Lesson 3
Virtualization

- Reading list
  - Cloud Computing bible, B.Sosinsky, John Wiley & Sons, 2010 (ch
  - Cloud Computing: principles and paradigms, R.Buyya, 2011
  - Service Oriented Architecture FOR dummies, Judith Hurwitz, Robin Bloor,Marcia Kaufman, and Dr. Fern Halper, 2ND EDITION

What is virtualization?

- Virtualize /verb/ →Convert something to a computer-generated simulation of reality

---

- When using CC, accessing pooled resources using a technique called virtualization
- Virtualization assigns a logical name for a physical resource and then provides a pointer to that physical resource when a request is made
- Virtualization is dynamic, it means mapping can be assigned based on rapidly changing conditions
- Virtualization is accessible, which means changes to a mapping assignment can be instantly (nearly)

What is virtualization?(continued)

- Virtualization is one of the hardware reducing, cost saving and energy saving technology that is rapidly transforming the IT landscape and fundamentally changing the way that people compute.
- With VMware virtualization solutions you can reduce IT costs while increasing the efficiency, utilization and flexibility of their existing computer hardware.
- With Virtualization it is possible to run multiple operating systems and multiple applications on the same SERVER at the same time, increasing the utilization and flexibility of hardware.

Energy saving

The Reality:

- Most servers only use **5-15%** of their capabilities on average, while consuming **60-90%** of their peak power.

The Solution - Virtualization:

- Use one server to host multiple applications.
- Reduce energy consumption
- Reduce CO2 emissions

Running fewer, highly utilized servers frees up space and power. Less space and power is better for environment and saves money.

Borrowed from : Fred Wuensch

Virtualization characters on CC

- Access → A client can request to a cloud service from any location
- Application → A cloud has multiple application instances and direct requests to an instance based on conditions
- CPU → computers can be partitioned into a set of VMs with each machine being assigned a workload.
- Storage → Data is stored across storage devices

Characteristics of virtualized environments

- There are three major components
  - Guest
    - Represents the system component
  - Host
    - Represents the original environment, where the guest is supposed to be managed
  - Virtualization layer OR Virtual machine Manager
    - Responsible for recreating the environment where he guest will operate

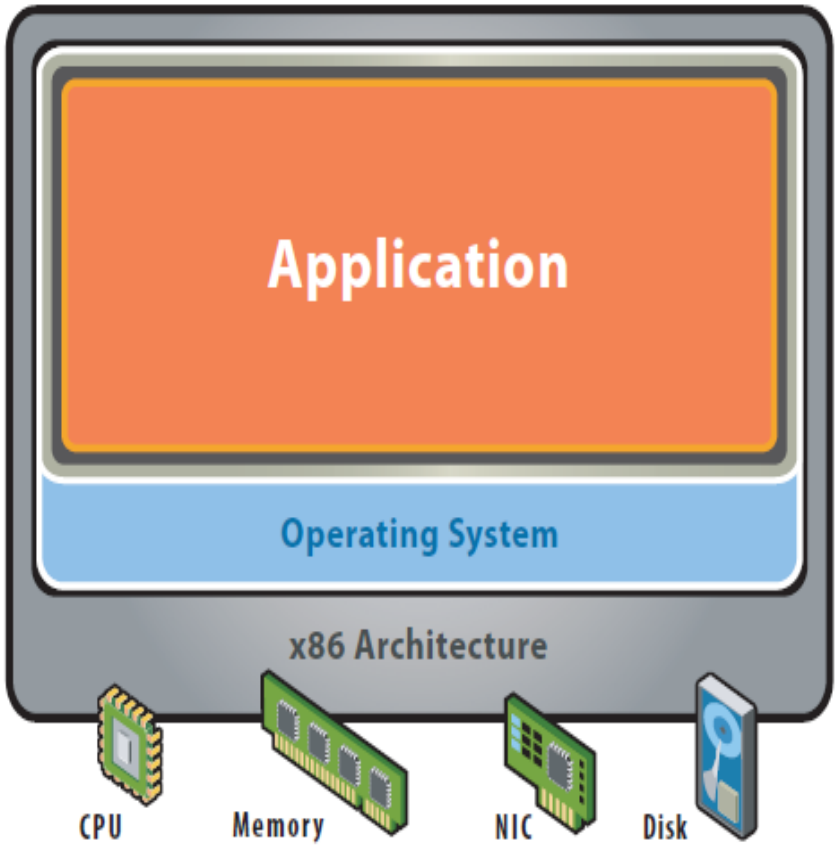
Why use virtualization

- Virtualization is the key enabler of the first four attributes of CC, which are
  - Service-based : where clients are abstracted from service providers through service interface
  - Scalable and elastic : Services can be altered to affect capacity and performance on demand
  - Shared Services : Resources are pooled to create greater efficiencies
  - Metered usages : pay as you go

Why use virtualization (continued)

- Systems and storages can be provided as much as needed from centralized infrastructure
- Costs are assessed by metered basis
- Multi tenancy is enabled
- Resources are scalable with agility

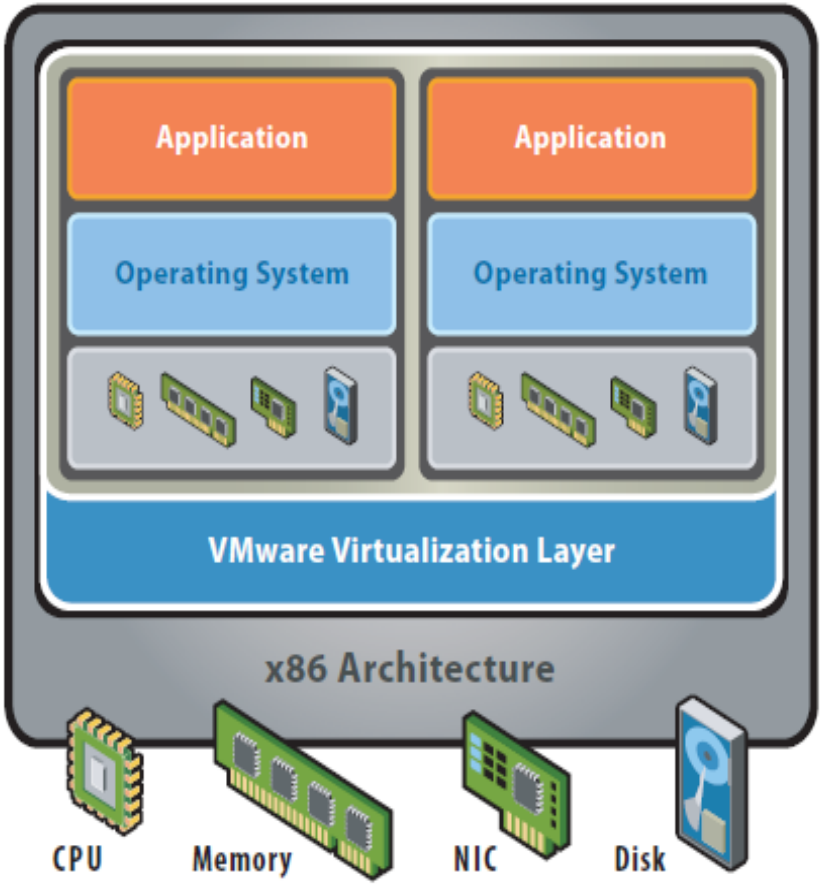
Before Virtualization



- Single OS image per machine
- Software and hardware tightly coupled
- Running multiple applications on same machine often creates conflict
- Inflexible and costly infrastructure

Borrowed from : MARKANA MEHUL K

After Virtualization



- Hardware-independence of operating system and applications
- Virtual machines can be provisioned to any system
- Can manage OS and application as a single unit by encapsulating them into virtual Machines

Borrowed from : MARKANA MEHUL K

First feature of virtualized environments

- **Increased security**
  - Control the execution of a guest
  - All operations are performed against the virtual machine
  - Virtual machine control and filter the activity
  - Prevent any harmful operation
  - E.g. applets run in a sandboxed version of JVM

Second feature of virtualized environments

- **Managed execution**
  - **Sharing**: creation of separate computing environments within the same host. E.g. in datacenters reduce the number of active servers and limit power consumption
  - **Aggregation**: a group of separate hosts can be tied together and represented to guests as a single virtual host
  - **Emulation**: a completely different environment with respect to the host can be emulated, thus allowing the execution of guest programs requiring specific characteristics that are not present in the physical host.
    - E.g. use of hardware to imitate the function of another hardware device for the purpose of connecting devices to one another or connecting to a mainframe computer.
  - **Isolation** : Virtualization allows providing guests—whether they are operating systems, applications, or other entities—with a completely separate environment, in which they are executed.
    - It allows multiple guests to run on the same host without interfering with each other.
    - It provides a separation between the host and the guest. The virtual machine can filter the activity of the guest and prevent harmful operations against the host.

Third feature of virtualized environments

- **Portability**
  - In the case of a hardware virtualization solution, the guest is packaged into a virtual image that, in most cases, can be safely moved and executed on top of different virtual machines. Except for the file size, this happens with the same simplicity with which we can display a picture image in different computers.
  - Virtual images are generally proprietary formats that require a specific virtual machine manager to be executed. In the case of programming-level virtualization, as implemented by the JVM or the .NET runtime, the binary code representing application components (jars or assemblies) can be run without any recompilation on any implementation of the corresponding virtual machine.
  - This makes the application development cycle more flexible and application deployment very straightforward: One version of the application, in most cases, is able to run on different platforms with no changes. Finally, portability allows having your own system always with you and ready to use as long as the required virtual machine manager is available.

Forth feature of virtualized environments

- performance tuning
  - control the performance of the guest by finely tuning the properties of the resources exposed through the virtual environment.
  - This capability provides a means to effectively implement a quality-of-service (QoS) infrastructure that more easily fulfills the service-level agreement (SLA) established for the guest.
    - For instance, software-implementing hardware virtualization solutions can expose to a guest operating system only a fraction of the memory of the host machine or set the maximum frequency of the processor of the virtual machine.

Category (Taxonomy) of virtualized techniques

- 1- Execution virtualization
  - *Process level*
    - Technique → Emulation , Model → Application
    - Technique → High-level VM , Model → programming language
    - Technique → Multiprogramming , Model → Operating System
  - *System level*
    - Hardware assisted virtualization
    - Full virtualization
    - Para virtualization
    - Partial virtualization

Hardware

Category (Taxonomy) of virtualized techniques

- 2- Storage

Category (Taxonomy) of virtualized techniques

- Network

Concepts

- Load balancing → the technology used to distribute service requests to resources
  - Can be implemented in hardware or in software
  - It is an optimization technique
  - Can be used to increase utilization, reduce response time, avoid system overload
  - Provides fault tolerance when coupled with a failover mechanism
  - Use different mechanism to assign service direction

Load balancing mechanism

- The load balancer listens to a network port for service requests
- When the request arrives, the load balancer use a scheduling algorithm to assign where the request is sent
- Typical scheduling Algorithms
  - Round robin
  - Weighted round robin
  - Fastest response time
  - Least connections
  - Weighted least connections

Round robin

- (RR) is one of the algorithms employed by process and network schedulers in computing.
- As the term is generally used, time slices are assigned to each process in equal portions and in circular order, handling all processes without priority (also known as cyclic executive).
- RR scheduling is simple, easy to implement, and starvation-free.
- RR scheduling can also be applied to other scheduling problems, such as data packet scheduling in computer networks. It is an operating system concept.
- The name of the algorithm comes from the round-robin principle known from other fields, where each person takes an equal share of something in turn.

Weighted round robin

- Each packet flow or connection has its own packet queue in a network interface controller.
- It is the simplest approximation of generalized processor sharing (GPS).
- While GPS serves infinitesimal amounts of data from each nonempty queue, WRR serves a number of packets for each nonempty queue
- number=normalized(weight/meanpacketsize)

Fastest response time

- The Fastest method passes a new connection based on the fastest response time of all servers.
- This method may be particularly useful in environments where servers are distributed across different logical networks.
- On the BIG-IP, only servers that are active will be selected.

Least connections

- The system passes a new connection to the server that has the least number of current connections.
- Least Connections methods work best in environments where the servers or other equipment you are load balancing have similar capabilities.
- This is a dynamic load balancing method, distributing connections based on various aspects of real-time server performance analysis, such as the current number of connections per node or the fastest node response time.

Weighted least connections

- Select pool members or nodes based on the number of active connections &server capacity.
- Specifies that the system uses the value you specify in **Connection Limit** to establish a proportional algorithm for each pool member.
- The system bases the load balancing decision on that proportion and the number of current connections to that pool member.
- This algorithm requires all pool members to have a non-zero connection limit specified.
- specifies that the system uses the value you specify in the node's **Connection Limit** setting and the number of current connections to a node to establish a proportional algorithm.
- If all servers have equal capacity, these load balancing methods behave in the same way as the Least Connections methods.

Application Delivery Controller (ADC)

- ADC is a combination of load balancer and application server
- It is a server placed between a firewall or router and provides web services
- ADC is assigned a Virtual IP address, that maps to a pool of servers based on application specific criteria
- Services provided by an ADC includes data compression, content caching, security, …

References

- Cloud Computing bible, Barrie Sosinsky, Wiley publication, 2011
- Mastering Cloud Computing, Rajkumar Buyya, Elsevier , 2013
- http://www.slideshare.net/markanamehul/virtualization-in-cloud-computing-ppt
- https://www.event-plans.com/Uploads/files/GSA/Fred%20Wuensch%20-%20Total%20Virtualization%20and%20Cloud%20Computing.ppt
- https://devcentral.f5.com/articles/intro-to-load-balancing-for-developers-ndash-the-algorithms