

EXPERIMENT NO . 3

Aim: To study and implement Bare-metal Virtualization using XEN

Theory:

Bare Metal Hypervisors

Bare-metal or Type 1 hypervisors are designed to run directly on the physical hardware of a host machine without the need for a host operating system. They perform several key functions to enable the creation and management of virtual machines (VMs) efficiently. Bare-metal hypervisors, such as VMware ESXi, Microsoft Hyper-V (when installed as a standalone hypervisor), and Xen, play a crucial role in virtualization by providing a stable and efficient platform for running multiple virtualized instances on a single physical server.

Here are the functions performed by bare-metal hypervisors:

1. **Resource Management:** Bare-metal hypervisors directly manage and allocate physical hardware resources, including CPU, memory, storage, and networking, among virtual machines. This direct control enhances resource efficiency and allows for optimal utilization.
2. **Virtual Machine Creation:** Bare-metal hypervisors are responsible for creating and managing virtual machines. They provide an environment for multiple operating systems to run independently on the same physical server.
3. **Hypervisor Layer:** The hypervisor layer itself is a critical function. It abstracts the underlying hardware and facilitates communication between virtual machines and physical resources.
4. **Isolation:** One of the fundamental functions is to ensure isolation between virtual machines. Each VM operates independently, and issues or failures in one VM do not affect others on the same host.
5. **Hardware Abstraction:** The hypervisor abstracts the physical hardware, presenting virtualized versions of CPUs, memory, and other resources to each VM. This abstraction allows VMs to run on different hardware without being tied to specific configurations.
6. **Performance Optimization:** Bare-metal hypervisors are optimized for performance, as they run directly on the hardware without the overhead of a host operating system. This leads to better resource utilization and improved VM performance.
7. **Security Management:** Bare-metal hypervisors incorporate security measures to ensure the integrity and isolation of virtual machines. This includes features like secure boot, encrypted storage, and access controls.

8. Live Migration: Some bare-metal hypervisors support live migration, allowing virtual machines to be moved between physical hosts without disruption to operations. This feature is beneficial for load balancing, maintenance, and disaster recovery.

9. Snapshot Management: Bare-metal hypervisors often support snapshot functionality, enabling the capture of the current state of a virtual machine. Snapshots can be used for backup, testing, and recovery purposes.

10. Networking and Connectivity: Managing networking resources is a crucial function. Bare-metal hypervisors provide tools and configurations to handle network connectivity for virtual machines, including virtual LANs, virtual switches, and network adapters.

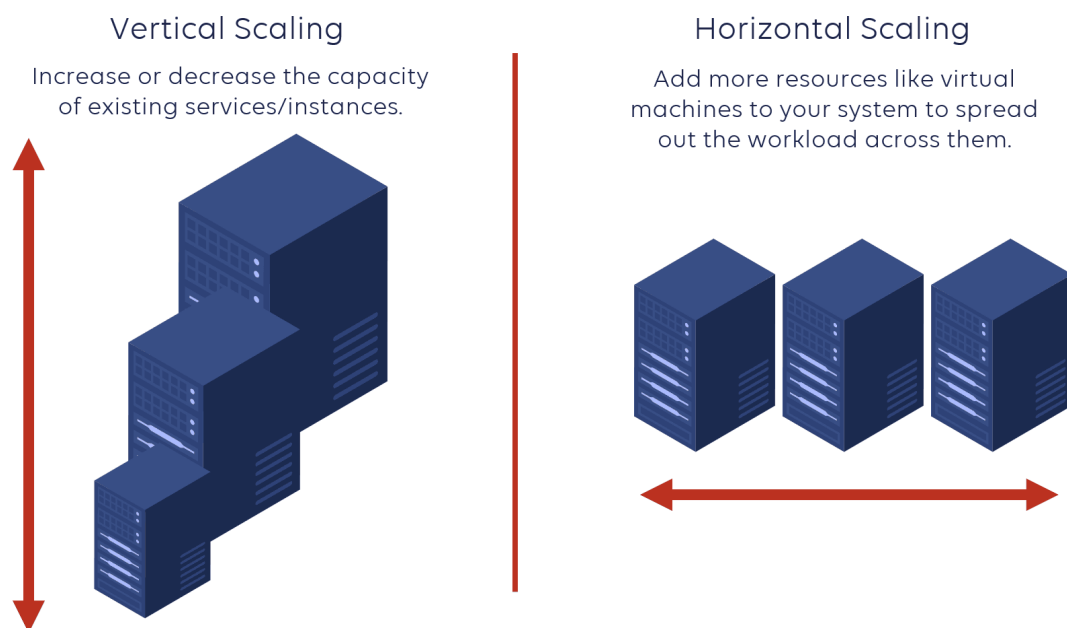
Hosted v/s Bare metal Hypervisors

Feature	Hosted Hypervisor (Type 2)	Bare-Metal Hypervisor (Type 1)
Installation	Installs on top of a host operating system	Installs directly on the physical hardware
Resource Management	Relies on the host OS for resource management	Directly manages and allocates physical resources
Performance	Typically has more overhead	Generally offers better performance
Resource Utilization	May have higher resource consumption due to running on top of an OS	Efficient resource utilization as it has direct access to hardware
Isolation	Limited isolation between VMs	Strong isolation between VMs
Platform Support	Supports various host operating systems	Primarily used with specific platforms (e.g., enterprise servers)
Ease of Use	User-friendly with graphical interfaces	Often managed through command-line tools; additional management tools available
Networking Options	Provides a range of networking options and configurations through a GUI	Networking configurations often handled through command-line tools, may require additional setup
Snapshot Management	Provides snapshot functionality for creating and managing snapshots of VM states	Offers snapshot functionality for creating and restoring VM states
USB Support	Supports USB passthrough for connecting USB devices to VMs	Offers USB passthrough, but setup may be more manual compared to Type 2 hypervisors
Community/Support	Large community support and documentation	Community support, especially strong within the Linux community
Performance Use Cases	Suited for development, testing, and desktop virtualization	Commonly used in server environments, data centers, and cloud infrastructure
Examples	VirtualBox, VMware Workstation	VMware ESXi, Microsoft Hyper-V (standalone), Xen

Horizontal and Vertical Scaling

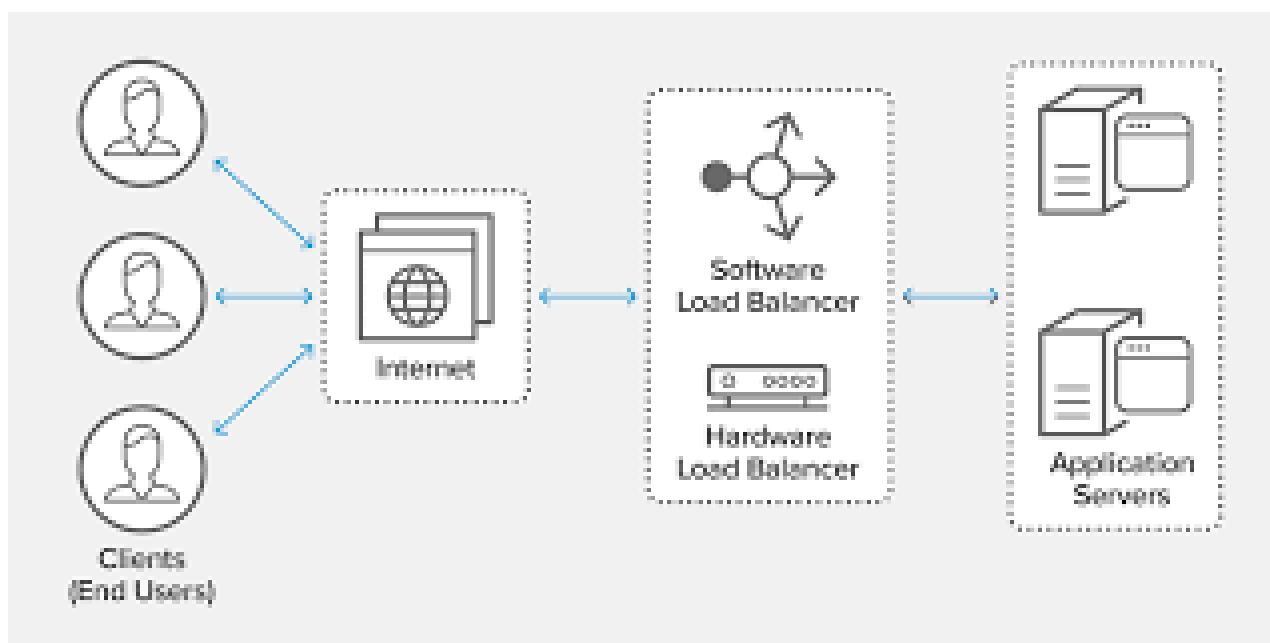
Horizontal Scaling: Horizontal scaling, also known as scaling out, refers to the practice of increasing the capacity or performance of a system by adding more machines or nodes to a distributed network. In horizontal scaling, new instances of servers or entire systems are added to distribute the load and improve overall performance. This approach is often associated with cloud computing, where additional virtual machines or containers are added dynamically to handle increased demand.

Vertical Scaling: Vertical scaling, or scaling up, involves increasing the capacity or performance of a single machine or server by adding more resources to it. This typically includes upgrading the existing hardware components, such as adding more RAM, CPU power, or storage capacity. Vertical scaling is suitable for applications or systems that require more processing power on a single machine, but it has limitations in terms of scalability compared to horizontal scaling.



Load Balancing:

Load balancing is a technique used to distribute incoming network traffic or workload across multiple servers or resources to ensure optimal utilization and prevent any single resource from being overwhelmed. The primary goal of load balancing is to enhance the performance, availability, and reliability of a system. Load balancers can operate at various layers of the network stack and distribute requests based on factors like server health, current load, or a predefined algorithm. This ensures efficient resource usage and helps prevent bottlenecks in a system, improving its overall responsiveness and scalability.



Auto Scaling

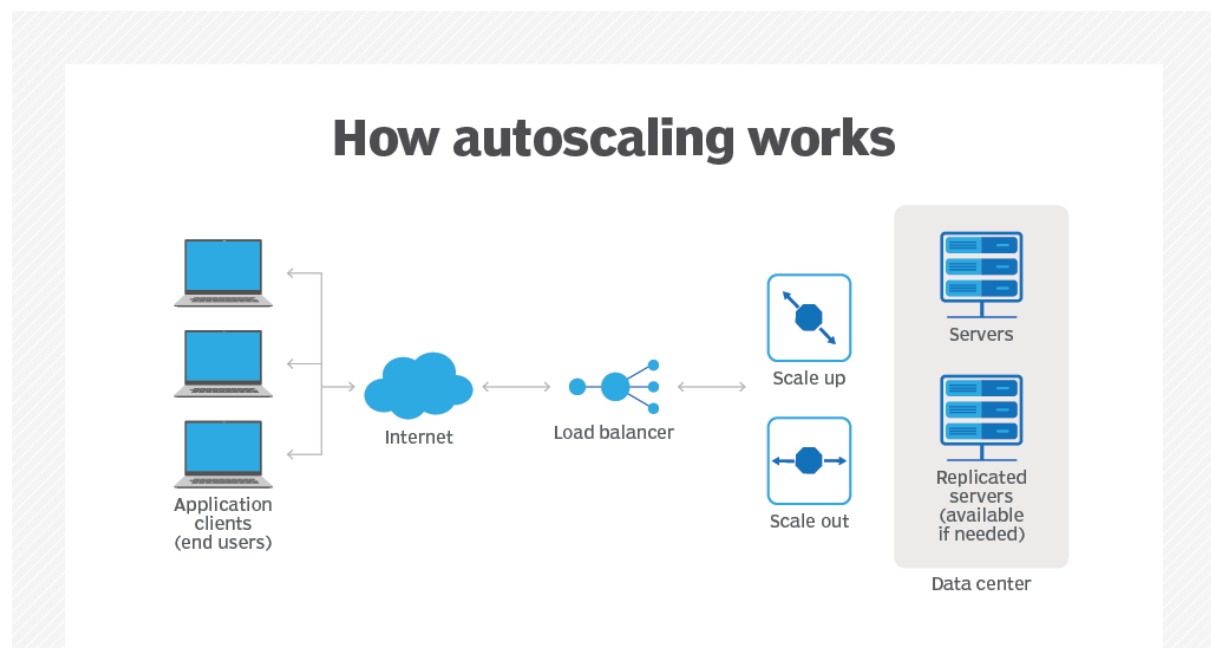
Auto scaling is a dynamic cloud computing feature designed to address the fluctuating demands of applications and services by automatically adjusting the allocated computing resources. The essence of auto scaling lies in its ability to monitor key performance metrics, such as CPU utilization, network traffic, and application response time. Based on predefined scaling policies, which outline conditions and actions for resource adjustments, the auto scaling system can seamlessly scale resources up or down.

When demand increases, a process known as "scaling out" occurs, where additional resources, such as virtual machines or containers, are automatically provisioned to handle the heightened workload. This ensures that the application or service can meet performance requirements during peak periods. Conversely, during periods of lower

demand, the auto scaling system engages in "scaling in," intelligently removing excess resources to optimize cost and resource utilization.

Auto scaling often integrates with load balancers to evenly distribute incoming traffic across the available resources, facilitating efficient load distribution among newly added instances. The flexibility of auto scaling allows users to configure policies based on specific criteria, such as time of day, day of the week, or custom metrics tailored to the application's unique performance characteristics.

One of the key benefits of auto scaling is its contribution to fault tolerance. By automatically replacing failed instances or resources, it enhances the system's resilience to unexpected failures, thereby ensuring high availability and reliability. Moreover, auto scaling plays a pivotal role in cost optimization by dynamically adjusting resources in response to the current demand, preventing unnecessary spending on idle resources and promoting efficiency in resource allocation. As a foundational feature provided by major cloud service providers, auto scaling empowers organizations to build and maintain responsive, resilient, and cost-effective applications and services in the dynamic cloud environment.



Steps of installation:

