Apache Mesos

incubator.apache.org/mesos

@ApacheMesos

Benjamin Hindman – @benh
origins

Berkeley research project including Benjamin Hindman, Andy Konwinski, Matei Zaharia, Ali Ghodsi, Anthony D. Joseph, Randy Katz, Scott Shenker, Ion Stoica

incubator.apache.org/mesos/research.html
motivation: static partitioning
static partitioning considered harmful
static partitioning considered harmful

hard to fully *utilize* machines (e.g., 72 GB RAM and 24 CPUs)
static partitioning considered harmful

harder to deal with *failures*
static partitioning considered harmful

harder to scale *elastically*
Mesos
level of indirection
Mesos:

1) efficiently *share* datacenter resources
better utilization
better utilization

Node
better utilization

- Hadoop
- Spark
- Node
- MPI
- service
easier to deal with failures
enables elasticity
enables elasticity
Mesos:

1) efficiently *share* datacenter resources

2) make it *easier* to build distributed services and analytics frameworks
a “kernel” for the datacenter
"kernel" primitives

messaging

mechanisms for high-availability

fault-detection

resource isolation (cgroups)

resource monitoring
anatomy of a framework

1. scheduler
scheduler

- requests resources, assigns tasks

- Mesos master allocation module
Hadoop

Mesos master allocation module

Spark

decides how to allocate resources
“two-level scheduling”

Mesos: controls resource allocations to schedulers

Schedulers: make decisions about what to run
requests and offers
tasks

abstraction representing some consumption of resources

first-class in Mesos, but no requirement on manifestation (i.e., can be a thread, a process, a work queue, etc)

*first-class* tasks enables *fine-grained* resource sharing!
tasks

Hadoop

Mesos master

allocation module

Mesos slave

Mesos slave

task

launches, isolates, and monitors tasks
anatomy of a framework

1. scheduler

2. executor\(^1\) (optional, if you want more control over how a task is executed)

\(^1\)also a consumer of resources
executors

Hadoop

Mesos master

allocation module

Mesos slave

Spark executor

task

launches, isolates, and monitors tasks and executors
executors

- Hadoop
- Spark

Mesos master

Mesos slave
- Hadoop executor
  - reduce
- Spark executor
  - task

allocation module

launches, isolates, and monitors tasks and executors
resource isolation

support for Linux control groups (cgroups)

isolates CPU, memory, disk I/O, network I/O
resource isolation is fine-grained

cgroup per executor or task (if no executor)

resource limits adjusted \textit{dynamically}, as tasks are launched and terminate over lifetime of executor!
Hadoop

scheduler == JobTracker
executor == TaskTracker
Hadoop

requires minor patch + contrib scheduler
Hadoop

$ wget ...hadoop-0.20.205.0.jar
$ tar zxvf hadoop-0.20.205.0.tar.gz
$ cd hadoop-0.20.205.0
Hadoop

$ wget ...hadoop-0.20.2-cdh3u3.jar
$ tar zxvf hadoop-0.20.2-cdh3u3.tar.gz
$ cd hadoop-0.20.2-cdh3u3
Hadoop

$ patch -p1 <.../hadoop-0.20.205.0.patch
Hadoop

$ patch -p1 <.../hadoop-0.20.2-cdh3u3.patch
Hadoop

```
$ diffstat hadoop/hadoop-0.20.205.0.patch
  JobInProgress.java | 4 ++
  Task.java | 3 +
  TaskRunner.java | 17 +++++++--
  TaskTracker.java | 68 +++++++++++++++++++++++++++++++++++++------
  TaskTrackerInstrumentation.java | 7 +++
5 files changed, 73 insertions(+), 26 deletions(-)
```

```
$ diffstat hadoop/hadoop-0.20.2-cdh3u3.patch
  JobInProgress.java | 4 ++
  Task.java | 3 +
  TaskRunner.java | 22 ++++++++--
  TaskTracker.java | 76 +++++++++++++++++++++++++++++++++++++------
  TaskTrackerInstrumentation.java | 7 +++
5 files changed, 97 insertions(+), 15 deletions(-)
```

committed in r1033804 and r987589
Hadoop

$ patch -p1 <.../hadoop-0.20.2-cdh3u3.patch
Hadoop

$ cp -r .../hadoop/mesos src/contrib
Hadoop

$ pwd
/path/to/hadoop-0.20.205.0/contrib/mesos
$ find .
./
./build.xml
./ivy
./ivy/libraries.properties
./ivy.xml
./src
./src/java
./src/java/org
./src/java/org/apache
./src/java/org/apache/hadoop
./src/java/org/apache/hadoop/mapred
./src/java/org/apache/hadoop/mapred/FrameworkExecutor.java
./src/java/org/apache/hadoop/mapred/FrameworkScheduler.java
./src/java/org/apache/hadoop/mapred/HadoopFrameworkMessage.java
./src/java/org/apache/hadoop/mapred/MesosScheduler.java
./src/java/org/apache/hadoop/mapred/MesosTaskTrackerInstrumentation.java
$ wc -l src/java/org/apache/hadoop/mapred/*
... 1256 total
tutorial

$ make hadoop-0.20.205.0
$ make hadoop-0.20.2-cdh3u3

$ ./TUTORIAL.sh
MPI

scheduler written in Python, wraps mpiexec

no executor!

included in distribution at mesos/mpi

assuming mpd is on every machine in your:

cluster:

$ ./mpiexec-mesos master_host:master_port
Mesos

Spark
Hadoop
MPI

...
Mesos at Twitter
Twitter framework

a framework that makes deploying and managing production servers easy

jobs/servers are submitted to the framework via a configuration file

provides:
  » rolling restarts/updates
  » re-launching processes after failures (if requested)
  » and more!
Mesos at Twitter

“Twitter’s kernel for the datacenter”

– Director of Engineering

> 1200 nodes, growing rapidly

5% - 10% higher utilization
demo
50,000+ lines of C++

libprocess for asynchronous actor style concurrency (github.com/libprocess)

APIs in C++, Java, Python

protobuf for data transport, data types
future

resource monitoring/collection

allocators (priority, weighted fair-sharing, etc)

revocation

scheduler management
try it out!

run on bare-metal or virtual machines – develop against Mesos API and run in private datacenter, or the cloud, or both!
operations

high-availability requires ZooKeeper (for now)

need to deploy the binaries yourself (e.g., puppet, monit)

can do rolling upgrades of masters/slaves (currently kills executors/tasks, in the future upgradable restarts will not kill running tasks/executors!)
questions?

incubator.apache.org/mesos

@ApacheMesos