1. INTRODUCTION

1.1 CLOUD COMPUTING

Jadeja. Y et al., [Jadeja. Y et al., 2012] say the latest emerging trends in information technology is cloud computing. The IT organizations transfer their information technology to cloud into something with greater business usage. The IT challenges listed below have made organizations think about the cloud computing model to provide better service to their customers:

- **Globalization**: IT must meet the business needs to serve customers worldwide, round the clock – 24x7x365.
- **Data center technology**: Converting the old storage method into a new data center storage method.
- **Reduce Cost**: Today the increase in commercial enterprise takes the purchase of own hardware and software products which is an expensive affair.
- **Storage Growth**: Explosion of storage consumption and usage.

Due to these causes the data centre computing has been inserted. In recent years, most of the IT companies use data center technology for storing information.

1.2 DATA CENTER

Katarina Stanoevska-Slabeva et al., [Katarina Stanoevska-Slabeva et al., 2010] say that data centers provide centralized digital data-processing capabilities needed to sustain an organization’s line of work. realized digital data-processing capabilities required to support an organization’s business.

Data center is a collection of servers with different applications. A data center can be Stationed anywhere in the universe and it should away from the customer, the
customer does not know which data centre the data is stacked away. The data center is classified into classis data center and virtualized data center.

FIGURE 1.1: JOURNEY TO CLOUD COMPUTING

Figure 1.1 illustrates the cloud journey. Organizations have to perform various steps to promote their existing data centers, to provide cloud services.

1.2.1 Classic Data Center (CDC)

A CDC is a facility containing physical IT resources, including compute, network, and storage. The core components of a CDC include compute (server), storage, network, application, and DBMS.

Compute: Compute is a physical computing machine that runs operating systems, applications and databases.

Storage: Storage refers to a device that stores data persistently for subsequent use.

Network: Network is a data path that helps communication between nodes and compute systems or between compute systems and memory.

Application: An application is a computer program that provides the logic for computing operations. Applications may use a DBMS, which uses operating system services to perform store or retrieve operations on storage devices.
DBMS: DBMS provides a structured way to store data in logically organized tables that are related. A DBMS optimizes the storage and retrieval of data.

These core components are typically seen and managed as separate entities. But, all these elements must work together to address data processing requirements. Other factors of a CDC are power supplies and environmental controls, such as air conditioning and fire suppression.

![Image of the structure of classic data center](image)

**FIGURE 1.2: THE STRUCTURE OF CLASSIC DATA CENTER**

Figure 1.2 shows the structure of the classic data center. In the classic data center the computing resources are typically dedicated to each application. This runs to complex management and underutilization of resources. The limitations of CDC can be overcome by the Virtualized Data Centers (VDCs). The classic data center uses normal storage disk drives.

**1.2.2 Virtualized Data Center (VDC)**

Gaurav Dhiman et al.,[Gaurav Dhiman et al., 2013] say a virtualization is a new idea used nowadays in a cloud computing. A VDC is a data center in which compute, storage, network and/or applications are virtualized.
Rimal B.P et al.,[Rimal B.P et al., 2009] says the core elements of Virtualized data center are

- Compute virtualization enables running multiple OS concurrently on a complete system.
- Storage virtualization provides a logical view of storage and presents it to the combat system.
- Network virtualization provides multiple logical networks are created on a physical network

**FIGURE 1.3: THE EVOLUTION OF CLOUD DATA CENTER**

Figure 1.3 shows the evolution of virtualization, which is the starting step towards building a cloud theme. Transforming a Classic Data Center (CDC) into a Virtualized Data Center (VDC) requires virtual core elements of the data center. A virtualized infrastructure enables a smooth transition from the classic data center to virtualized data center.
1.2.3 Key requirements of a Data Center

Nuaimi K.A et al., [Nuaimi K.A et al., 2012] say that data center infrastructure should meet the following attributes to ensure that data is accessible to users effectively and efficiently all the time.

[Diagram: Availability, Performance, Manageability, Security, Flexibility, Scalability]

**FIGURE 1.4: KEY REQUIREMENTS OF A DATA CENTER**

**Availability:** All data centre elements should be planned to assure availability. The inability of users to access data can have a significant negative impact on a business.

**Performance:** All the elements of the data center should provide high performance and service to all processing requests of the client with high speed.

**Scalability:** Data center should allocate more processing of storage space without interrupting business operations. Data center should be able to provide additional resources on demand without interrupting availability of the current process.

**Data Integrity:** Data integrity ensures that data is not changed or modified and the data are stored in original copy. Any variation in data during its retrieval implies corruption, which may affect the operations of the client process in an organization.
**Manageability:** Manageability can be achieved through automation and reduction of manual intervention in common tasks. The additional memory resource requirements should primarily be managed by reallocating or improving utilization of existing resources, rather than by adding new resources. The cost of data center management is one of the key concerns. Organizations are looking towards optimizing their IT spending on data center maintenance, so that they can invest in innovation and new application deployment.

### 1.3 COMPUTE VIRTUALIZATION

According to Xie, Silian et al.,[Xie, Silian, et al., 2012] the compute virtualization is a technique of converting or allocation of physical resources from the operating system and enabling multiple operating systems to run on a single or multiple physical machine(s) called virtual machine. The virtual machine (VM) is a portable machine from physical machine.

![FIGURE 1.5: COMPUTE VIRTUALIZATION](image)

Figure 1.5 illustrated the process of compute virtualization. The virtualization layer is also known as a hypervisor. The hypervisor provides standardized hardware resources (for Example: CPU, Memory, Network, etc.) to all the virtual machines.

Hu Wu et al.,[[Hu Wu et al., 2012] say that virtual machine is a virtual entity like a physical machine. Each operating system runs on its own virtual machines. In compute virtualization, a virtualization layer lies between the hardware and virtual machine.
Physical to Virtual machine Conversion (P2V)

Physical to virtual machine (VM) conversion is a process through which a physical machine is converted into a virtual machine. When converting a physical machine, the “Converter Application” (Converter) clone data on the hard disk of the source machine and transfers that data to the destination virtual disk. Cloning is the process of creating a copy of the virtual disk that is an exact copy of the source physical disk. After cloning is complete, reconfiguration steps are performed on the destination virtual machine.

The system reconfiguration is the process of configuring the migrated operating system to enable it to function on a virtual hardware. This configuration is performed on the target virtual disk after cloning, and enables the target virtual disk to function as a bootable system disk in a virtual machine. Because the migration process is non-destructive to the source, it can continue to be in use after the conversion is complete.

From a user’s perspective, a virtual machine (VM) is a logical compute system just like a physical machine that runs an operating system (OS) and the application. An operating system that runs within a virtual machine is called a guest operating system. At a time, only one supported guest operating system can run on a single virtual machine. Each virtual machine is independent and can run its own application. From a hypervisor perspective, a virtual machine is a discrete set of files. The set includes a configuration file, virtual disk files, virtual BIOS file, virtual machine swap file, and a log file.

FIGURE 1.6: PHYSICAL TO VIRTUAL MACHINE CONVERSION (P2V)
Figure 1.6 shows the concept of physical to virtual machine conversion on a cloud. After converting P2V, the virtual machine can do its process individually.

**FIGURE 1.7: EXAMPLE P2V CONVERSION**

Figure 1.7 illustrates an example where the resources of a parent resource pool are distributed among child resource pools and virtual machines.

**Recovery-Point Objective (RPO)**

RPO is the point in time to which systems and data must be recovered after an outage. It defines the amount of data loss that a business can endure. A large RPO signifies high tolerance to information loss in a business. Based on the RPO, organizations plan for the minimum frequency with which a backup or replica must be made.

For example, if the RPO is six hours, backups or replicas must be made at least once in 6 hours. An organization may plan for an appropriate BC technology solution on the basis of the RPO it sets for example, if RPO is 24 hours, that means that backups are created on an off-site tape drive every midnight. The corresponding recovery strategy is to restore data from the set of last backup tapes. Similarly, for zero RPO, data is mirrored synchronously to a remote site.
Recovery-Time Objective (RTO)

The time within which systems, applications, or functions must be recovered after an outage. It defines the amount of downtime that a business can endure and survive. Businesses can optimize disaster recovery plans after defining the RTO for a given data center or network.

For example, if the RTO is two hours, then use a disk backup because it enables a faster restore than a tape backup. However, for an RTO of one week, tape backup will most likely meet the requirements. Few examples of RTOs and the recovery strategies to ensure data availability is listed below:

An RTO of 72 hours: Restore from the backup tapes at a cold site.

An RTO of 12 hours: Restore from tapes at a hot site.

An RTO of 4 hours: Use a data vault to a hot site.

RTO of 1 hour: Cluster production servers with controller-based disk mirroring.

A virtual machine (VM) can be configured with the following virtual components:

Virtual Central processing Unit (vCPU): A virtual machine can be configured with one or more vCPU when it is created. The number of vCPUs can be increased or decreased based on requirements.

- Virtual Random Access Memory (vRAM): vRAM is the amount of memory allocated to a virtual machine. It is visible to a guest operating system. This memory size can be changed based on requirements.

- Virtual Disk: A virtual disk stores the virtual machine's operating system, program files, application data, and other data associated with the virtual machine. A virtual machine should have at least one virtual disk.

- Virtual Network Adaptor (vNIC): It provides connectivity between virtual machines on the same compute system, between virtual machines on
different compute systems between virtual and physical machines. vNIC functions exactly like a physical NIC.

- **Virtual DVD/CD-ROM** and floppy drives: These devices enable to map the virtual machines drive to either the physical drive or to the image file (such as CD/DVD and floppy) for the storage.

- **Virtual SCSI Controller**: A virtual machine uses a virtual SCSI controller to access virtual disks.

- **Virtual USB Controllers**: Enable a virtual machine to connect to the physical USB controller and to access the USB device connected.

**FIGURE 1.8: VM HARDWARE COMPONENTS**

**Optimizing Virtual machines CPU Resources**

Modern CPUs are equipped with multiple cores per CPU and hyper-threading features. A multi-core CPU is an integrated circuit to which two or more processing units (cores) have been attached for enhanced performance and more efficient, simultaneous processing of multiple processes. Hyper-threading makes a physical CPU appear as two or more logical CPUs. Today’s data centers deploy servers with
multi core and hyper-threading features in their environment. The role of the hypervisor scheduler is to assign a physical CPU resource to the virtual CPU in a way that meets system objectives, such as responsiveness, throughput, and utilization. A conventional Operating System schedules a process or thread on a CPU, while a hypervisor schedules virtual CPUs of virtual machines on the physical machines.

A hypervisor supports and optimizes the CPU resources using modern CPU features such as multi-core and hyper-threading. They also support CPU load balancing. The resource pool is used for load balancing.

A resource pool is a logical abstraction of aggregated physical resources that are managed centrally. Each physical machine and cluster has a parent resource pool that groups the resources of that physical machine or cluster. Administrators may create child resource pools from the parent resource pool. Each child resource pool owns some of the parent’s resources. A parent resource pool can contain child resource pools, virtual machines, or both. For each resource pool and virtual machine, reservation, limit, and share (discussed later) can be specified. Reservation, limit, and share are used to control the resources consumed by a child resource pool or virtual machines.

**FIGURE 1.9:** (A) BEFORE VIRTUALIZATION (B) AFTER VIRTUALIZATION
TABLE 1.1: NEED FOR VIRTUALIZATION

<table>
<thead>
<tr>
<th>Before Virtualization</th>
<th>After Virtualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Runs single operating system per machine at a time.</td>
<td>1. Runs multiple operating systems per machine concurrently.</td>
</tr>
<tr>
<td>2. Couples software and hardware tightly.</td>
<td>2. Makes operating system and hardware independent.</td>
</tr>
<tr>
<td>3. Conflicts when multiple applications are running on the same machine.</td>
<td>3. Isolates the virtual machine from each other, and hence there is no conflict.</td>
</tr>
<tr>
<td>4. Underutilizes resources.</td>
<td>4. Improves resource utilization.</td>
</tr>
<tr>
<td>5. Is inflexible and expensive.</td>
<td>5. Offers flexible infrastructure at low cost.</td>
</tr>
</tbody>
</table>

Table 1.1 reveals the difference between cloud virtualization techniques and the distributed techniques. The one operating system (OS) per compute system (physical machine) is deployed because the operating system and the hardware are tightly coupled and cannot be separated.

Traditionally, one OS per physical machine is deployed because the operating system and the hardware are tightly coupled and cannot be separated. Only one application is deployed per system to minimize the potential resource conflict. This causes organizations to purchase new physical machines for every application they deploy, resulting in expensive and inflexible infrastructure. Further, these compute systems remain underutilized - it is very common to find compute systems running at 15 – 20% utilization. This compounding over many machines within a Classic Data Center (CDC) leads to poor utilization of physical machines. Compute virtualization enables to overcome these challenges by allowing multiple operating systems and
applications to run on a single physical machine. This technique significantly reduces acquisition cost and improves utilization.

1.3.1 Benefits of compute virtualization

Compute virtualization offers the following benefits:

- **Server Consolidation**: Compute virtualization enables running multiple virtual machines on a physical server. This reduces the requirement for physical servers.

- **Isolation**: While virtual machines can share the physical resources of a physical machine, they remain completely isolated from each other as if they were separate physical machines. When there are two virtual machines from single physical machine, if one virtual machine fails, the other virtual machine is used for further processing.

- **Encapsulation**: A virtual machine contains a complete configuration of virtual hardware resources, an operating system, and applications. Encapsulation makes virtual machines, portable and easy to manage. For example, a virtual machine can be moved and copied from one location to another just like a file.

- **Hardware Independence**: At virtual machine is configured with virtual components of CPU, memory, network card, storage, SAN disk, ISCSI, NIC and SCSI controller that are completely independent of the underlying physical hardware. This gives the freedom to move a virtual machine from one machine to another virtual machine without making any change to the hardware and software.

- **Reduced Cost**: Compute virtualization reduces the following direct costs

- **Space** (leased or owned) for physical machines, power and cooling, Hardware (including switches and Fiber Channel HBA), and annual maintenance.
The National Institute of Standards and Technology (NIST) defines “cloud computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”.

A disk drive is a core element of storage that governs the performance of any storage system. The older disk array technologies could not overcome performance constraints due to the limitations of a disk drive and its mechanical components. An intelligent storage system consists of four key components:

1. Front-end
2. Cache
3. Back-end
4. Physical Disks

**FIGURE 1.10: THE COMPONENTS OF AN INTELLIGENT STORAGE SYSTEM**

Figure 1.10 illustrates these components and their interconnections. An I/O request received from the compute system at the front-end port is processed through a cache and the back end to enable the storage and retrieval of data from the physical disk. In cloud, RAID technology is the most recent and an important contribution to enhance storage performance and reliability, but disk drives, even with a RAID implementation could not meet performance requirements of today’s applications.
With advancements in technology a new storage solution known as an Intelligent Storage System, has evolved.

These intelligent storage systems are rich RAID arrays that provide highly optimized I/O processing capabilities. These storage systems are configured with a large amount of memory and multiple I/O paths and use sophisticated algorithms to meet the requirements of performance-sensitive applications. These arrays have an operating environment that intelligently and optimally handles the resource utilization.

Ying Shao, et al.,[Ying Shao, et al., 2011] say read request can be serviced directly from cache if the requested data is found in the cache. Cloud computing, RAID technology helps to store the data in a data centre in the form of the Virtualized Data Centre. Normally Cloud computing

• depends on grids.

• is globally distributed and self managed.

• has mostly, no central infrastructure.

• has full metering and billing.

• has service assurance and service level agreements.

• does monitoring and troubleshooting.

• has dynamic on-demand provisioning.

1.4 SERVICES IN CLOUD COMPUTING

According to Moreno-Vozmediano et al., [Moreno-Vozmediano, et al., 2013] say that the cloud service models can be classified into three categories:

• Infrastructure-as-a-Service (IaaS)

• Platform-as-a-Service (PaaS)

• Software-as-a-Service (SaaS)
Infrastructure as a Service - IaaS

Infrastructure as a Service provides the computing infrastructure such as physical or virtual machines and other resources like virtual-machine disk, block and file-based storage, firewalls, virtual local area networks etc. The service providers are Amazon EC2, Windows Azure, Rackspace.

Platform as a Service – PaaS

Platform as a Service provides computing platforms which includes OS, programming language execution environment, database, web server, etc. The service providers are AWS Elastic Beanstalk, Heroku, Force.com, Google App Engine.
Software as a Service

Software as a Service provides access to application software often referred to as high cost software. The user does not have to worry about the installation, set up and running of the application. A service provider will do all these for the user and just the user has to pay and use it through some client. The service providers are Google Apps, Microsoft Office 365.

FIGURE 1.13: THE SOFTWARE AS A SERVICE

1.5 DEPLOYMENT OF CLOUD SERVICES

Lee. R et al.,[Lee. R, et al., 2011] say that cloud Computing can be classified into three deployment models: private, public, and hybrid. These models provide a basis for how the cloud infrastructures are constructed and consumed.

PUBLIC CLOUD

In a public cloud, IT resources are made available to the general public or organizations and are owned by the cloud service provider. The cloud services are accessible to everyone via standard Internet connections.

In a public cloud, a service provider makes IT resources, such as applications, storage capacity, or server compute cycles, available to any consumer. This model can be thought of as an “on-demand” and as a “pay-as-you-go” environment, where there are no on-site infrastructure or management requirements.
However, for organizations, these benefits come with certain risks: no control over the resources in the cloud, the security of confidential data, network performance issues, and interoperability. Popular examples of public clouds include Amazon’s Elastic Compute Cloud (EC2), Google Apps, and Salesforce.com.

**FIGURE 1.14: PUBLIC CLOUD**

PRIVATE CLOUD

Qiang Guan, et al.,[Qiang Guan, et al., 2012] say that the private cloud infrastructure is operated solely for one organization and is not shared with other organizations. This cloud model offers the greatest level of security and control. There are two variations to a private cloud.

**On-premise Private Cloud**: On-premise private clouds is also known as internal clouds, are hosted by an organization within their own data centers. This cloud model provides a more standardized process and protection, but is limited in terms of size and scalability. Organizations would also need to incur the capital and operational costs for the physical resources. This model is best suited for applications which requires complete control and configurability of the infrastructure and security.

**Externally-hosted Private Cloud**: The private cloud is hosted externally with a Cloud provider, where the provider facilitates an exclusive cloud environment for a specific organization with full guarantee of privacy or confidentiality. This model is best suited for organizations that do not prefer a public cloud due to data privacy/security concerns. Like a public cloud computing, a private cloud computing
also enables provisioning an automated service request rather than a manual task processed by IT.

In on-premise private cloud, organizations will have to run their own hardware, storage, networking, hypervisor, and cloud software. Many enterprises, including EMC, Cisco, IBM, Microsoft, Oracle, and VMware, now offer cloud platforms and services to build and manage a private cloud.

FIGURE 1.15: PRIVATE CLOUD

HYBRID CLOUD

In the hybrid cloud environment, the organization consumes resources from both private and public clouds. The ability to augment a private cloud with the resources of a public cloud can be utilized to maintain service levels in the face of rapid workload fluctuations. Organizations use their computing resources in a private cloud for normal usage, but access the public cloud for high/peak load requirements. This ensures that a sudden increase in computing requirement is handled gracefully.

FIGURE 1.16: HYBRID CLOUD
For example, an organization might use a public cloud service, such as Amazon Simple Storage Service (Amazon S3), for archiving data, but continues to maintain in-house storage for operational customer data. Ideally, the hybrid approach allows a business to take advantage of the scalability and cost-effectiveness that a public cloud offers without exposing mission-critical applications and data to third-party vulnerabilities.

**COMMUNITY CLOUD**

The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations).

An example where a community cloud could be useful is in a state government setting. If various agencies within the state government operate under similar guidelines, they could all share the same infrastructure and bear the cost among them.

![Diagram of Community Cloud](image)

**FIGURE 1.17: COMMUNITY CLOUD**

A community cloud may be managed by the organizations or by a third party. With the costs spread over to fewer users than a public cloud, this option is more expensive, but may offer a higher level of privacy, security, and/or policy compliance. The community cloud computing offers organizations access to a vast pool of resources than that in the private cloud.
FIGURE 1.18: CLOUD OFFERING EXAMPLES

Figure 1.18 illustrates the cloud computing service offering organizations. Organizations might need to rapidly expand their businesses, which may enforce them to enlarge the IT infrastructure by adding new servers, storage devices, network data, bandwidth, etc. Critical business data must be protected and should be available to the intended user, which, in turn, requires data security and disaster recovery infrastructure.

The capital expenditure rises to fulfill the requirements, the risk associated with the investment too increases. For small and medium size businesses, this may be a big challenge, which eventually restricts their business to grow. As an individual, it may not be sensible or affordable every time to purchase new applications if they are required only for a brief period. Instead of purchasing new resources, Cloud resources are hired based on pay-per-use without involving any capital expenditures. Cloud
service providers offer on-demand network access to the computing resources, such as servers, storage, network, and cloud applications.

Consumers (organization or individual) can scale up or scale down the demand of computing resources with minimal management effort or service provider interaction. Consumers can leverage Cloud service provider’s expertise to store, protect, backup, and replicate data empowered by the most advanced technology, which otherwise would cost more.

1.6 CHARACTERISTICS OF CLOUD COMPUTING

Individual user on request

A user can use the resources from any place and at any time via network.

Wide range of network access capacities

The user can access the cloud services through a network and can be accessed from different devices such as computers, mobiles and notebooks.

Allocation of resources

In cloud computing, the computer resources are grouped in order to serve a large number of simultaneous users. The mechanism of processing the amount of resources is such that the system dynamically allocates these parameters according to customer requirements. All the data can be stored in the cloud data center.

Measurable service payment, pay-per-use

Cloud systems automatically control and provide necessary resources depending on the needs and request of users to required types of services, power of processor, amount of RAM. All these services are identified and their usage is known for both providers and clients because the financing plays an important role. The service is provided and the user pays as per what the user uses based on hour, month and year.
1.7 ADVANTAGES OF CLOUD COMPUTING

- Lower computer costs
- Reduced software costs
- Instant software updates
- Improved document format compatibility
- Unlimited storage capacity
- Increased data reliability
- Universal document access

1.8 LOAD BALANCING IN CLOUD COMPUTING

Tin-Yu et al., [Tin-Yu, et al., 2012] says the load balancing is a key objective of managing network traffic. It is a technique to distribute workload across multiple physical or virtual machines and parallel network links, to prevent overutilization or underutilization of the resources and to optimize performance. It is provided by a dedicated software or hardware. In VDC, network administrators can apply a policy for distribution of network traffic across VMs and network links. The network traffic management techniques can also be used to set a policy to take over network traffic across network links. In the event of a network failure, the traffic from the failed link will go over to another available link based on a predefined policy. Network administrators have the flexibility to change a policy, when required.

Mohan, N.R.R et al., [Mohan, N.R.R, et al., 2012] say when multiple VM traffic share bandwidth, network traffic management techniques ensure guaranteed service levels of traffic generated by each VM. Traffic management techniques allow an administrator to set priority for allocating bandwidth for different types of network traffic, such as VM, VM migration, IP storage, and management. Load balancing is a process of reassigning the total load to the individual data center of the collective VM
to make resource utilization effective and to improve the response time of the job. The cloud load balancing algorithm can check the conditions in which some of the nodes are over loaded while some others are under loaded.

Tiwari et al., [Tiwari, et al., 2010] say the load balancing algorithm cannot consider the previous state of the system. It considers only the present state. The load balancing can balance certain types of information such as the number of jobs waiting in ready queue, job arrival rate, CPU processing rate, and so forth at each processor for improving the overall performance. This process is called scheduling.

The load balancing can do efficiently by using scheduling methods in cloud. The scheduling is the method of data given access to system resources such as processor time and bandwidth. This load balance is done effectively to achieve a target quality of service. The scheduling in most modern systems perform multitasking.

Olivier Beaumont et al., [Olivier Beaumont et al., 2013] say that the load balancing algorithm is divided into two types

- **Static**: It does not depend on the current state of the system. The user can know whether the system is needed in advance.

- **Dynamic**: Decisions on load balancing are based on the current state of the system. No prior knowledge is needed. So it is better than the static approach.

The load balancing policies are divided into three types

**Transfer Policy**: The dynamic load balancing algorithm which selects a job for transferring from a local node to a remote node.

**Location Policy**: The dynamic load balancing algorithm which selects a destination node for a transferred task.

**Information Policy**: The dynamic load balancing algorithm responsible for collecting information about the nodes in the system.
1.8.1 Load balancing using Multi-core Processors

Xiaonian Wu et al., [Xiaonian Wu et al., 2013] say that the multi-core CPUs provide many advantages to a hypervisor that performs multitasking of virtual machines. A dual-core CPU, for example, can provide almost double the performance of a single-core CPU by allowing two virtual CPUs to execute at the same time.

Bojanova et al., [Bojanova, et al., 2011] say that the load balancing is achieved by migrating a thread from one logical CPU (over utilized) to another (underutilized) to keep the load balanced. The hypervisor intelligently manages the CPU load by spreading it smoothly across the CPU cores in the compute system.

At a regular intervals, the hypervisor looks to migrate the CPU of a virtual machine (virtual CPU) from one logical CPU to another to keep the load balanced. If the logical CPU has no work assigned, it is put into a halted state. This action frees its physical resources and allows the virtual machine running on the logical CPU of the same core to use all the physical resources of that core.
VDC network traffic must be managed in order to optimize both performance and availability of networked resources. Although the network traffic management techniques described in this lesson are related to VDC, some of the techniques are similar to those used in Classic Data Center (CDC).

Load balancing is a key objective of managing network traffic. It is a technique to distribute workload across multiple physical or virtual machines and parallel network links, to prevent overutilization or underutilization of these resources and to optimize performance. It is provided by a dedicated software or hardware.

In VDC network administrators can apply a policy for distribution of network traffic across VMs and network links. Network traffic management techniques can also be used to set a policy to failover network traffic across network links. In the event of a network failure, the traffic from the failed link will failover to another available link based on a predefined policy. Network administrators have the flexibility to change a policy, when required. When multiple VMs in cloud traffic share bandwidth, network traffic management techniques ensure guaranteed service levels of traffic generated by each VM. The traffic management techniques allow an administrator to set priority for allocating bandwidth for different types of network traffic, such as VM, VM migration, IP storage, and management.

1.8.2 Benefits of Load balancing

- Distributing a workload across multiple computers for improved performance
- The failure is avoided.
Improving the VM to process efficiently.

Increasing consumption of system resources

1.8.3 Challenges for Load balancing

Pawar.C.S, et al., [Pawar, C.S, et al.,2012 discuss the principle challenges of load balancing cloud environment is given below

- **Throughput**: It is the total number of tasks that have completed execution for a given scale of time. It is required to have high throughput for better performance of the system.

- **Fault tolerance**: It can be defined as the ability to perform load balancing by the appropriate algorithm without arbitrary link or node failure. Every load balancing algorithm should have a good fault tolerance approach.

- **Migration time**: It is the amount of time for a process to be transferred from one system node to another node for execution. For better performance of the system this time should be always less.

- **Response time**: The time taken by a particular request to respond from the client and the server load balancing technique to respond. The minimum response time gives better performance.

- **Resource Utilization**: This parameter gives information about resources utilized such as memory usage. For efficient load balancing in system in cloud optimum resources should be utilized.

- **Scalability**: It is the ability of the load balancing algorithm for a system with any finite number of processors and machines.

- **Performance**: It is the overall efficiency of the system. If all the parameters are improved, then the overall system performance can be improved.
- **Reliability**: The reliability identified the efficient and failure VM at the time of loading the data.

1.9 RELIABILITY

Chen et al., [Chen, et al., 2011] say that cloud computing is a collection of interconnected and virtualized computing resources that are managed by one or more unified computing resources. Failures of the virtual machines and disaster cause data loss. As a result, the cloud environment requires some capability for an adaptive data replication management.

Reliability is an important component of trust. It is also called the success rate. The reliability is a component to perform the stated functions under stated conditions for a specified period of time. The reliability of a cloud resource is a measure of successful completion of accepted jobs by the cloud resource. The key issue is to enhance reliability by minimizing failures. Reliability is also defined as the probability that the data is available in the system for a certain period. The data replication in cloud supports data reliability. Replication strategy is adopted in cloud computing environments to enhance reliability.

Thanadech Thanakornworakij et al., [Thanadech Thanakornworakij et al., 2013] say that the client’s data is replicated in multiple virtual machines to overcome the data loss. When a client accesses the data the cloud server can search all the VM where the data is stored, including all the replication VMs and identify the reliable VM to process the client data. For example, if the client data is replicated in 10 VMs, the server can send request to all 10 VMs from this ten VM the reliable VM is identified based on reliability methods.

JiSu Park et al., [JiSu Park et al., 2012] say that the reliability method also used to identify the failure VM. To identify the efficient VM, this method overcomes the drawbacks of the existing method. The reliable VM is identified at the time of two situations.
**Replication of process:** In a VDC environment, to restart operations, a user needs to primarily restart VMs on the secondary (backup) site. VM replication makes this possible by making the copies of these VMs available at the backup site. The VM replication is performed at the hypervisor level and relies on replication software that can copy all the changes made to a VM disk to another server.

VM replication requires a secondary site which is a network connectivity linking these sites. VM replication is a critical requirement for successful BC and DR processes in a VDC. This is because, in a VDC environment, to restart operations, a user needs to primarily restart VMs on the secondary (backup) site. VM replication makes this possible by making the copies of these VMs available at the backup site. VM replication is performed at the hypervisor level and relies on replication software that can copy all the changes made to a VM disk to another server. VM replication requires a secondary site which is up and a network connectivity linking these sites.

A virtual disk snapshot taken at the hypervisor level temporarily redirects incoming writes to a separate data file. The delta file maintains these I/Os to be written to the virtual disk after the snapshot is taken. The virtual disk is then mounted by the replication software and any updates since the last replication cycle are copied to another identical virtual disk on a VM at the secondary site.

To ensure data integrity of VM is necessary before the replication process starts. Pauses currently running applications within a VM and forcibly flushes all data in the memory to the disk. To ensure application-level consistency, applications (for example, Microsoft Exchange or SQL Server) are instructed to complete any pending transactions and write the pending data to the disk. There are two main methods to replicate VMs.

The first approach uses compute based replication. In this method, replication happens between similar/dissimilar storage devices. For example, using compute based replication, a VM is replicated from a local disk drive to a storage array in a different location. Compute based replication creates either a VM snapshot, or a VM
clone or a VM template. Another method is to use storage arrays to replicate the VM either to an array at the primary site itself or to a remote array (fiber channel, storage array) at the secondary site. This method is similar to the traditional array based replication method used in a CDC. It works by copying LUNs of the source VM to the target array. Replication to a remote array may be either synchronous or asynchronous.

**Virtual machine's failure**: Using the reliability method the physical server can identify the failure of VM.

![Diagram of VMs and reliability algorithm](image)

**FIGURE 1.21: THE RELIABILITY**

Figure 1.21 shows the general structure of the reliability in cloud computing. Here, file 1 is replicated in three VMs by using the reliability algorithm the reliable VM is identified.
1.9.1 Failure analysis of cloud

There are many types of failure that may affect the success/reliability of a cloud service, including queue overflow, response timeout, data missing, computing resource missing, software failure, database failure, hardware failure, and network failure.

**Queue Overflow:** The request queue should have a limited capacity on the maximum number of requests waiting in the queue, so new requests have to wait for too long a time in the queue, which could make the request timeout failures much more dominant. Therefore, if the queue is full when a new job request arrives, it is simply stopped and the user is unable to get the service, which is called an overflow failure.

**Response Timeout:** The cloud service usually has its due time set by the user or the service monitor. If the waiting time of the request in the queue is over the due time, the Timeout failure occurs. The time out requests will be dropped from the queue so as not to affect other following requests.

**Data missing:** In cloud the data resource manager registers all data resources. However, it is possible that some previously registered data are removed, but the DRM is not updated. The result causes those data resources to be assigned in a certain job request that will cause the data resource missing failure.

**Computing resource missing:** The computing resource missing is nothing but the computing resource such as the PC can turn off.

**Software failure:** The software programs running on different cloud computing resources might fail.

**Database failure:** The database that stores the required data resources may also fail, causing that the subtasks when running cannot access the required data.

**Hardware failure:** The computing resources and data resources in general may have hardware failures.
Network failure: When data accesses the remote computer, the communication channels may be broken either physically or logically, which causes the network failure, especially for those long time transmissions of large datasets.

1.9.2 Benefits of Reliability

- Identify the reliable VM among all VM.
- Identify the failure VM.

1.10 OBJECTIVES OF THE THESIS

This thesis proposes a new algorithm to improve the efficiency of cloud computing. The new approaches are “Efficient Dynamic Fuzzy Resource Level Scheduling”, “Efficient Dynamic Resource Availability Procedure” and “Improved Efficient Dynamic Resource Optimization” in which, an efficient VM is identified by load data in a balanced way and also the reliable and failure VM at the time of data loading in the cloud data center.

Why Load Balancing, Reliability and Resource Optimization for the cloud

In cloud the resource optimization is the most important process. Nowadays, many organizations use cloud storage. Many requests can occur at a time in a cloud this gives more problems in the cloud. The important problem is server crash, failover and outage. Some outages are lengthy and it may take more hours or days over solve the problem.

VMs are isolated from one another and are not dependent on the underlying x86 hardware configurations available at the failover location. The inbuilt properties of VMs allow mapping of resources from one virtual environment to another and at the same time, enable failover of an IT service. In addition, storage is replicated to the failover site, and then presented to the virtual infrastructure at the failover location where the VMs may be activated.
When there is an outage at the production site, the source LUN is made read-only and the target is made write-enabled. Applications resume operations at the secondary site. There exist certain challenges, which might prevent complete and/or correct failover of VMs in case of an outage at the production site. For example, there may exist VM placement rules specifying that selected VMs should be placed on the same server/clustered servers.

A scenario in which such constraints might be present, is when a group of VMs is communicating with one another heavily, because of which they required to be placed on the same server or within clustered servers. These constraints might prevent a VM to failover completes and/or correctly to a different server on the failover site if all of the specifies constraints are not satisfied at the failover site. To avoid failover in cloud computing efficient load balancing, reliability and resources optimization algorithms were implemented.

**TABLE 1.2: OUTAGE IN DIFFERENT CLOUD SERVERS**

<table>
<thead>
<tr>
<th>Service and Outage</th>
<th>Duration</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Azure: malfunction in Windows Azure</td>
<td>22 hours</td>
<td>March 13-14 2008</td>
</tr>
<tr>
<td>Gmail and Google Apps Engine</td>
<td>2.5 hours</td>
<td>Feb 24, 2009</td>
</tr>
<tr>
<td>Google search outage programming error</td>
<td>40 min</td>
<td>Jan 31, 2009</td>
</tr>
<tr>
<td>Google App Engine partial outage: programming error</td>
<td>5 hours</td>
<td>June 17, 2008</td>
</tr>
<tr>
<td>Flexi Scale: core network failure</td>
<td>18 hours</td>
<td>Oct 31, 2008</td>
</tr>
</tbody>
</table>

Table 1.2 shows the different failover the cloud server in different locations. At the simplest level, VMs run Guest OS, which consequently runs applications that provide IT services to the businesses. During an outage at the primary site, it is this service that must be protected from a BC point of view.
The main objectives of the thesis is to present a new approach of load balancing and reliability to improve the quality of VM by overcoming the issues of existing priority scheduling and reliability assessment algorithms. The objective of the present method is to suggest load balancing and reliability for resource optimization with such virtual servers for higher performance rate.

- To load the client data in VM based on the resource level percentage.
- To improve the QoS.
- To identify the reliability and failure VM.
- Resource Utilization efficiently.

1.11 CONTRIBUTION TO THE THESIS

The main contribution of the thesis is to develop efficient methods for load balancing reliability and resource optimization for cloud computing. The proposed algorithms is the mainly based on resource optimization.

The three proposed methods EDFRLS, EDRAP and IEDRO provide better performance than the existing methods. The present method developed in CLOUDSIM 3.0 gives the best performance.

1.12 ORGANIZATION OF THE THESIS

This thesis is organized into three phases

1. Load balancing
2. Reliable VM identification.
3. Resource Optimization

Phase1: The load balancing is used to load the resource efficiently by using the present resource level scheduling method.
Phase 2: The server identifies the failure VM by using the reliability method. The present method uses reliability in two ways, one by identifying the failure VM and the other by identifying the reliable VM after the data is loaded in the VM.

Phase 3: The phase 3 gives the comparative analysis of phase 1 and phase 2. Phase three links phase 1 and phase 2 and produces the result as resource optimization. In both phases, the resource such as memory is the most common parameter.

The aim of this work is to increase the efficiency of the cloud. In phase 1, the memory is optimized efficiently and in phase 2 the optimized memory is taken to identify the reliable and failure VM. To achieve the goals, three algorithms have been proposed and the specific contribution is structured in the following three phases

Phase I: Efficient Dynamic Fuzzy Resource Level Scheduling Algorithm (EDFRLS)
- Reduce processing time and execution time
- Balance load
- Increase speed

Phase II: Efficient Dynamic Resource Assessment Algorithm (EDRAP)
- Reduce delay
- Increase speed
- Identify reliable and failure VM

Phase III: Improved Efficient Dynamic Resource Optimization Algorithm (IEDRO)
- Reduce processing time
- Balance load
Increase speed
Reduce delay
Identify Reliable & Failure VM

The thesis is organized into the following chapters:

Chapter 1 explains the scheme of the thesis followed by the introduction of cloud computing, services, deployment in cloud, virtualization, characteristics, advantages, load balancing, reliability and objective of the thesis.

Chapter 2 deals with the literature survey and general background study of load balancing and reliability algorithm in cloud computing.

Chapter 3 explains the present methodology in two different phases. It illustrates the existing load balancing and its limitations. Chapter 4 and Chapter 5 show the significance of each phase of the present approach for cloud computing.

Chapter 4 gives the details of the present algorithm “Efficient Dynamic Fuzzy Resource Level Scheduling”. It presents the performance and the comparison of the PISA and EDFRLS. It gives the efficient load balancing and improves the QoS.

Chapter 5 deals with the present algorithm “Efficient Dynamic Resource Assessment Procedure”. It presents the performance and a comparison of the RAP and EDRAP.

Chapter 6 deals with the “Improved Efficient Dynamic Resource Optimization” algorithm to further improvements of load balancing and reliability algorithms. This algorithm uses resource optimization technique. It deals with the load balancing by the identification of failure and reliable VM.

Chapter 7 explains the conclusion and scope for future work.

The simulation of the result is performed in cloud simulator with 200 VM. The simulator achieves the efficient and the expected result in better manner.
1.13 SUMMARY

In recent years, cloud computing is the major need for all the organizations. In cloud computing, the data loading and accessing is a primary requirement. The load balancing and reliability plays a vital role in this process. However, the major outage and the speed have not been attempted in the existing load balancing and reliability algorithms. As a result, an attempt is made to develop balance load, improve the speed, minimize the delay and optimize the resource in cloud computing using the present efficient load balancing, reliability and resource optimization algorithms. To analyze the limitations in the existing algorithm some of the algorithms are discussed.