JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

NAME :MWANG’OMBE DAFTON KIRIGHA

REG NO :BIT-001-0738/2013

UNIT NAME :CLOUD COMPUTING

UNIT CODE :BIT 2225

**Cloud security**

This is a term used to define an evolving sub-domain of computer security, network security and more broadly information security. It is a broad set of policies, technologies and controls deployed to protect data, applications and the associated infrastructure of cloud computing.

**Network security issues in cloud computing**

Network security concerns in cloud computing can be categorised into two broad categories; those faced by the cloud provider and those faced by the customers.

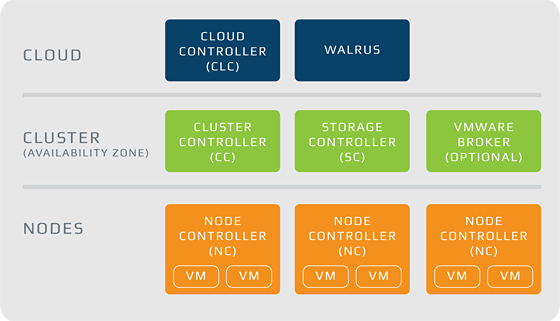
Probable loss of confidentiality for the customers. This is because once a firm stores confidential information on a cloud, it loses its physical ability to access the servers hosting the information. This exposes the data stored to risks from inside attacks.

 The provider must ensure that their infrastructure is secure and that their clients’ data and applications are protected while the user must take measures to fortify their application and use strong passwords and authentication measures.

In order to conserve resources, cut costs, and maintain efficiency, Cloud Service Providers often store more than one customer's data on the same server. As a result there is a chance that one user's private data can be viewed by other users (possibly even competitors). To handle such sensitive situations, cloud service providers should ensure proper data isolation and logical storage segregation.

Virtualization alters the relationship between the OS and underlying hardware - be it computing, storage or even networking. This introduces an additional layer - virtualization - that itself must be properly configured, managed and secured. Specific concerns include the potential to compromise the virtualization software.

**Five eucalyptus cloud components**



**The node controller (NC)**

NC is written in C and hosts the virtual machine instances and manages the virtual network endpoints. It downloads and caches images from Walrus as well as creates and caches instances. There is no theoretical limit to the number of Node Controllers per cluster, performance limits do exist.

**The Cloud Controller (CLC)**

This is a java program that offers EC2-compatible interfaces, as well as a web interface to the outside world. In addition to handling incoming requests, the CLC acts as the administrative interface for the cloud managements and performs high-level resource scheduling and system accounting. The CLC accepts user Application Programming Interface (API) requests for command-line interfaces like euca2tools or GUI-based tools like Eucalyptus User Console and manages network resources.

**Walrus**

Walrus, also written in java is the Eucalyptus equivalent to AWS Simple Storage Service (S3). It offers persistent storage to all of the virtual machines in the Eucalyptus cloud and can be used as a simple HTTP put/get storage as a service solution. There are no data type restrictions for walrus, and it can contain images, volume snapshots and application data. Only one walrus can exist per cloud.

**The cluster controller**

It is written in C and acts as the front end for a cluster within a eucalyptus cloud and communicates with the storage controller and node controller. It manages instance execution and service level agreements per cluster.

**The storage controller**

It is written in java and is the eucalyptus equivalent to AWS EBS. It communicates with the cluster controller and the node controller and manages eucalyptus block volumes and snapshots to the instances within its specific cluster. Iof an instance requires writing persistent data to memory outside of the cluster; it would need to write to Walrus which is available to any instance in any cluster.

**The VMware Broker**

It is an optional component that provides an AWS-compatible interface for VMware environments and physically runs on the Cluster Controller. The VMware Broker overlays existing EXN/ESXi hosts and transforms eucalyptus Machine Images (EMIs) to VMware virtual disks. The VMware broker mediates interactions between the cluster controller and VMware and can connect directly to either ESX/ESXi hosts or to vCenter server.o

The front end of the eucalyptus cloud architecture is implemented by the Cloud controller and the Walrus, while the back end is implemented by the cluster controller, the storage controller and the node controller.

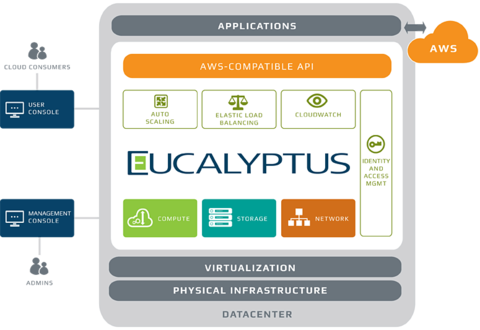
**Compare Eucalyptus, HPCC and HADOOP**

**EUCALYPTUS**

EUCALYPTUS - Elastic Utility Computing Architecture Linking Your Programs To Useful Systems - is an open source software infrastructure for implementing on-premise clouds on existing Enterprise IT and service provider infrastructure. It is a free and open- source computer software for building Amazon Web Services- compatible private and hybrid cloud computing environments marketed by the company Eucalyptus Systems.

**Software Architecture**

Eucalyptus commands can manage either Amazon or Eucalyptus instances. Users can also move instances between a Eucalyptus private cloud and the Amazon Elastic Compute Cloud to create a hybrid cloud. Hardware virtualization isolates applications from computer hardware details.

[](http://en.wikipedia.org/wiki/File:Eucalyptus_Architecture.png)

**HPCC**

High performance computing environments which utilize supercomputers and computer clusters to address complex computational requirements, support applications with significant processing time requirements, or require processing of significant amounts of data. Supercomputers have generally been associated with scientific research and compute-intensive types of problems, but more and more supercomputer technology is appropriate for both compute-intensive and data-intensive applications.

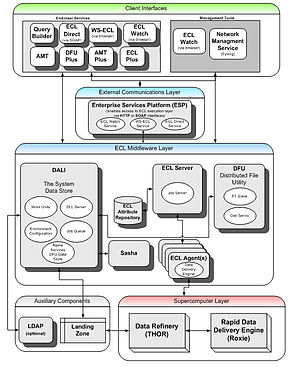
A new trend in supercomputer design for high-performance computing is using clusters of independent processors connected in parallel. Many computing problems are suitable for parallelization, often problems can be divided in a manner so that each independent processing node can work on a portion of the problem in parallel by simply dividing the data to be processed, and then combining the final processing results for each portion.

This type of parallelism is often referred to as data-parallelism, and data-parallel applications are a potential solution to terabyte scale data processing requirements.Data-parallelism can be defined as a computation applied independently to each data item of a set of data which allows the degree of parallelism to be scaled with the volume of data. The most important reason for developing data-parallel applications is the potential for scalable performance in high-performance computing, and may result in several orders of magnitude performance improvement.

**System Architecture**

The HPCC system architecture includes two distinct cluster processing environments, each of which can be optimized independently for its parallel data processing purpose.

The first of these platforms is called a data refinery whose overall purpose is the general processing of massive volumes of raw data of any type for any purpose but typically used for data cleansing and hygiene, extract, transform and load the processing of the raw data, record linking and entity resolution, large-scale ad-hoc complex analytics, and creation of keyed data and indexes to support high-performance structured queries and data warehouse applications. The data refinery is also referred to as ‘Thor’ a reference to the mythical Norse god of thunder with the large hammer symbolic of crushing large amounts of raw data into useful information. A Thor cluster is similar in its function, execution environment, file systems, and capabilities to the Google and Hadoop MapReduce platforms.

[](http://en.wikipedia.org/wiki/File:Fig4b_HPCC.jpg)

The second of the parallel data processing platforms is called Roxie and functions as a rapid data delivery engine. This platform is designed as an online high-performance structured query and analysis platform or data warehouse delivering the parallel data access processing requirements of online applications through Web services interfaces supporting thousands of simultaneous queries and users with sub-second response times. Roxie utilizes a distributed indexed file system to provide parallel processing of queries using an optimized execution environment and file system for high-performance online processing. A Roxie cluster is similar in its function and capabilities to Hadoop with HBase and Hive capabilities added, and provides for near real time predictable query latencies. Both Thor and Roxie clusters utilize the ECL programming language for implementing applications, increasing continuity and programmer productivity.

**HADOOP**

Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software’s ability to detect and handle failures at the application layer.

Hadoop makes it possible to run applications on systems with thousands of nodes involving thousands of [terabyte](http://searchstorage.techtarget.com/definition/terabyte)s. Its distributed file system facilitates rapid [data transfer rate](http://searchunifiedcommunications.techtarget.com/definition/data-transfer-rate)s among nodes and allows the system to continue operating uninterrupted in case of a node failure. This approach lowers the risk of catastrophic system failure, even if a significant number of nodes become inoperative.

## High-level architecture

Apache Hadoop has two pillars:

* **YARN**- Yet Another Resource Negotiator (YARN) assigns CPU, memory, and storage to applications running on a Hadoop cluster. The first generation of Hadoop could only run MapReduce applications. YARN enables other application frameworks (like Spark) to run on Hadoop as well, which opens up a wealth of possibilities.
* [**HDFS**](http://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/) - Hadoop Distributed File System (HDFS) is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system.

Hadoop is supplemented by an ecosystem of Apache projects, such as Pig, Hive, and Zookeeper that extend the value of Hadoop and improves its usability.

Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant.