

# *CLOUD COMPUTING*

## Module-2

### Using Cloud Computing Platforms

# *Abstraction and Virtualization*

- *Virtualization Technologies*
- *Load Balancing and Virtualization*
- *Hypervisors*
- *Machine Imaging*
- *Porting Applications*

# VIRTUALIZATION

Virtualization is the process of creating a software-based, or virtual, representation of something, such as virtual applications, servers, storage and networks.

It is the single most effective way to reduce IT expenses while boosting efficiency and agility for all size businesses.

Virtualization can increase IT agility, flexibility and scalability while creating significant cost savings.

Greater workload mobility, increased performance and availability of resources, automated operations – they're all benefits of virtualization that make IT simpler to manage and less costly to own and operate.

# Benefits of virtualization

Additional benefits include:

- Reduced capital and operating costs.
- Minimized or eliminated downtime.
- Increased IT productivity, efficiency, agility and responsiveness.
- Faster provisioning of applications and resources.
- Greater business continuity and disaster recovery.
- Simplified data center management.

# Virtualization technologies

Majority of cloud-based systems combine their resources into pools that can be assigned on-demand to users.

Without resource pooling, it is impossible to attain efficient utilization, provide reasonable costs to users, and proactively react to demand.

Pooled resources can be accessed 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.

The mapping of virtual resources to physical resources can be both dynamic and facile; which helps efficient resource management.

Virtualized components for cloud computing:

*Access:* A client can request access to a cloud service from any location.

*Application:* A cloud has multiple application instances and directs requests to an instance based on conditions.

*CPU:* Computers can be partitioned into a set of virtual machines with each machine being assigned a workload. Alternatively, systems can be virtualized through load-balancing technologies.

*Storage:* Data is stored across storage devices and often replicated for redundancy.

*To enable these characteristics, resources must be highly configurable and flexible.*

Mobility patterns for Cloud Computing:

P2V: Physical to Virtual

V2V: Virtual to Virtual

V2P: Virtual to Physical

P2P: Physical to Physical

D2C: Datacenter to Cloud

C2C: Cloud to Cloud

C2D: Cloud to Datacenter

D2D: Datacenter to Datacenter

Virtualization is a key enabler for the following attributes of cloud computing:

*Service-based:* A service-based architecture is where clients are abstracted from service providers through service interfaces.

*Scalable and elastic:* Services can be altered to affect capacity and performance on demand.

*Shared services:* Resources are pooled in order to create greater efficiencies.

*Metered usage:* Services are billed on a usage basis.



# LOAD BALANCING

Clients in a distributed environment randomly generate a request in any processor in the network.

The major drawback of it is associated with the assignment of tasks.

The unequal assignment of the task to the processor creates imbalance i.e., some of the processors are heavily overloaded and some of them are underloaded.

The main objective of load balancing is to transfer the load from overloaded process to an underloaded process.

Load Balancing is very essential for efficient operations in a distributed environment.

To achieve better performance, minimum and fast response time and high resource utilization we need to transfer the tasks between different nodes in the cloud network. The load balancing technique is used to distribute tasks from overloaded nodes to underloaded or idle nodes.

# LOAD BALANCING IN CLOUD COMPUTING


Cloud Load balancing is basically the process of distributing or dividing the workloads and different computing resources across one or more available servers. This kind of distribution ensures that maximum throughput in a minimum response time.

The workload is divided among two or more servers, hard drives, network interfaces or other different computing resources, which helps to enable better resource utilization and improves system response time. Thus, for a website with high traffic rate, effective use of the cloud load balancing can ensure better business continuity.

Load balancing is an optimization technique; it can be used to increase utilization and throughput, lower latency, reduce response time, and avoid system overload.

The common objectives of using load balancers are:

- To maintain system firmness.
- To improve system performance.
- To protect against system failures.



Cloud providers like **Amazon Web Services (AWS)**, **Microsoft Azure** and **Google** offer cloud load balancing to facilitate easy distribution of workloads.

For exp: **Amazon Web Services (AWS)** offers **Elastic Load balancing (ELB) technology** to distribute traffic among Elastic Compute Cloud (EC2) instances.

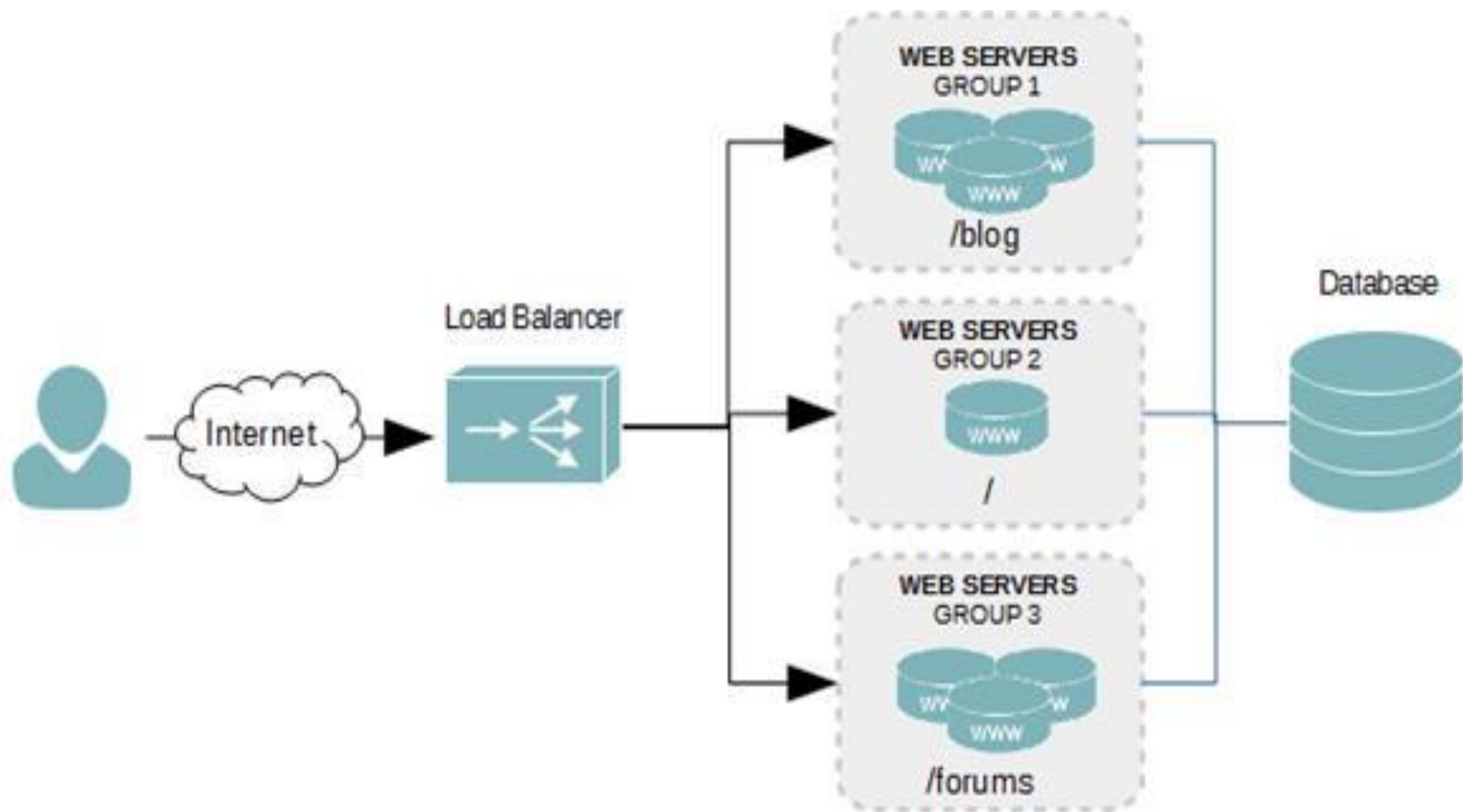
Most of the **Amazon Web Services (AWS)** powered applications have Elastic Load Balancers (ELBs) installed as the key architectural component.

Similarly, **Azure's Traffic Manager** allocates its cloud servers' traffic across multiple data centers.

## LOAD BALANCED RESOURCES

The different network resources that can be load balanced are as follows:

- ❑ Storage resources
- ❑ Connections through intelligent switches
- ❑ Processing through computer system assignment
- ❑ Access to application instances
- ❑ Network interfaces and services such as DNS, FTP, and HTTP
- ❑ In Load balancing Scheduling algorithms are used to assign resources
- ❑ The various scheduling algorithm that are in use are round robin and weighted round robin fastest response time, least connections and weighted least connections, and custom assignments.



## Working of Load Balancing

Actually Load does not refer to only the website traffic but it also includes the CPU load, network load and memory capacity of every server.

A load balancing technique always makes sure that each and every system connected to the network has the same amount of workload at any instant of time. This ensures that neither any of them is excessively over-loaded, nor under-utilized.

The load balancer always distributes data depending upon how busy each server or node is at a particular time.

In the absence of a load balancer, the client must wait while his process gets processed, which might be too tiring and demotivating for him and it is not recommended at any point in time.

Various types of information like jobs waiting in the queue, CPU's processing rate, job arrival rate, etc. are exchanged between the processors during the load balancing process.

Failure in the right implementation of the load balancers can lead to serious problems, and data getting lost is one of them.

Different companies may use different load balancers and multiple load balancing algorithms like static and dynamic load balancing.

One of the most commonly used methods is **Round-robin load balancing**.

It forwards the client requests to each connected server in turn.

On reaching the end, the load balancer loops back and repeats the list again. The major benefit is its ease of implementation.

The load balancers check the system heartbeats during set time intervals to verify whether each node is performing well or not.

## Advantages of Load Balancing

- High Performing applications
- Increased scalability
- Ability to handle sudden traffic spikes
- Business continuity with complete flexibility



# ADVANCED LOAD BALANCING

Workload managers are more sophisticated load balancers.

They determine the current utilization of the resources in their pool, the response time, the work queue length, connection latency and capacity, and other factors in order to assign tasks to each resource.

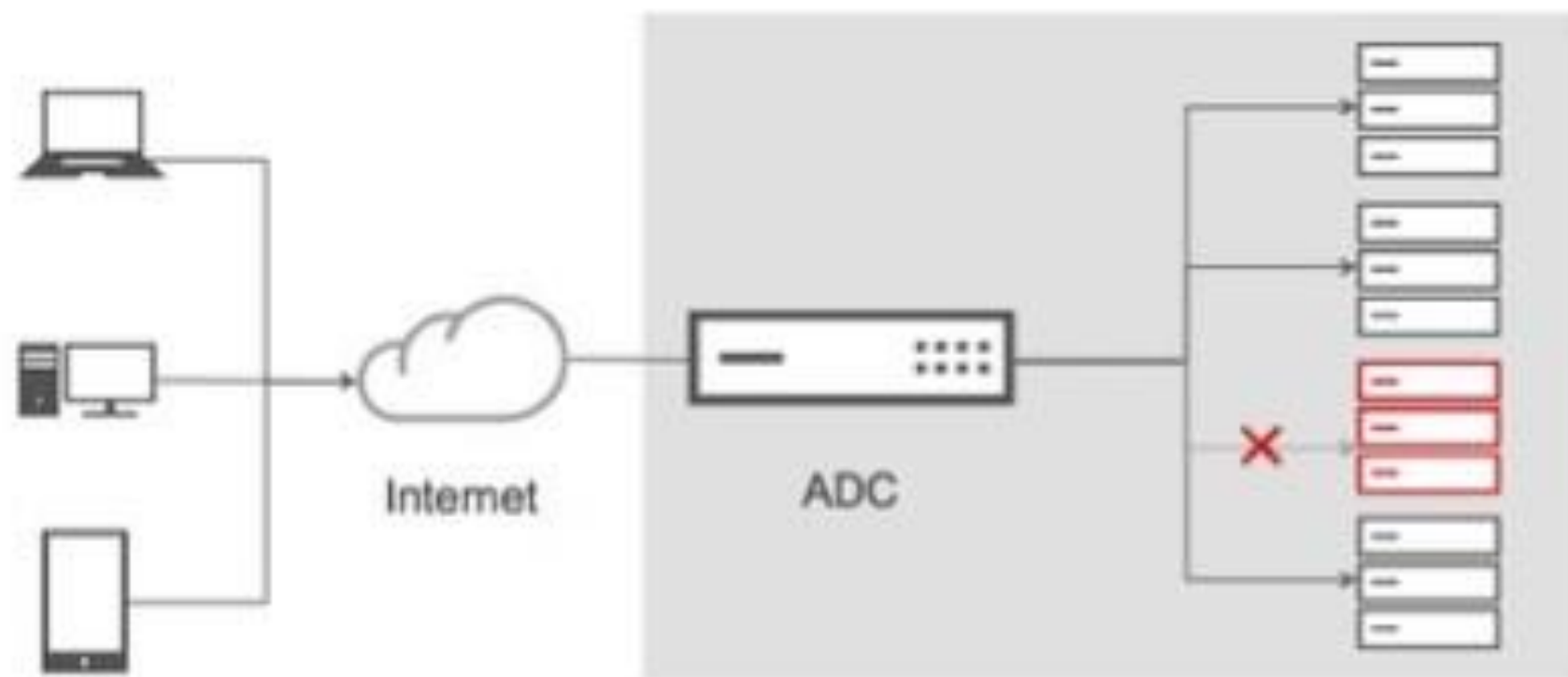
An Application Delivery Controller (ADC) is a combination of load balancer and application server. It is placed between a firewall/router and a data centre providing the web services.

ADCs are also referred to as a content switch, multilayer switch, or web switch.

An ADC is assigned a virtual IP address (VIP) that it maps to a pool of servers based on application specific criteria.

An ADC is considered to be an advanced version of a load balancer and lowers the workload of the Web servers.

Services provided by an ADC include data compression, content caching, security, server health monitoring, Secure Sockets Layer[SSL] offload and advanced routing based on current conditions.



An ADC is considered to be an application accelerator, which focus on two areas of technology:

- Network optimization
- Application/framework optimization.

An architectural layer containing ADCs is described as an Application Delivery Network (ADN) which provide WAN optimization services.

The purpose of an ADN is to distribute content to resources based on application specific criteria.

ADN provide a caching mechanism to reduce traffic, traffic prioritization and optimization, and other techniques.

# HYPERVISORS

Virtual systems are created out of physical systems. Using a computer system with a certain set of resources, you can set aside portions of those resources to create a virtual machine.

A virtual machine has all the attributes and characteristics of a physical system but is strictly software that emulates a physical machine.

A system/hardware virtual machine has its own address space in memory, processor resource allocation, and device I/O using its own virtual device drivers.

Some virtual machines are designed to run only a single application or process and are referred to as process virtual machines.

Virtual machines provide the capability of running multiple machine instances, each with their own operating system.

A **hypervisor**, also known as a virtual machine monitor or VMM, is software that creates and runs virtual machines (VMs).

A hypervisor allows one host computer to support multiple guest VMs by virtually sharing its resources, such as memory and processing.

It is a low-level program that allows multiple operating systems to run concurrently on a single host computer.

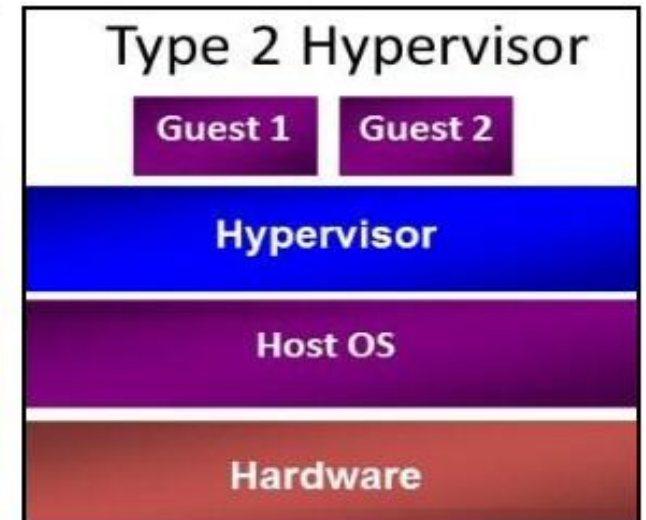
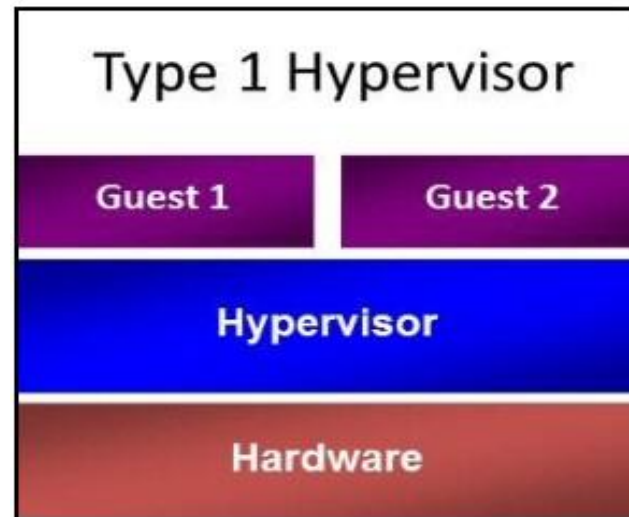
Hypervisors use a thin layer of code in software or firmware to allocate resources in real-time.

A computer on which a hypervisor runs one or more virtual machines is called a host machine, and each virtual machine is called a guest machine.

Hypervisor/Virtual Machine Types:

➤ Type 1

➤ Type 2



## *Type 1 Hypervisor*

Type 1 hypervisors run directly on the system hardware. They are often referred to as a "native" or "bare metal" or "embedded" hypervisors.

Type 1 VMs have no host operating system because they are installed on a bare system.

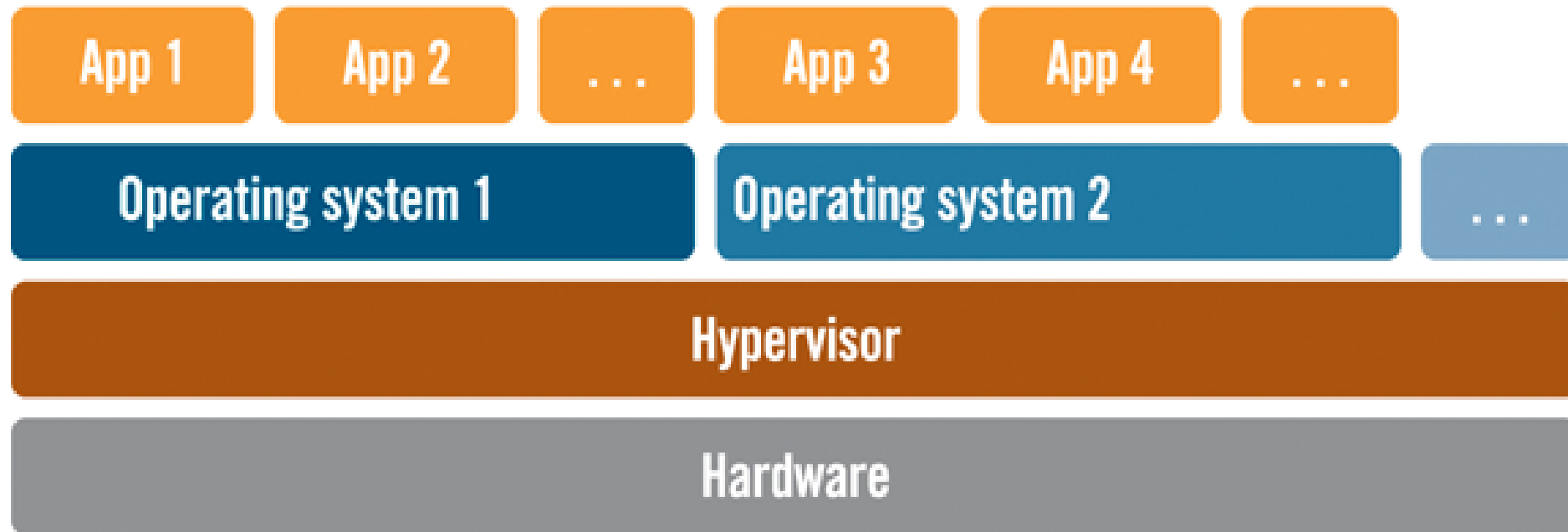
Because they run directly on the hardware, Type 1 hypervisors support hardware virtualization.

They provide higher performance, availability, and security.

Eg: Oracle VM, VMware ESX and ESXi



## Type 1 hypervisor



**Figure 2. A Type 1 or bare-metal hypervisor sits directly on the host hardware.**

## *Type 2 Hypervisor*

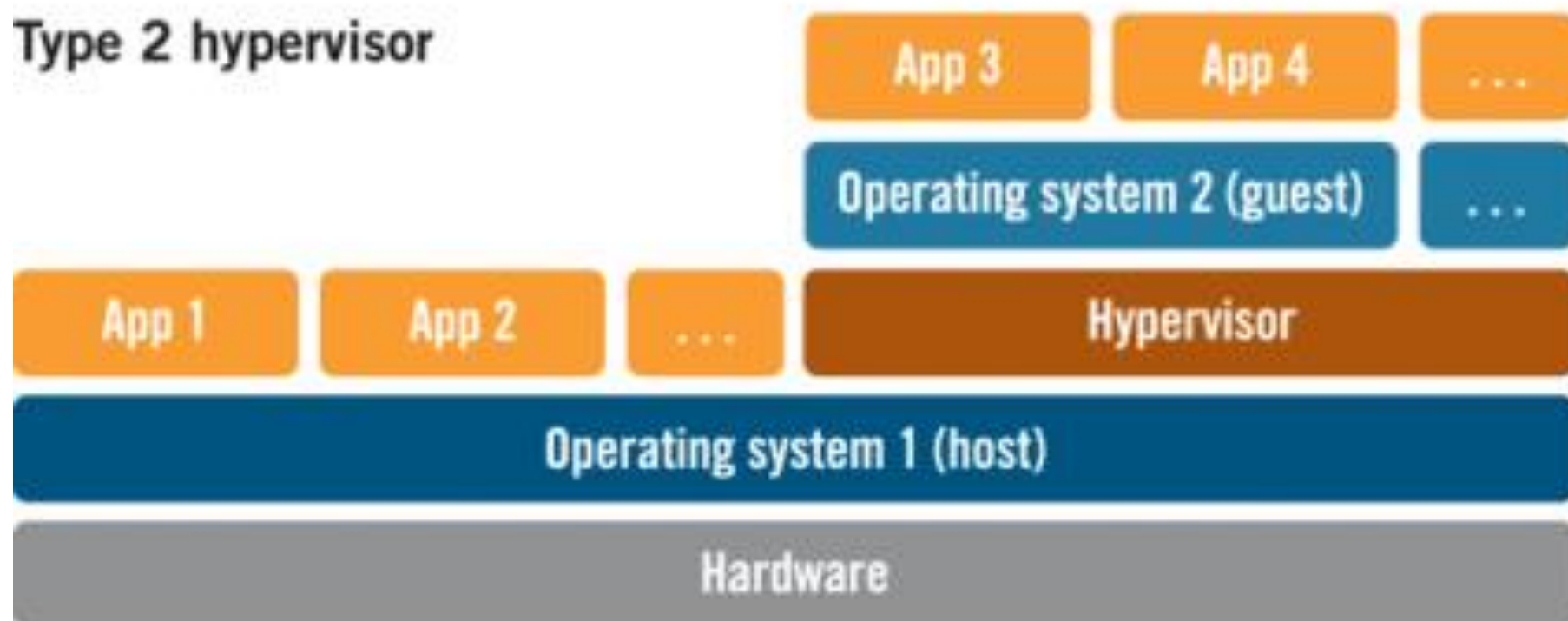
It run on a host operating system.

Because they run as an application on top of an operating system, Type 2 hypervisors perform software virtualization.

Additional applications can be installed, executed and can be monitored; Such as browser, word editor, application softwares etc.

Eg: Microsoft Hyper V, VMware Workstation 6.0

Type 2 hypervisor



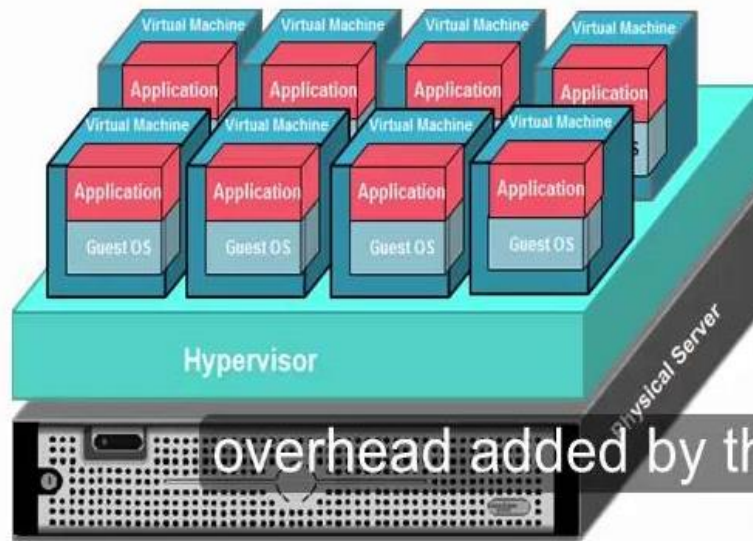
**Figure 1. A Type 2 hypervisor runs as an application on a host operating system.**

## Type-1 Hypervisor

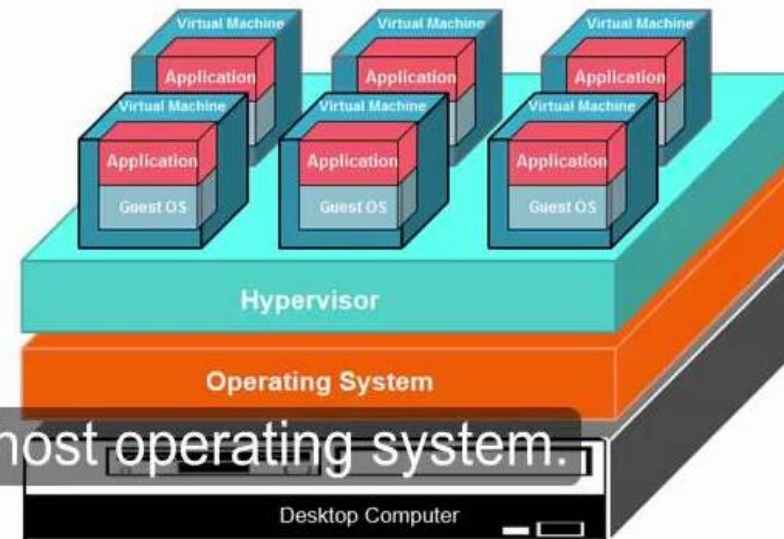
Higher performance and scalability because of being bare-metal type

## Type-2 Hypervisor

Lower performance as a result of host operating system overhead



Type-1 Hypervisor



Type-2 Hypervisor

overhead added by the host operating system.

# BENEFITS OF HYPERVISORS

There are several benefits to using a hypervisor that hosts multiple virtual machines:

- **Speed:** Hypervisors allow virtual machines to be created instantly, unlike bare-metal servers. This makes it easier to provision resources as needed for dynamic workloads.
- **Efficiency:** Hypervisors that run several virtual machines on one physical machine's resources also allow for more efficient utilization of one physical server. It is more cost- and energy-efficient to run several virtual machines on one physical machine than to run multiple underutilized physical machines for the same task.
- **Flexibility:** Bare-metal hypervisors allow operating systems and their associated applications to run on a variety of hardware types because the hypervisor separates the OS from the underlying hardware, so the software no longer relies on specific hardware devices or drivers.

**Portability:** Hypervisors allow multiple operating systems to reside on the same physical server (host machine). Because the virtual machines that the hypervisor runs are independent from the physical machine, they are portable.

IT teams can shift workloads and allocate networking, memory, storage and processing resources across multiple servers as needed, moving from machine to machine or platform to platform.


When an application needs more processing power, the virtualization software allows it to seamlessly access additional machines.

# MACHINE IMAGING

A mechanism to provide system portability, instantiate applications and provision and deploy systems in the cloud through storing the state of a systems using a system image.

A system image makes a copy or a clone of the entire computer system inside a single container such as a file.

The system imaging program is used to make this image and can be used later to restore a system image.



Some imaging programs allow you to view the files contained in the image and do partial restores.

Eg: Amazon Machine Image (AMI) used by Amazon Web Services: An AMI is a file system image that contains an operating system, all appropriate device drivers, and any applications and state information that the working virtual machine would have.



# PORTING APPLICATIONS

Cloud computing applications have the ability to run on virtual systems and these virtual systems to be moved as needed to respond to demand.

Systems (VMs running applications), storage, and network assets can all be virtualized.

Cloud Developers want the ability to port their applications from one cloud vendor to another. But the major cloud vendors don't have technologies that interoperate with one another.

Some technologies used for porting applications:

- *The Simple Cloud API* by Zend Technologies
- *Virtual Application Appliance* by AppZero

## *The Simple Cloud API*

Porting an application build on Microsoft Azure platform to AWS or GoogleApps is difficult.

The *Simple API for Cloud Application Services* is an open source initiative by Zend Technologies to create a common application program interface that will allow applications to be portable.

Among the founding supporters are IBM, Microsoft, Nivanix, Rackspace, and GoGrid.

Simple Cloud API has common interfaces for:

- File Storage Services
- Document Storage Services
- Simple Queue Services

# *AppZero Virtual Application Appliance*

Applications that run in datacenters are captive to the operating systems and hardware platforms that they run on.

Also, the applications are tightly coupled with the operating systems on which they run due to the use of Dynamic Link Libraries (DLL) and registry files.

So moving an application from one platform to another is not simple.

An application developed by AppZero company, called the *Virtual Application Appliance* (VAA), gives the ability to run an application from whatever platform you want.

It is created as an architectural layer between the Windows or the UNIX operating system and the applications.

The virtualization layer serves as the mediator for file I/O, memory I/O, and application calls and response to DLLs.

The running application in AppZero changes none of the registry entries or any of the files on the Windows Server.

*Virtual Application Appliance* creates a container that encapsulates the application and all the application's dependencies within a set of files; it is essentially an Application Image for a specific OS.

Dependencies includes DLL, service settings, necessary configuration files, registry entries, and machine and network settings

This container forms an installable server-side application stack that can be run after installation, but has no impact on the underlying operating system.

VAAs are created using the AppZero Creator wizard, managed with the AppZero Admin tool, and may be installed using the AppZero Director, which creates a VAA runtime application.

AppZero Dissolve removes the VAA virtualization layer from the encapsulated application and installs that application directly into the operating system.



# *Capacity Planning*

- *Capacity Planning*
- *Defining Baseline and Metrics*

*Baseline Measurements | System Metrics | Load Testing*

*Resource Ceiling | Server and Instance Types*

- *Network Capacity*
- *Scaling*


# CAPACITY PLANNING

Capacity planning examines the available systems, measures their performance, and determines patterns in usage that enables the planner to predict demand.

A system uses processor, memory, storage, and network capacity to satisfy cloud computing demands.

Each of these resources has a utilization rate, and these resources reaches a ceiling that limits performance when demand increases.

Resources are provisioned and allocated to meet demand. The goal of capacity planning is to accommodate the workload.



Capacity planning measures the maximum amount of work that can be done using the current technology and then adds resources to do more work as needed.

It is the goal of a capacity planner to identify the critical resource that has resource ceiling and add more resources to move the bottleneck to higher levels of demand.



### *Capacity planning steps:*

1. Determine the characteristics of the present system.
2. Measure the workload for the different resources in the system: CPU, RAM, disk, network, and so forth.
3. Load the system until it is overloaded, determine when it breaks, and specify what is required to maintain acceptable performance.
4. Predict the future based on historical trends and other factors.
5. Deploy or tear down resources to meet your predictions.
6. Iterate Steps 1 through 5 repeatedly.

# DEFINING BASELINE AND METRICS

In business, the current system capacity or workload should be determine as a measurable quantity over time.

Many developers create cloud-based applications and Web sites based on a LAMP solution stack.

*LAMP* stands for:

- *Linux*: the operating system
- *Apache HTTP Server*: the Web server.
- *MySQL*: the database server
- *PHP [Hypertext Preprocessor]*: the scripting language

These four technologies are open source products.

# 1. *BASELINE MEASUREMENTS*

Two important overall workload metrics in this LAMP system:

- *Page views or hits* on the Web site, as measured in hits per second.
- *Transactions* completed on the database server, as measured by transactions per second or perhaps by queries per second.

The total workload might be served by a single server instance in the cloud, a number of virtual server instances, or some combination of physical and virtual servers.

## 2. *SYSTEM METRICS*

Capacity planning must measure system-level statistics, determining what each system is capable of, and how resources of a system affect system-level performance.

A machine instance (physical or virtual) is primarily defined by four essential resources:

- CPU
- Memory (RAM)
- Disk
- Network connectivity

In Linux/UNIX, *sar* command display the level of CPU activity. In Windows, the *Task Manager* serves this purpose.

Linux performance measurement tool RRDTool (Round Robin Database tool) capture time-dependent performance data from resources such as a CPU load, network utilization (bandwidth), and so on and store the data in a circular buffer. It is commonly used in performance analysis work.

Some LAMP Performance Monitoring Tools are:

- Alertra: Web site monitoring service
- Collectd: System statistics collection daemon

### *3. LOAD TESTING*

The aim of Load Testing is to check what happens to a system when the load increases.

It is also referred to as performance testing, reliability testing, stress testing, and volume testing.

Upon reaching the maximum load, cloud can create virtual clone of the system to perform tasks.

Examples of load generation tools: HP LodeRunner, IBM Rational Performance Tester, JMeter

Load balancers serves more requests to more powerful systems and fewer requests to less powerful systems.

Load testing seeks to answer the following questions:

- What is the maximum load that my current system can support?
- Which resource(s) represents the bottleneck in the current system that limits the system's performance?
- Can I alter the configuration of my server in order to increase capacity?
- How does a server's performance relate to other servers that might have different characteristics?

## *4. RESOURCE CEILING*

Among several components (like the CPU, RAM, Network I/O and Disk I/O) of a particular server, if any component reaches its maximum utilization while functioning, this factor is the current system resource ceiling.

Usually, in such a scenario, since a particular component reaches its maximum utilization, more utilization of other resources might not be possible even though the system is not fully loaded.




## *5. SERVER AND INSTANCE TYPES*

Capacity planning makes the growth and shrinkage of capacity predictable.

In cloud computing, you can increase (automatically or manually) capacity on demand quickly and efficiently.

It can be made possible by standardizing on a few hardware types and then well characterizing those platforms.

Assign servers standardized roles and populate those servers with identical services.



A server with the same set of software, system configuration, and hardware should perform similarly if given the same role in an infrastructure.


Capacity planning compare the capability of different systems and choose the solution that is right sized and provides the service with the best operational parameters at the lowest cost.

Server instances of various sizes are created to performs different kinds of tasks.

# NETWORK CAPACITY

There are three aspects to assessing network capacity:

- Network traffic to and from the network interface at the server, be it a physical or virtual interface or server(Server side network)
- Network traffic from the cloud to the network interface (Measurement of WAN traffic)
- Network traffic from the cloud through your ISP to your local network interface (your computer)



To measure network traffic at a server's network interface, a network monitor is used, which is a form of packet analyzer.

Eg: Microsoft includes a utility called the Microsoft Network Monitor as part of its server utilities.

# SCALING

Scalability is a key feature provided in cloud computing.

This is achieved with the help of providing adequate infrastructure.

You can either scale vertically (scale up) or scale horizontally (scale out), and each method is broadly suitable for different types of applications.

## *Vertical scaling*

In this method, we add resources to a system to make it more powerful. Results in single powerful supercomputer.

Eg: Replacing a dual-processor machine instance equivalence with a quad-processor machine instance equivalence

Vertical scaling allows you to use a virtual system to run more virtual machines (operating system instance), run more daemons on the same machine instance, or take advantage of more RAM (memory) and faster compute times.

## *Horizontal scaling*

It adds capacity to a system by adding more individual nodes.

Eg: In a dual-processor machine instance, you add more dual-processor machines instances.

Scaling out indefinitely leads you to an architecture with a large number of servers, which is the model that many cloud and grid computer networks use.

It allows you to run distributed applications more efficiently and is effective in using hardware more efficiently because it is both easier to pool resources and to partition them.

# *Exploring Platform as a Service*

- *SaaS versus PaaS*
- *Application Development*
- *Using PaaS Application Frameworks*  
*Drupal / Squarespace / Eccentex*  
*LongJump/ WaveMaker /Wolf Frameworks*



# EXPLORING PLATFORM AS A SERVICE

The PaaS model provides the tools and environment that are needed to create applications that can run in a SaaS model.

Applications developed in PaaS systems can be composite business applications, data portals, or mashups with data derived from multiple sources.

Application frameworks are powerful tool for creating cloud computing applications.

The services provided by PAAS model are:

- *Application development.*
- *Collaboration:* allows multiple individuals to work on the same projects.
- *Data management:* Tools are provided for accessing and using data.
- *Instrumentation, performance, and testing:* Tools are available for measuring your applications and optimizing their performance.
- *Storage:* Data can be stored in either the PaaS vendor's service or accessed from a third-party storage service.
- *Transaction management.*

# *SAAS VERSUS PAAS* SALESFORCE.COM VERSUS FORCE.COM

Salesforce.com is a Web application suite that is a SaaS.

Force.com is its PaaS platform for building their own services.

The Salesforce.com team created hosted software based on a cloud computing model.

Since Salesforce.com is browser-based, it is platform-independent. Its services are available to mobile devices that runs on different platforms.

# APPLICATION DEVELOPMENT

A PaaS provides the tools needed to construct different types of applications that can work together in the same environment.

These are among the common application types:

- Composite business applications.
- Data portals.
- Mashups of multiple data sources.

Eg: Google AppEngine, Microsoft Windows Azure Platform, Eccentex AppBase, LongJump, and Wolf



All PaaS application development must take into account lifecycle management.

As an application ages, it must be upgraded, migrated, grown, and eventually phased out or ported.

Many PaaS vendors offer systems that are integrated with lifecycle development platforms.

That is, the vendor provides a full software development stack for the programmer to use, and it isn't expected that the developer will need to go outside of the service to create his application.

An integrated lifecycle platform includes the following:

- The virtual machine and operating system (often offered by an IaaS)
- Data design and storage
- A development environment with defined Application Programming Interfaces
- Middleware
- Testing and optimization tools
- Additional tools and services

# USING PAAS APPLICATION FRAMEWORKS

Application frameworks provide a means for creating SaaS hosted applications using a unified development environment or an integrated development environment (IDE).

Many Web sites are based on the notion of information management and organization; they are referred to as content management systems (CMS).

A database is a content management system [CMS], but implementing it as a Web site adds a number of special features such as rich user interaction, multiple data sources, and extensive customization and extensibility.

These services they have these common characteristics:

They separate data-handling from presentation (user interface).

They offer tools for establishing business objects or entities and the relationships between them.

They support the incorporation of business rules, logic, and actions.

They provide tools for creating data entry controls (forms), views, and reports.

They provide instrumentation, tools for measuring application performance.

They support packaging and deployment of applications.



- 
- Drupal
  - Squarespace
  - Eccentex
  - LongJump
  - WaveMaker
  - Wolf Frameworks