Abstract

Cloud computing offers IT resources on-demand. The IaaS (Infrastructure-as-a-Service) layer provides virtual machines (VMs), storage and network. In order to overcome the difficulties carried by complex workflow managements, we have used the SlipStream solution. This service allows to automatically deploy environments on interoperable Cloud platforms such as StratusLab. We have deployed an Hadoop cluster in order to demonstrate the Cloud suitability for Big Data applications. Benchmarks were realised to check the cluster performance.

SlipStream


- Recipes for specific nodes (master and slave):
  - Disk image customisation.
  - Shared parameters.
- Deployment of complex workflows:
  - Node number.
  - Multi-Cloud.

StratusLab Cloud

StratusLab [7] offers an academic public IaaS Cloud since 2010. Virtual computing resources such as VMs are:

- generated by KVM hypervisor and managed by OpenNebula virtual infrastructure manager.
- supported by physical machines SLAM: 16 nodes, 440 CPUs, 573 GB of memory, 15 TB of storage, 1 GbE/s network.
- provided in OS disk images by the MarketPlace [4] and manageable by the StratusLab client.

SlipStream Web interface: Hadoop master node recipe.

SlipStream Web interface: Hadoop deployment.

SlipStream Web interface: Hadoop virtual cluster in running phase.

Hadoop cluster

Hadoop 2.0 is a recent version using Yet Another Resource Negotiator (YARN).

Hadoop cluster with YARN services.

Hadoop Job: scientific example

Big data application with MapReduce algorithm on meteorological data (NCDC) [8].

- Input data: 36 GB of semi-structured, record oriented data.
- Algorithm: max(temperture).

Hadoop history

The Hadoop solution has been developed since 2008. Hadoop cluster nodes hosted both computing and storage services:

- Distributed File System HDFS
- Distributed algorithm MapReduce
- Huge ecosystem (Hive, Pig,...)

Hortonworks distribution

Advantages [2]:

- Packaged distribution.
- Easy to install with shell scripts.
- Optimised Hadoop parameters.

Virtual cluster setup

- OS tuning [3]: files, FS, network, huge pages.
- Master node:
  - Services: NameNode, SecondaryNameNode, ResourceManager, JobHistoryServer.
  - Resources: 1 node, 4 CPUs, 8 GB of memory, 5 GB of local disk, 10 GB of ephemeral disk, 50 GB of persistent disk.
- Slave node:
  - Services: DataNode, NodeManager.
  - Resources: 3 nodes, 4 CPUs, 16 GB of memory, 5 GB of local disk, 150 GB of persistent disk (2 disks).
- Hadoop parameters tuning.

Algorithm comparison

Analysis of meteorological data (see [8]):

101 maps & 1 reduces (default value).

MapReduce job:

- Java is the most efficient method.
- Virtual cluster does not under perform in comparison with EMR AWS.
- Streaming adds a performance overhead (factor 2 of time).

References