UOC Universitat Oberta de Catalunya

Proyecto Fin de Carrera

OpenNebula y Hadoop: Cloud Computing con herramientas Open Source

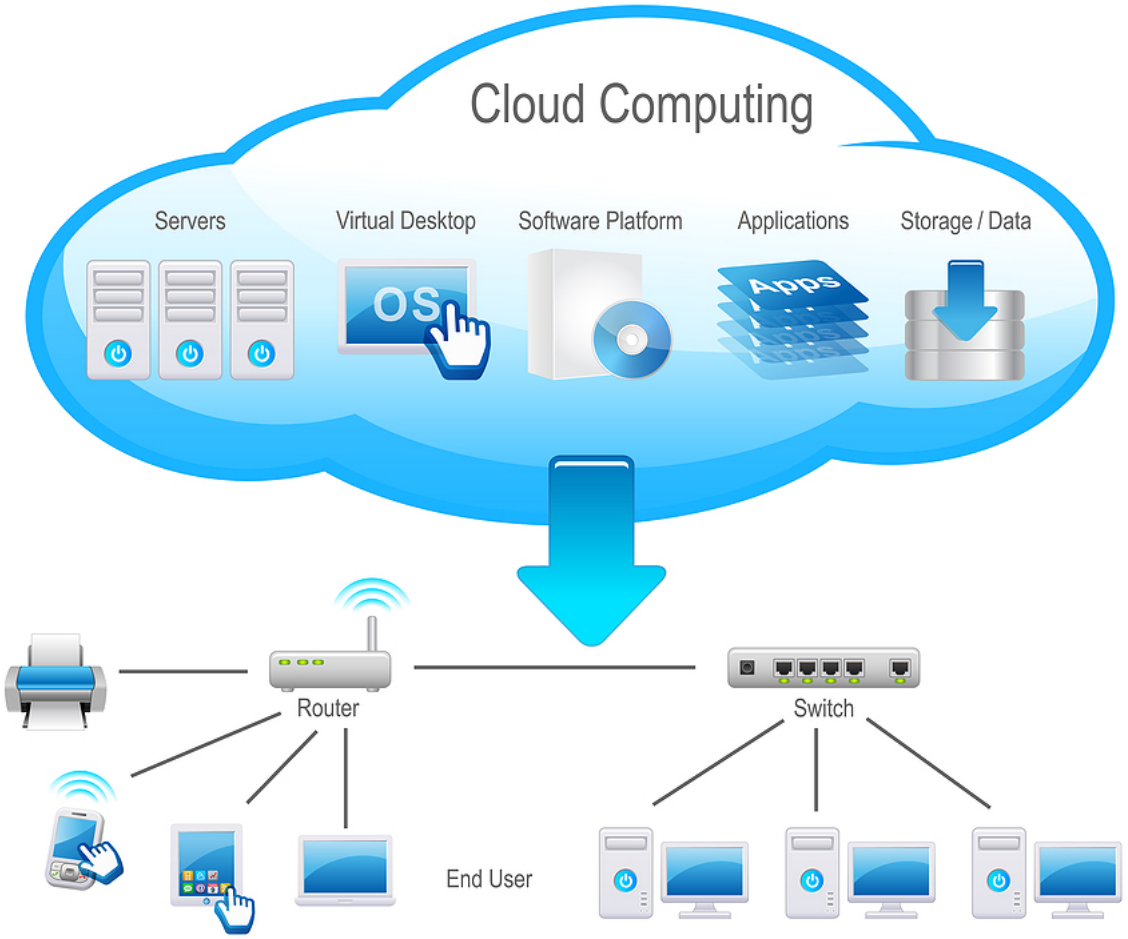
Francisco Magaz Villaverde
Consultor: Víctor Carceler Hontoria

Junio 2012

Contenido

- Introducción
- ¿Qué es Cloud Computing?
- *IaaS*: OpenNebula
- *PaaS*: Hadoop
- Aplicación Práctica

Cloud Computing



Cloud Computing

- Definición
  - Paradigma computacional que pretende el uso compartido de recursos (procesamiento, almacenamiento, servicios) a través de Internet (también redes privadas) de la forma más transparente posible para el usuario.

Cloud Computing

- Características principales
  - Servicio bajo demanda
  - Elasticidad
  - *Pool* de recursos
  - Acceso por red
  - Recursos compartidos

Cloud Computing

Ventajas	Inconvenientes
Escalabilidad	Dependencia del proveedor
Coste	Localización de los datos
Confiabilidad	Protección de la información
Integración	Fiabilidad
Rapidez de despliegue	Cuestiones legales
Simplicidad	

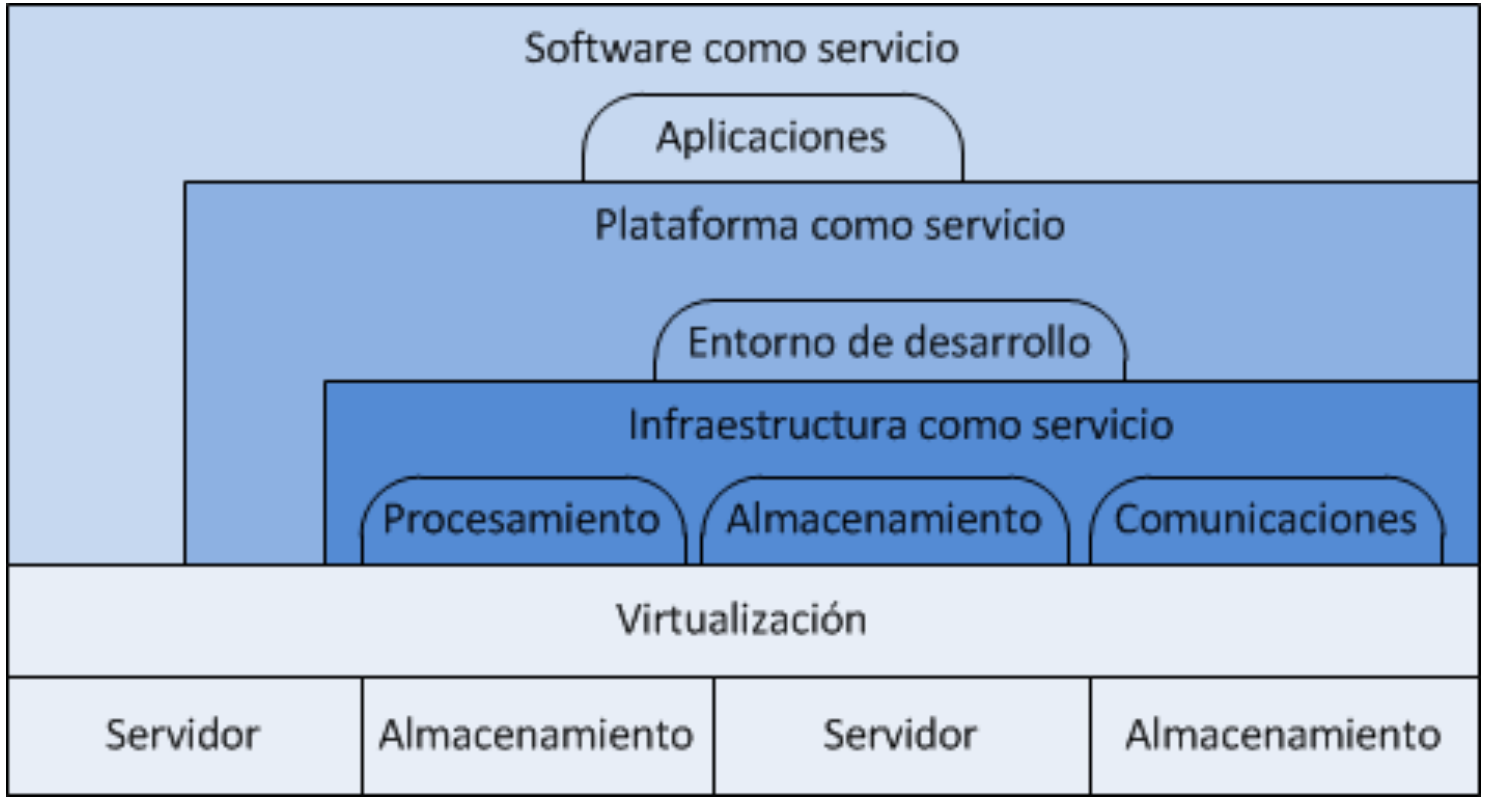
Modelos de despliegue

- **Público**: Todos los recursos proporcionados por el proveedor del servicio.
- **Híbrido**: Unión de una nube pública y privada.
- **Privado**: Toda la infraestructura pertenece al usuario.

Modelos de servicio

- **SaaS**: Software como servicio
  - Proporciona aplicaciones que serán utilizadas directamente por el usuario.
- **PaaS**: Producto como servicio
  - Proporciona Sistemas Operativos, herramientas específicas (ej.: SGBD) y de desarrollo.
- **IaaS**: Infraestructura como servicio
  - Proporciona una infraestructura de computación.

Modelos de servicio



IaaS: OpenNebula

OpenNebula

- Solución Open Source (bajo licencia Apache v2) que permite implementar fácilmente infraestructuras Cloud Computing privadas (también híbridas) según el modelo *IaaS.*

- Proyecto iniciado por la por la Universidad Complutense de Madrid en 2008.

OpenNebula.org
The Open Source Toolkit for Cloud Computing

OpenNebula

- Plataforma de Cloud Computing escalable, segura y rápida de desplegar.
- Consiste en un software que permite desplegar máquinas virtuales sobre un *pool* de máquinas físicas u hipervisores.
- Hipervisores soportados:
  - Xen
  - KVM
  - VMWare

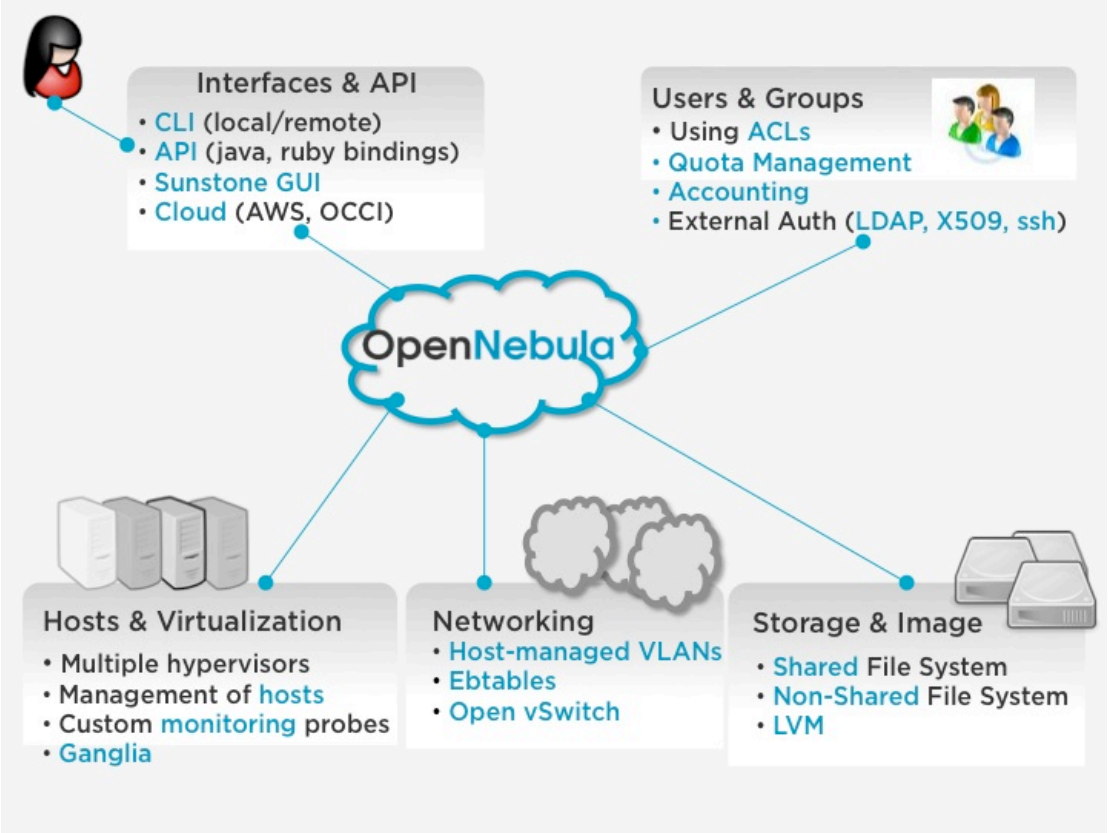
OpenNebula.org
The Open Source Toolkit for Cloud Computing

OpenNebula – Pilares básicos

- Almacenamiento
- Repositorio de plantillas
- Redes virtuales
- Manejo de máquinas virtuales
- Clústeres
- Usuarios y grupos
- API

OpenNebula.org
The Open Source Toolkit for Cloud Computing

OpenNebula – Pilares básicos



OpenNebula.org
The Open Source Toolkit for Cloud Computing

OpenNebula – Hoy en día

- Soporte comercial a través de la empresa C12G.
- Disponible en los repositorios oficiales de Debian, Ubuntu y OpenSuse.
- Uitlizado por Telefonica, FermiLab o el CERN.

OpenNebula.org
The Open Source Toolkit for Cloud Computing

PaaS: Hadoop

Hadoop

- Es un *framework* que permite el tratamiento distribuido de grandes cantidades de datos (del orden de peta bytes) y trabajar con miles de máquinas de forma distribuida.
- Inspirado en la documentación sobre MapReduce y Google File System publicada por Google.

Hadoop - Características

- Económico
- Escalable
- Eficiente
- Confiable

Hadoop - Aspecto clave

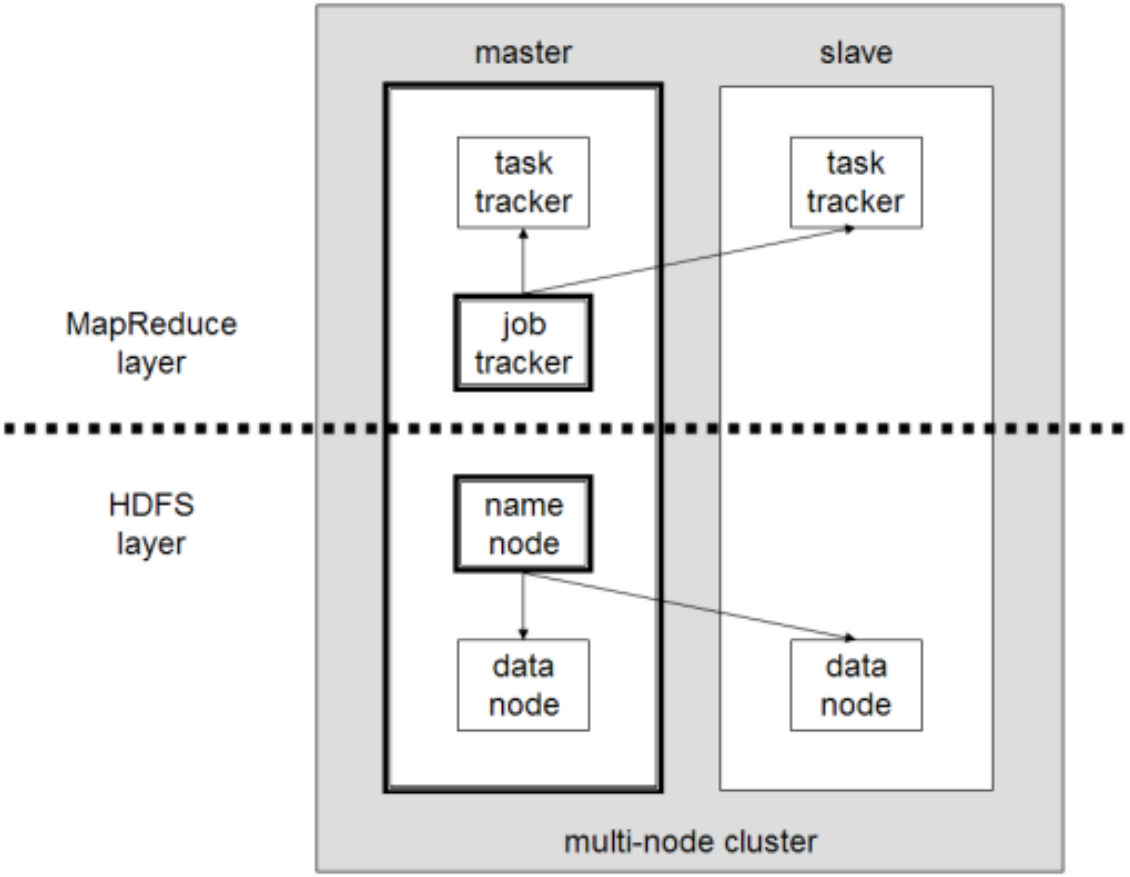
- Hadoop, en lugar de mover los datos hacia donde se hace el procesamiento, Hadoop mueve el procesamiento (Tasks) a donde están los datos.
- Esto reduce el tráfico de información a través de las redes.

Hadoop - Capas

- MapReduce: Procesamiento de la información de forma distribuida.

- Hadoop Distributed File System (HDFS): Almacenamiento de todos los datos repartiéndolos entre cada nodo de la red Hadoop.

Hadoop - Capas



Hadoop – Tipos de nodos

- NameNode:
  - Almacena los metadatos del sistema de ficheros HDFS y donde se almacenan los bloques de datos.
- DataNode:
  - Almacena los bloques de datos de HDFS.
- JobTracker:
  - Gestiona las tareas MapReduce eligiendo que nodo ejecuta cada una.
- TaskTracker:
  - Ejecuta las tareas Map o Reduce siguiendo instrucciones del JobTracker.

Hadoop - MapReduce

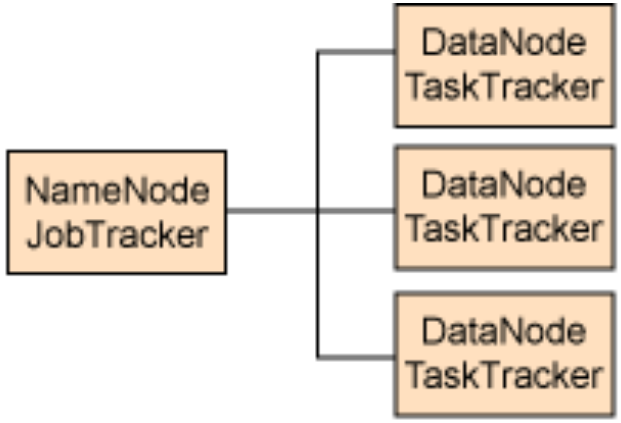
- Map and Reduce es un algoritmo de la categoría *divide y vencerás*.
- Se basa en la programación funcional, en las funciones Map y Reduce de los lenguajes funcionales.
  - Map(k1,v1) → list(k2,v2)
  - Reduce(k2, list (v2)) → list(v3)

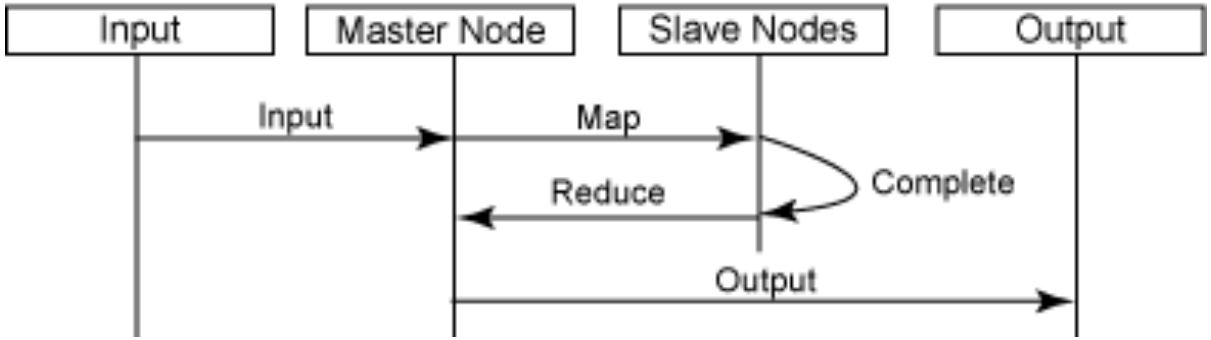
Hadoop - MapReduce

- MapReduce transforma una lista de clave/ valor en una lista de valores.

- Un ejemplo de aplicación: Contar palabras en un documento de entrada (útil para la indexación de ficheros)

Hadoop - MapReduce





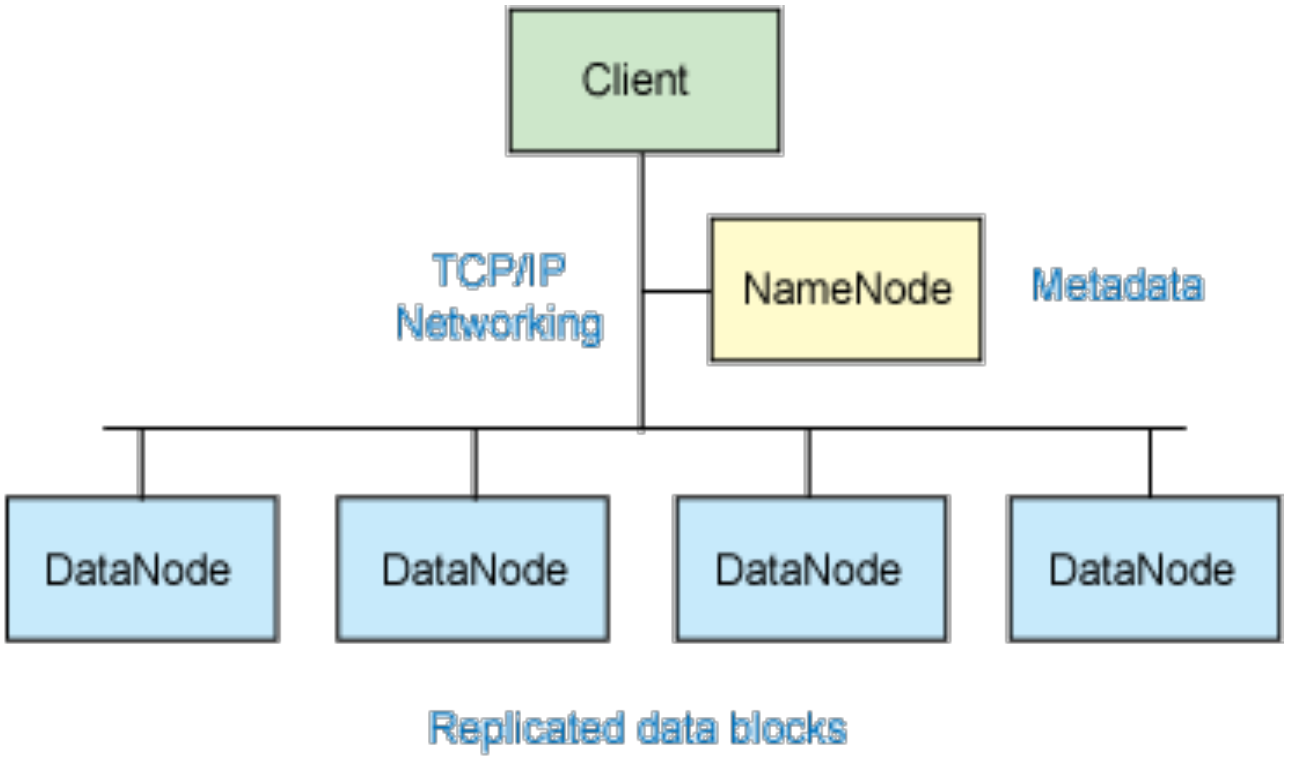
Hadoop - HDFS

- Principal sistema de almacenamiento utilizado por Hadoop.
- Crea múltiples replicas de los bloques de datos y los distribuye entre los nodos de un clúster.
- Desde la perspectiva del usuario, HFS se muestra como un sistema de ficheros tradicional pudiendo llevar a cabo operaciones CRUD.

Hadoop - HDFS

- Recuperación antes fallos de hardware
- Acceso en streaming
- Grandes volúmenes de datos
- Coherencia simple
- Portabilidad

Hadoop - HDFS


Client
TCP/IP
Networking
NameNode
Metadata
DataNode
DataNode
DataNode
DataNode
Replicated data blocks

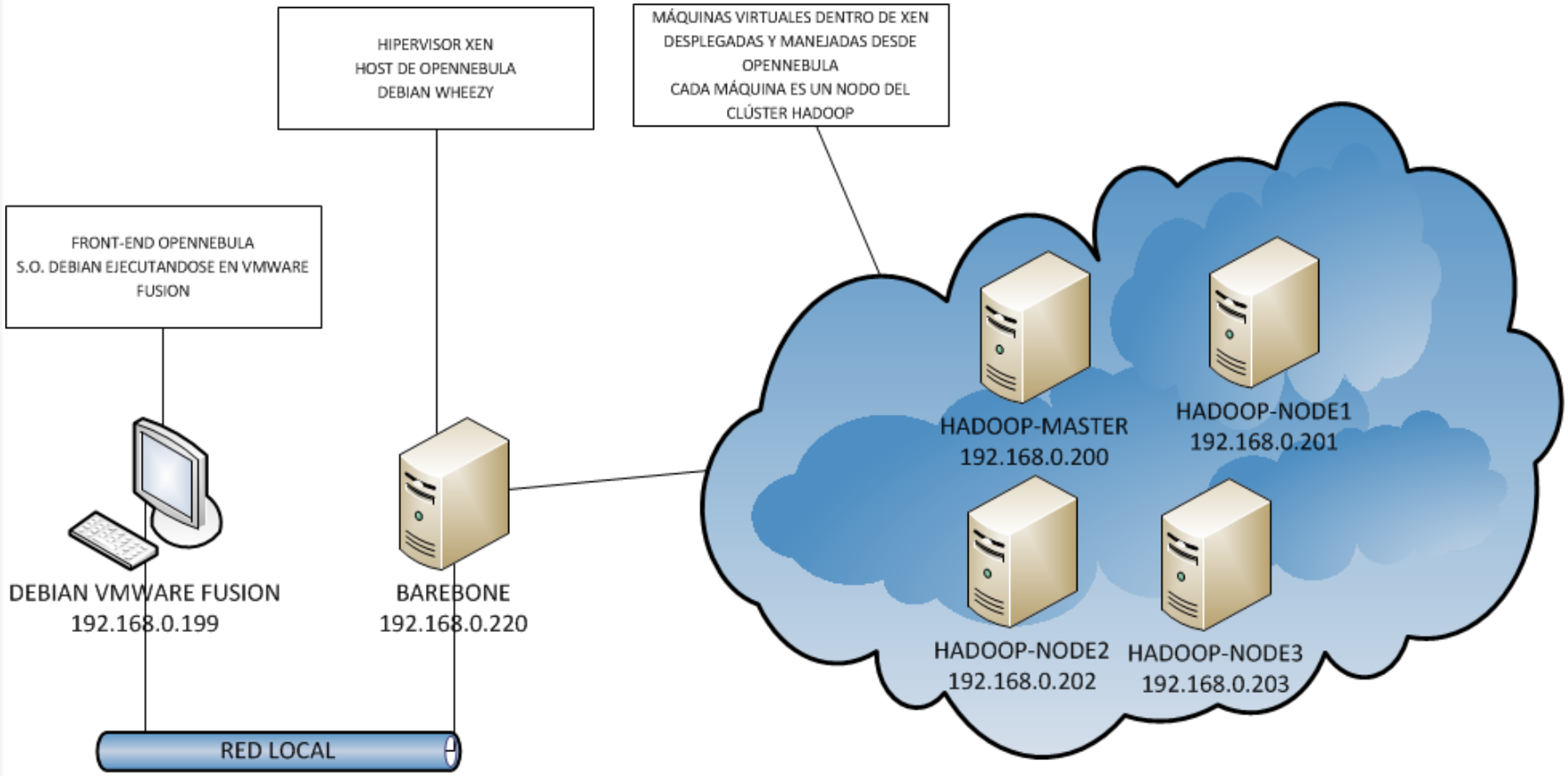
Aplicación Práctica

Despliegue de un Clúster Hadoop utilizando OpenNebula con hipervisor Xen

Aplicación Práctica - Material

- Ordenador Barebone. Se ejecuta el hipervisor Xen controlado por OpenNebula.
  - Hardware:
    - Procesador Intel Core i3 con VT-x a 3,10 GHz.
    - 8GB de memoria RAM.
    - 1 TB de disco duro.
    - Un Interfaz de red.
  - Software:
    - Debian Wheezy
    - Xen 4.1
- Mac-Mini: Una máquina virtual Debian corriendo en VMWare Fusion hará el rol de front-end de OpenNebula.
  - Procesador Intel Core 2 Duo. (1 procesador para la máquina virtual)
  - 2 GB de memoria RAM (512 MB para la máquina virtual)
  - 350 GB de disco duro (15 GB para la máquina virtual)
  - Un Interfaz de red.

Aplicación Práctica



Aplicación Práctica

- Principal hándicap:
  - Limitaciones de una maqueta de laboratorio

- Aún así se ha podido desplegar las máquinas virtuales en OpenNebula.

- Sobre estas máquina virtuales (nodos del clúster Hadoop) se han ejecutado ejemplos de demo de tareas MapReduce.

Aspectos a destacar en el despliegue

- El inicio es difícil ya que hay que integrar varios sistemas:
  - Virtualización con OpenNebula
  - Hadoop con OpenNebula
- Debido a la naturaleza Open Source de las tecnologías elegidas la documentación disponible es abundante y útil.
- Una vez desplegados y configurados los primeros nodos, los siguientes son más sencillos, siendo posible automatizar dicho proceso.

En el mundo real

- Esta infraestructura implantada en un entorno real puede utilizarse para:
  - Indexación masiva de archivos.
  - Almacenamiento de grandes volúmenes de datos con tolerancia a fallos.
  - Minería de datos.

Posibles mejoras

- Utilización de redes virtuales.
- Uso de la contextualización.

Conclusiones

- Cloud Computing no es solo una tendencia, es una realidad.
- Cuenta con el apoyo de grandes empresas (Amazon, Google, etc.)
- Las nubes privadas son las preferidas por las grandes empresas actualmente.
- Es posible desplegar y dar servicio a terceros mediante una infraestructura Cloud Computing utilizando herramientas Open Source.
- Existen mercados por explotar dentro de esta tecnología.

Gracias por su atención