos(default)

Twister2 on Mesos

- Integration of Mesos and Twister2 provided a mechanism that Twister2 jobs can be run on Mesos.
- There are two choices when implementing Twister2 workers in Mesos
  - one worker per executor: Exactly one worker runs on each executor.
  - multiple workers per executor:Multiple workers runs on each executor.
- Initialization of executors are slow in Mesos, therefore second option is a faster solution.
- We use Docker container to run Twister2 jobs.
  - it has all Twister2 and dependency library files
  - this container is downloaded from docker hub
    - name is given in the configuration file

Twister2 on Mesos

- Twister2 job package
  - User job java files. The codes user wants to run on Twister2.
  - Configuration files for the job and the environment.
  - Serialization of the Job object. Job specific dynamic values.

- Twister2 core package and job package are transferred to the node where the Mesos master is running
  - transferred to master using uploaders provided in Twister2
  - each worker needs the core and the job package
  - workers fetch these files to their sandboxes
  - uses HTTP protocol to do that

Twister2 on Mesos

- Worker Discovery
  - Worker controller watches the agents and provides the following services
    - unique id assignment to workers
    - worker address discovery
    - getting IP addresses after workers start
- Persistent storage
  - provides persistent storage for worker logs
  - Network File System(NFS) is used for this purpose
- NodePort Service
  - provides extra ports for each worker on demand
- Node Mapping
  - you can map workers to specific machines in the cluster.
  - it is done through a configuration parameter

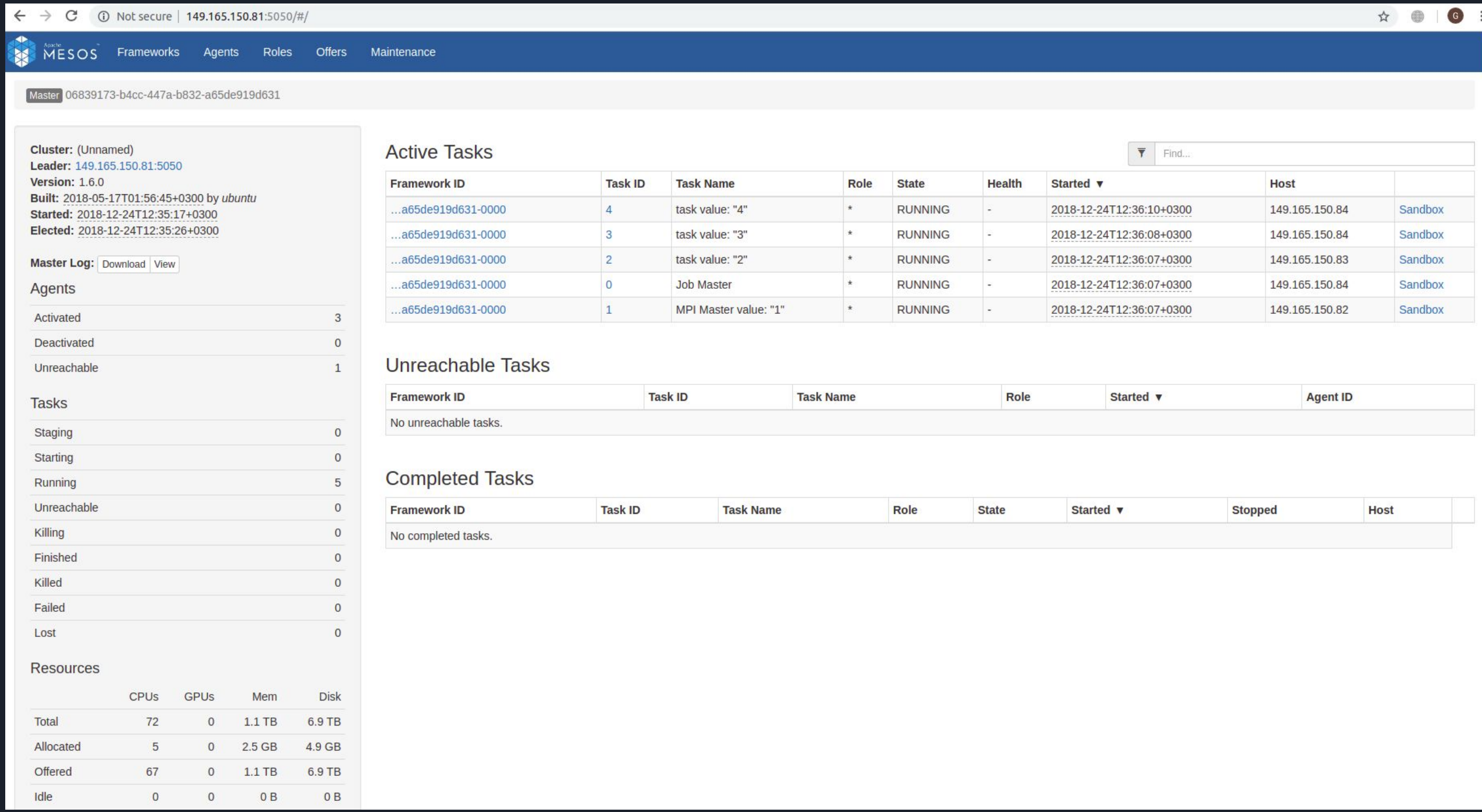
Twister2 on Mesos

- Twister2 can run OpenMPI on Mesos
  - Only OpenMPI implementation on Mesos
  - Uses Docker containerization for this purpose
  - Needs Docker swarm installation for overlay network among containers
  - Password-free SSH access should be enabled among Docker containers
  - OpenMPI and OpenSSH libraries should be installed on Docker containers
  - Hostfile generation is done MPI master before running the MPI command
    - Uses Zookeeper to get worker IP addresses and port numbers

- Next slides shows the screenshots of twister2 jobs running on Mesos
  - First one shows the terminal output where you submit a job.
  - Second one is the dashboard of Mesos that shows the running jobs.

```
I1224 04:26:12.882903  4491 group.cpp:831] Syncing group operations: queue size (joins, cancels, datas) = (0, 0, 0)
I1224 04:26:12.882925  4491 group.cpp:419] Trying to create path '/mesos' in ZooKeeper
I1224 04:26:12.885380  4491 detector.cpp:152] Detected a new leader: (id='1535')
I1224 04:26:12.885639  4502 group.cpp:700] Trying to get '/mesos/json.info_0000001535' in ZooKeeper
I1224 04:26:12.893975  4502 zookeeper.cpp:262] A new leading master (UPID=master@149.165.150.81:5050) is detected
I1224 04:26:12.894122  4502 sched.cpp:336] New master detected at master@149.165.150.81:5050
I1224 04:26:12.894733  4502 sched.cpp:351] No credentials provided. Attempting to register without authentication
I1224 04:26:12.896603  4494 sched.cpp:749] Framework registered with ff33c193-f7b4-489f-9ff9-266b0d59be4f-0000
Dec 24, 2018 4:26:12 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler registered
INFO: Registeredvalue: "ff33c193-f7b4-489f-9ff9-266b0d59be4f-0000"

Dec 24, 2018 4:26:12 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer comes from host ...:149.165.150.84
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer from host 149.165.150.84has been accepted.
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer comes from host ...:149.165.150.83
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer from host 149.165.150.83has been accepted.
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer comes from host ...:149.165.150.82
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer from host 149.165.150.82has been accepted.
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_STARTING from 2
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_STARTING from 0
Dec 24, 2018 4:26:13 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_STARTING from 1
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer comes from host ...:149.165.150.84
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer from host 149.165.150.84has been accepted.
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_RUNNING from 0
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_STARTING from 3
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_RUNNING from 2
Dec 24, 2018 4:26:15 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_RUNNING from 1
Dec 24, 2018 4:26:16 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_RUNNING from 3
Dec 24, 2018 4:26:17 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer comes from host ...:149.165.150.84
Dec 24, 2018 4:26:17 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler resourceOffers
INFO: Offer from host 149.165.150.84has been accepted.
Dec 24, 2018 4:26:17 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
INFO: Status update: TASK_STARTING from 4
Dec 24, 2018 4:26:18 AM edu.iu.dsc.tws.rsched.schedulers.mesos.MesosScheduler statusUpdate
```


Not secure | 149.165.150.81:5050/#/
Apache MESOS
Frameworks
Agents
Roles
Offers
Maintenance
Master 06839173-b4cc-447a-b832-a65de919d631
Cluster: (Unnamed)
Leader: 149.165.150.81:5050
Version: 1.6.0
Built: 2018-05-17T01:56:45+0300 by ubuntu
Started: 2018-12-24T12:35:17+0300
Elected: 2018-12-24T12:35:26+0300
Master Log: Download View
Agents
Activated 3
Deactivated 0
Unreachable 1
Tasks
Staging 0
Starting 0
Running 5
Unreachable 0
Killing 0
Finished 0
Killed 0
Failed 0
Lost 0
Resources
CPUs GPUs Mem Disk
Total 72 0 1.1 TB 6.9 TB
Allocated 5 0 2.5 GB 4.9 GB
Offered 67 0 1.1 TB 6.9 TB
Idle 0 0 0 B 0 B
Active Tasks
Find...
Framework ID Task ID Task Name Role State Health Started ▼ Host
…a65de919d631-0000 4 task value: "4" * RUNNING - 2018-12-24T12:36:10+0300 149.165.150.84 Sandbox
…a65de919d631-0000 3 task value: "3" * RUNNING - 2018-12-24T12:36:08+0300 149.165.150.84 Sandbox
…a65de919d631-0000 2 task value: "2" * RUNNING - 2018-12-24T12:36:07+0300 149.165.150.83 Sandbox
…a65de919d631-0000 0 Job Master * RUNNING - 2018-12-24T12:36:07+0300 149.165.150.84 Sandbox
…a65de919d631-0000 1 MPI Master value: "1" * RUNNING - 2018-12-24T12:36:07+0300 149.165.150.82 Sandbox
Unreachable Tasks
Framework ID Task ID Task Name Role Started ▼ Agent ID
No unreachable tasks.
Completed Tasks
Framework ID Task ID Task Name Role State Started ▼ Stopped Host
No completed tasks.