1.3 ESSENTIAL CHARACTERISTICS

1. On-Demand Self-Service

With cloud computing, you can provision computing services, like server time and network storage, automatically. You won't need to interact with the service provider. Cloud customers can access their cloud accounts through a web self-service portal to view their cloud services, monitor their usage, and provision and de-provision services.

2. Broad Network Access

Another essential cloud computing characteristic is broad network access. You can access cloud services over the network and on portable devices like mobile phones, tablets, laptops, and desktop computers. A public cloud uses the internet; a private cloud uses a local area network. Latency and bandwidth both play a major role in cloud computing and broad network access, as they affect the quality of service.

3. Resource Pooling

With resource pooling, multiple customers can share physical resources using a multi-tenant model. This model assigns and reassigns physical and virtual resources based on demand. Multi-tenancy allows customers to share the same applications or infrastructure while maintaining privacy and security. Though customers won't know the exact location of their resources, they may be able to specify the location at a higher level of abstraction, such as a country, state, or data center. Memory, processing, and bandwidth are among the resources that customers can pool.

4. Rapid Elasticity

Cloud services can be elastically provisioned and released, sometimes automatically, so customers can scale quickly based on demand. The capabilities available for provisioning are practically unlimited. Customers can engage with these capabilities at any time in any quantity. Customers can also scale cloud use, capacity, and cost without extra contracts or fees. With rapid elasticity, you won't need to buy computer hardware. Instead, can use the cloud provider's cloud computing resources.

5. Measured Service

In cloud systems, a metering capability optimizes resource usage at a level of abstraction appropriate to the type of service. For example, you can use a measured service for storage, processing, bandwidth, and users. Payment is based on actual consumption by the customer via a pay-for-what-you-use model. Monitoring, controlling, and reporting resource use creates a transparent experience for both consumers and providers of the service.

6. Resiliency and Availability

Resilience in cloud computing refers to the ability of a service to recover quickly from any disruption. Cloud resiliency is measured by how fast its servers, databases, and networks restart and recover after any damage. To prevent data loss, cloud services create a copy of the stored data. If one server loses data for any reason, the copy version from the other server restores.

Availability is a related key concept in cloud computing. The benefit of cloud services is that you can access them remotely, so there are no geographic restrictions when using cloud resources.

7. Flexibility

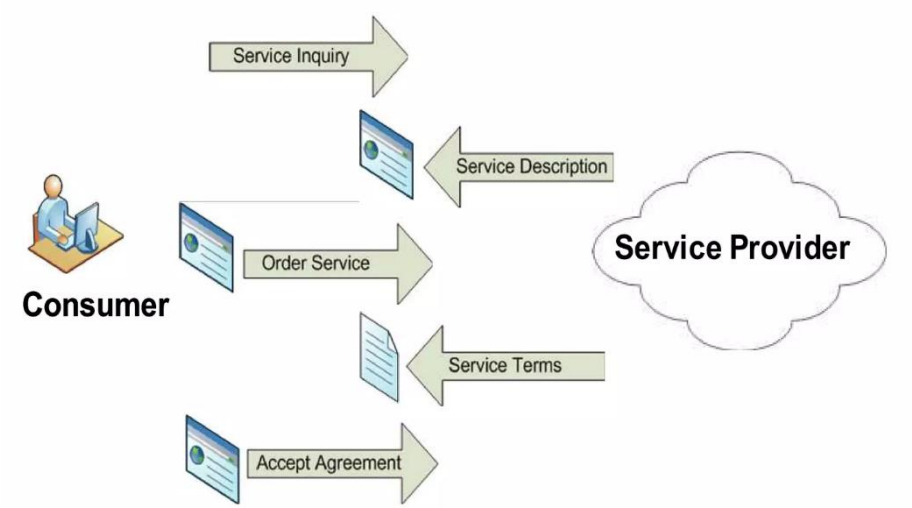
Companies need to scale as their business grows. The cloud provides customers with more freedom to scale as they please without restarting the server. They can also choose from several payment options to avoid overspending on resources they won't need.

8. Remote Work

Cloud computing helps users work remotely. Remote workers can safely and quickly access corporate data via their devices, including laptops and smartphones. Employees who work remotely can also communicate with each other and perform their jobs effectively using the cloud.

1.4 ON-DEMAND SELF-SERVICE

On-demand self-service is one of the essential characteristics of cloud computing, allowing users to access and manage computing resources without direct interaction with the service provider. This feature empowers users to control and provision resources as needed, reducing the need for human intervention and enabling a more flexible and efficient computing environment.



Here's a closer look at on-demand self-service in cloud computing:

1.User Empowerment:

On-demand self-service puts the power of resource provisioning and management into the hands of cloud users. Users can independently request, configure, and release computing resources without having to go through a lengthy process involving IT administrators or service providers.

2.Resource Provisioning:

Cloud service providers offer a range of computing resources, such as virtual machines, storage, databases, and networking, through a user-friendly web interface or API. Users can choose the required resources and deploy them instantly, often in a matter of minutes.

3.Flexible Resource Scaling:

With on-demand self-service, users can scale their resources up or down based on their needs. During periods of high demand, additional resources can be quickly provisioned to handle increased workloads. Conversely, when demand decreases, excess resources can be released, optimizing cost efficiency.

4.No Upfront Commitments:

Cloud computing's on-demand model eliminates the need for upfront commitments or long-term contracts. Users can start using the services as required and stop using them whenever needed. This pay-as-you-go approach enables cost control and aligns expenses with actual usage.

5. Automation and APIs:

Cloud providers often offer Application Programming Interfaces (APIs) that allow users to automate resource provisioning and management. This automation facilitates seamless integration with existing applications and infrastructure, streamlining workflows and improving overall efficiency.

6.Self-Service Capabilities:

The cloud provider's management portal typically provides a user-friendly interface that enables users to perform a wide range of tasks, such as creating, configuring, and terminating virtual machines, managing storage, and monitoring usage.

7.Resource Monitoring and Reporting:

On-demand self-service often includes features for monitoring resource usage and generating detailed usage reports. Users can keep track of their consumption, identify trends, and optimize resource allocation accordingly.

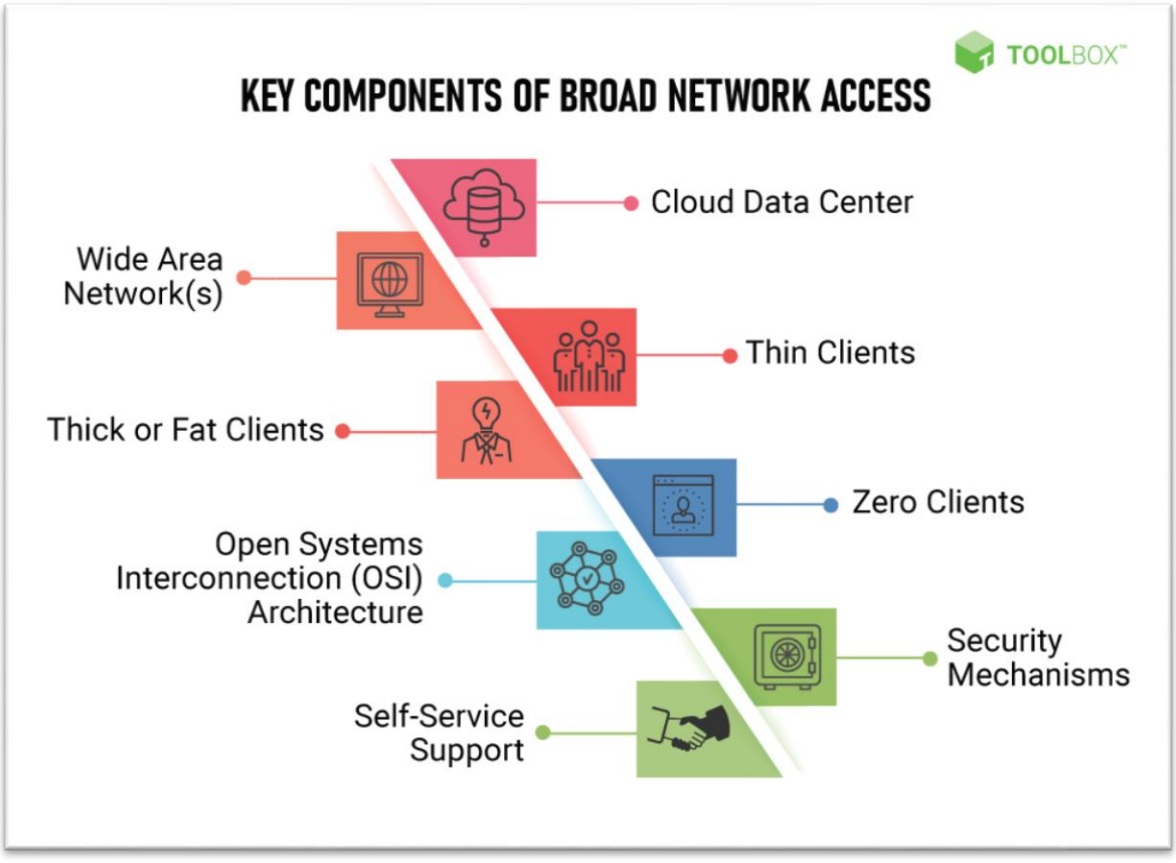
By offering on-demand self-service, cloud computing empowers users to be more agile, responsive, and independent in managing their computing resources. This flexibility is particularly beneficial for businesses and organizations with dynamic workloads that require the ability to scale resources up and down rapidly to meet changing demands.

1.5 BROAD NETWORK ACCESS

Broad network access is the ability of network infrastructure to connect with a wide variety of devices, including thin and thick clients, such as mobile phones, laptops, workstations, and tablets, to enable seamless access to computing resources across these diverse platforms. It is a key characteristic of cloud technology.

- The term broad network access can be traced back to the early days of cloud computing, when accessing resources was a complex and costly affair.
- Resources were finite and, for the most part, extremely limited as devices could only access networking and storage systems that were hosted locally.
- The cloud introduced a radical shift by democratizing access to compute, storage, and network resources.
- Broad network access is a defining characteristic of the cloud, without which the private and public cloud services we know today would not exist.

There are 8 key components in Broad Network Access. They are,



1. Cloud data center

- The cloud data center is the component housing all your network resources. It is the foundation of your broad network's access capabilities and may be managed by you, a managed services partner, or a cloud vendor like AWS, GCP, Microsoft, IBM, HPE, etc.
- The cloud data center must be set up to optimize and deliver resources across various devices without any compatibility issues.
- You should be able to log and monitor traffic from diverse platforms and optimize your cloud costs accordingly.
- Keep in mind that the "cloud" data center in this scenario could be situated in any location – your office campus, your headquarter city, your country, or on the opposite side of the world. The location will be determined by compliance and latency requirements.

2. Wide area network(s)

- The cloud data center reaches your devices (also known as clients) through one or more wide area networks.
- A wide area network (WAN) refers to a network system where the connected clients are situated at a significant distance from each other, typically more than a kilometer apart.
- Depending on the nature of your company, the cloud data center will require connectivity with multiple WANs to provide uninterrupted broad network access.
- For instance, a global organization can have dedicated cloud data centers in every continent. Each data center connects with multiple regional WANs to provide network access to devices in that perimeter.

3.Thin clients

- Now let us discuss the devices through which broad network access will be delivered. An enterprise with sufficient digital maturity will have three types of clients: thin, thick, and zero.
- A thin client refers to a computing system with very little local resources, only enough to run apps and processes when connected to a server. It is a bare-metal device without the ability to store data.
- It must connect with a remote server to fetch computing resources and have basic OS and configuration settings housed locally.

You can use thin clients in several ways. A thin client installed in your lobby can help visitors check in and out without requiring the full investment of on-device storage, operating memory, a full-featured OS, multiple apps, etc.

Virtual machines, which are also thin clients, can be used as temporary shells for software development and DevOps. Chromebooks are also an evolution of thin clients as they cannot run without network connectivity.

4. Thick or fat clients

- Most devices that end-users interact with on a daily basis fall into this category. A computing system capable of running on its own is called a thick client, or fat or rich client.
- While thin clients support internet access, it’s perfectly equipped to run as a standalone device without support from your cloud data center.
- Our laptops, office workstations, mobile phones, tablets, etc., are all thick clients as they have built-in storage, operating memory, and an OS of their own.

Your cloud data center has to be able to support thick clients of all shapes and sizes. As the device marketplace evolves, broad network access must adapt with the entry of new form factors, the retiring of once-popular devices, like Blackberry, and the rise of high-capacity consumer electronics.

5. Zero clients

- Zero clients are an emerging device category that is increasingly popular in the age of IoT.
- Zero clients are defined as devices with no storage capacity and can run only ultra-lite and low bandwidth applications.
- IoT systems like wearables that collect real-time information or equipment sensors fall into this category. Unlike a thin client, it does not have operating systems or on-device configurations.
- Its only purpose is to collect data from its surroundings and relay data generated by your central cloud data center to users by acting as a terminal. Broad network access has to be able to power zero clients as well.

6. Open systems interconnection (OSI) architecture

- OSI is at the heart of the deployment of broad network access. It tells the cloud data center how it should connect with the WAN and how the WAN, in turn, should connect with the end-thin, thick, or zero clients.
- OSI architecture, as defined by the International Organization for Standards (ISO) and the International Electrotechnical Commission (IEC), defines OSI as having seven layers that allow a network to receive and transmit data. These are:
    - Physical layer
    - Data link layer
    - Network layer
    - Transport layer
    - Session layer
    - Presentation layer
    - Application layer

How you (or your cloud vendor or managed services provider) set up each of these layers will determine the performance of your broad network access landscape.

7. Security mechanisms

- A key challenge of broad network access is security, which is why it must be considered as a key component of your network environment.
- By definition, broad network access allows multiple devices to log in to a network and gain from cloud-delivered resources.
- If left unattended, this could open backdoors and vulnerabilities that a malicious entity can exploit.
- Therefore, network security must be woven into your SLAs at the web, application, network, and device levels.

You may want to collaborate with stakeholders across your IT supply chain to address any security risks across the end-to-end broad network access ecosystem.

8. Self-service support

- Finally, self-service is a vital characteristic that makes your cloud *genuinely* a cloud.
- For example, a public or private cloud infrastructure may require users to submit a complaint or service ticket instead of solving the issue themselves.
- This goes against one of the five defining characteristics of the cloud as laid down by NIST and could significantly hinder broad network access.

Let’s say that your cloud environment supports access from mobile phones – but when an employee tries to clock in from their phone in the morning before they can open their workstation, the service is unavailable.

In such scenarios, the employee should be able to quickly look up and resolve the issue via a self-service knowledge base. In the absence of self-service, true and effective broad network access isn’t possible.

1.6 LOCATION INDEPENDENT RESOURCE POOLING

Resource pooling is an information technology (IT) term used in cloud computing environments to describe a situation in which suppliers deliver temporary and expandable services to numerous clients, customers, or "tenants." These services can be adjusted to meet each client's demands without requiring the client or end-user to notice any changes.

Cloud Computing platforms are accessible via internet connection. It can also be shared, maintained, or developed platforms that provide specific services. These are also cutting-edge technologies that provide clients with greater flexibility and scalability. In the cloud computing resource sharing paradigm, the service provider serves numerous clients simultaneously. To handle and deal with such clients, they employ a multi-tenant approach.

What is resource pooling in cloud computing?

Cloud computing platforms can be accessed through an internet connection. It can also be shared, managed, or developed platforms to provide specialized services. Furthermore, these are cutting-edge technologies that provide clients with flexibility and scalability.

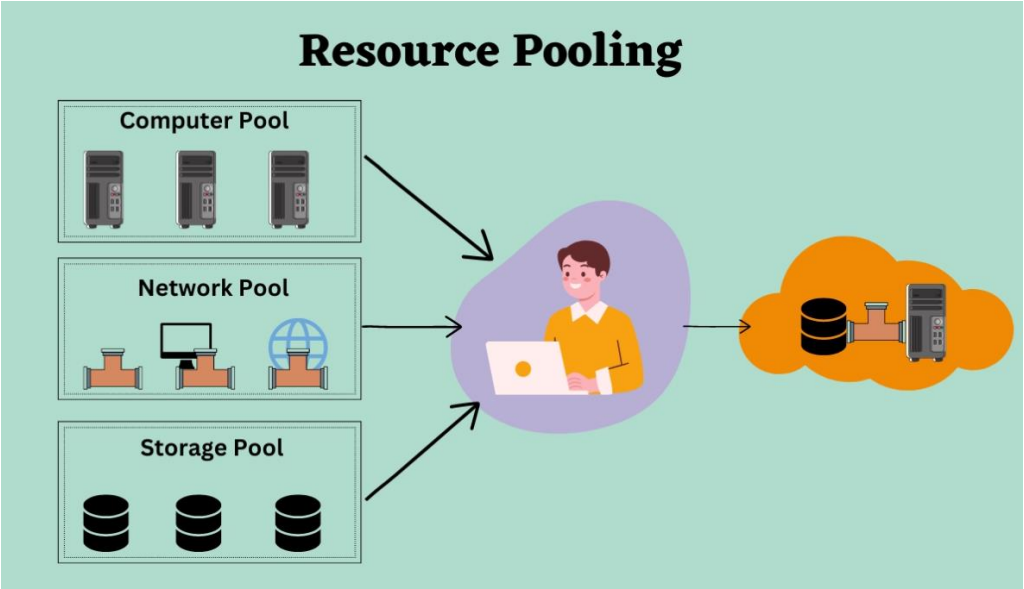
The service provider serves multiple clients at the same time in the cloud computing resource pooling model. To handle and deal with such clients, they employ a multi-tenant model.

How does resource pooling work?

- The user can choose the best resource division for his needs in this private cloud as a service. The most crucial factor while resources pooling is cost-effectiveness. It also ensures that the company offers new service delivery options.
- It is utilized a lot in wireless technology like radio communication. And it's here that single channels come together to establish a strong link. As a result, the connection can transport data without being interrupted.
- Resource pooling is a multi-tenant operation in the cloud-dependent on user demand.
- Additionally, as more people begin to use the software as a service(SaaS) services, the charges for these services usually remain relatively inexpensive.
- As a result, owning such technology has become more accessible than previously.

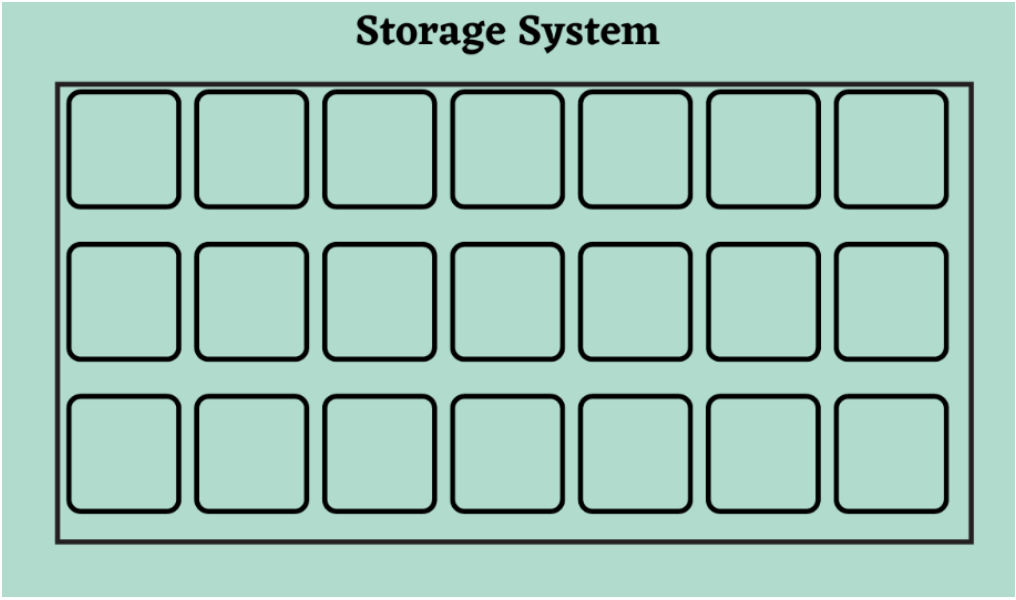
The pool is built in a private cloud, and cloud computing resources are assigned to the user's Internet protocol (IP) address. As a result, the resources continue to send data to an ideal cloud service platform by visiting the IP address.

Types of Resource pooling

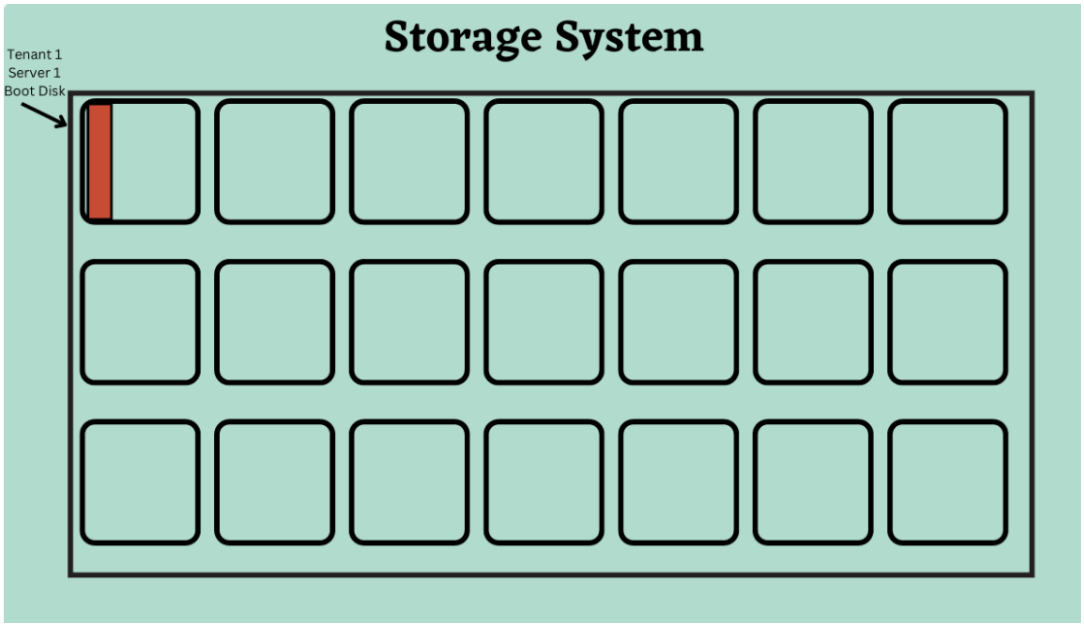


Storage Pooling

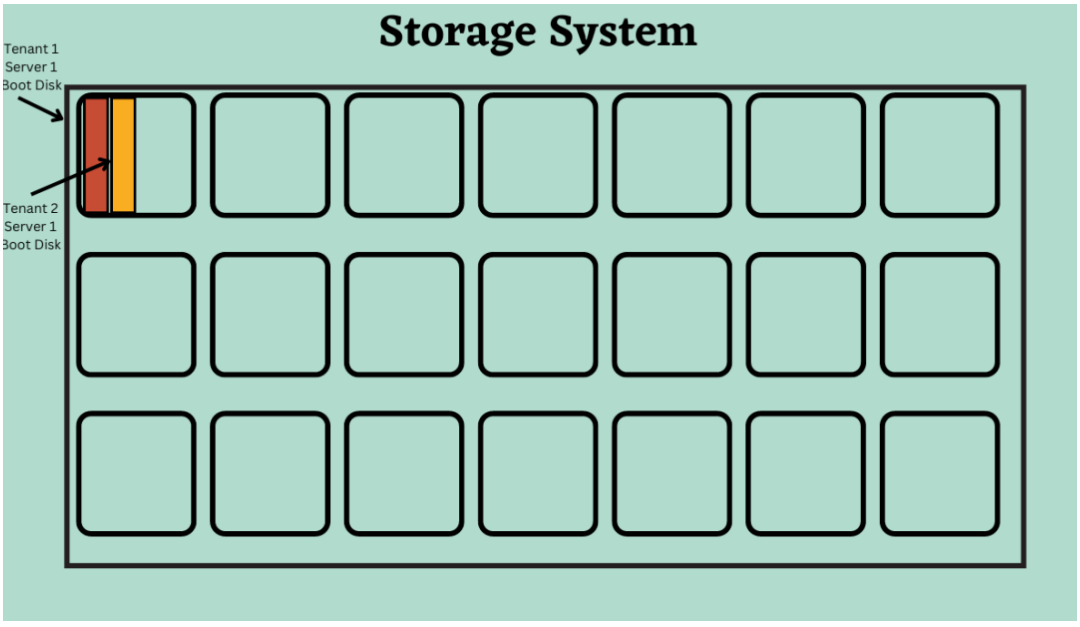
We're going to look at the storage pooling. The big blue box in the diagram below illustrates a storage system with multiple hard drives. Each of the tiny white squares represents one of the hard drives.



We can slice up our storage anyway we like using our centralized storage and offer virtual machines their own small portion of that storage for however much space they desire. We'll use a slice of the first disc as the boot disc for 'Tenant 1, Server 1' in the example below.



We will take another slice of our storage and provision that as the boot disk for ‘Tenant 2, Server 1’.



Instead of allocating entire discs to separate servers, we can offer them the amount of storage they require with centralized shared storage. Storage efficiency strategies like thin provisioning, deduplication, and compression can help us save even more money.

Network Infrastructure Pooling

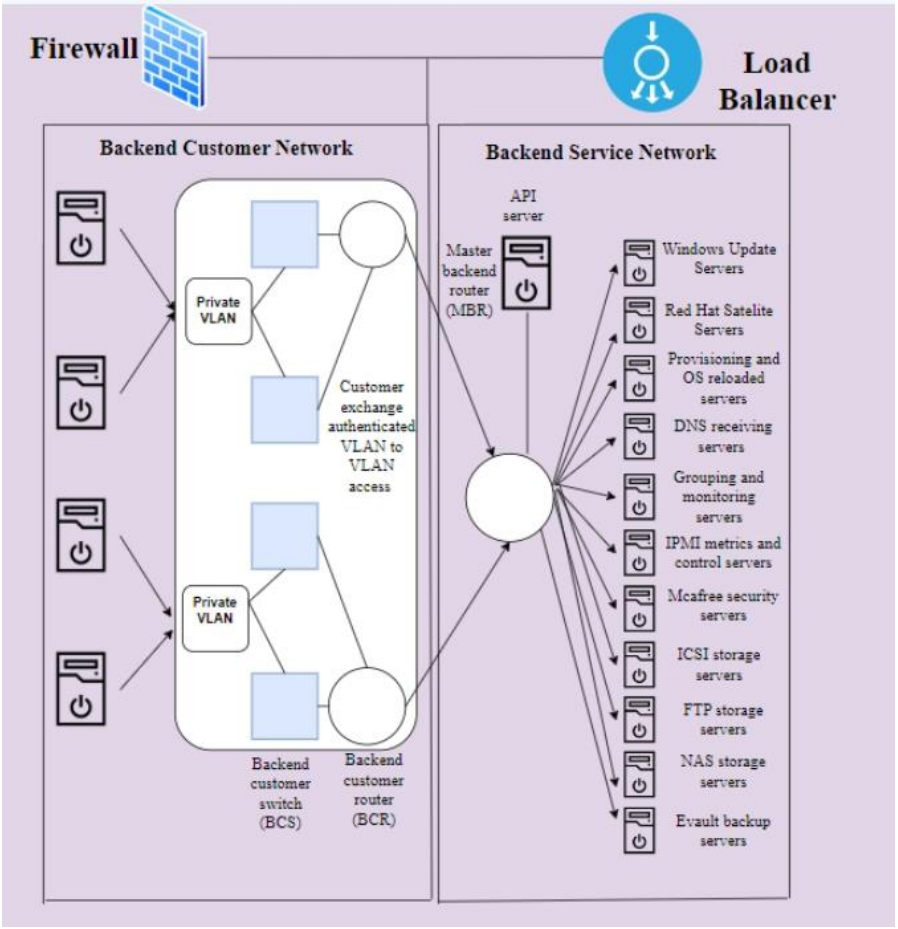
Next, we are going to look at network infrastructure pooling.

Firewall rules will restrict what traffic is allowed to each tenant's virtual machines, such as Remote Desktop Protocol(RDP) for management and Hypertext Transfer Protocol(HTTP) traffic on port 80 if the virtual machine is a web server.

We don't have to provide each customer with a physical firewall; instead, we can share one physical firewall among multiple clients. Incoming connection load balancers can also be virtualized and shared across numerous clients.

Several switches and routers may be seen in the central part on the left side of the figure. These switches and routers are shared, with traffic traveling to different customers over the same device.

A physical firewall is present at the top of the diagram below.



Service pooling

On the right side of the diagram, we can see that the cloud provider also offers various services to its customers. Windows Update and Red Hat Update Server handles operating system patching, Domain Name System(DNS), and other tasks. Customers are relieved of needing to offer their DNS solution by keeping DNS as a centralized service.

The benefits of using resource pooling in cloud computing are:

- High availability rate
- Balanced load on the server
- Provides high computing experience
- Stored data Virtually and physically
- Flexibility for businesses
- Handling virtual host failure

1.7 RAPID ELASTICITY

Elasticity is a 'rename' of scalability, a known non-functional requirement in IT architecture for many years already. Scalability is the ability to add or remove capacity, mostly processing, memory, or both, from an IT environment.

Ability to dynamically scale the services provided directly to customers' need for space and other services. It is one of the five fundamental aspects of cloud computing.

It is usually done in two ways:

- **Horizontal Scalability**: Adding or removing nodes, servers, or instances to or from a pool, such as a cluster or a farm.
- **Vertical Scalability**: Adding or removing resources to an existing node, server, or instance to increase the capacity of a node, server, or instance.

Most implementations of scalability are implemented using the horizontal method, as it is the easiest to implement, especially in the current web-based world we live in. Vertical Scaling is less dynamic because this requires reboots of systems, sometimes adding physical components to servers.

A well-known example is adding a load balancer in front of a farm of web servers that distributes the requests.

Why call it Elasticity?

Traditional IT environments have scalability built into their architecture, but scaling up or down isn't done very often. It has to do with Scaling and the amount of time, effort, and cost.

Servers have to be purchased, operations need to be screwed into server racks, installed and configured, and then the test team needs to verify functioning, and only after that's done can you get the big There are. And you don't just buy a server for a few months - typically, it's three to five years. So it is a long-term investment that you make.

Three forms for scalability

1. **Manual Scaling**
    - Manual scalability begins with forecasting the expected workload on a cluster or farm of resources, then manually adding resources to add capacity.

- Ordering, installing, and configuring physical resources takes a lot of time, so forecasting needs to be done weeks, if not months, in advance.
- It is mostly done using physical servers, which are installed and configured manually.
- Another downside of manual scalability is that removing resources does not result in cost savings because the physical server has already been paid for.

2. **Semi-automated Scaling**
   - Semi-automated scalability takes advantage of virtual servers, which are provisioned (installed) using predefined images. A manual forecast or automated warning of system monitoring tooling will trigger operations to expand or reduce the cluster or farm of resources.
   - Using predefined, tested, and approved images, every new virtual server will be the same as others (except for some minor configuration), which gives you repetitive results.
   - It also reduced the manual labor on the systems significantly, and it is a well-known fact that manual actions on systems cause around 70 to 80 percent of all errors.
   - There are also huge benefits to using a virtual server; this saves costs after the virtual server is de-provisioned. The freed resources can be directly used for other purposes.

3. **Elastic Scaling (fully automatic Scaling)**
   - Elasticity, or fully automatic scalability, takes advantage of the same concepts that semi-automatic scalability does but removes any manual labor required to increase or decrease capacity.
   - Everything is controlled by a trigger from the System Monitoring tooling, which gives you this "rubber band" effect. If more capacity is needed now, it is added now and there in minutes. Depending on the system monitoring tooling, the capacity is immediately reduced.

Scalability vs. Elasticity in Cloud Computing

Imagine a restaurant in an excellent location. It can accommodate up to 30 customers, including outdoor seating. Customers come and go throughout the day. Therefore restaurants rarely exceed their seating capacity.

- The restaurant increases and decreases its seating capacity within the limits of its seating area. But the staff adds a table or two to lunch and dinner when more people stream in with an appetite. Then they remove the tables and chairs to de-clutter the space.
- A nearby center hosts a bi-annual event that attracts hundreds of attendees for the week-long convention.
- The restaurant often sees increased traffic during convention weeks. The demand is usually so high that it has to drive away customers. It often loses business and customers to nearby competitors. The restaurant has disappointed those potential customers for two years in a row.
- Elasticity allows a cloud provider's customers to achieve cost savings, which are often the main reason for adopting cloud services.
- Depending on the type of cloud service, discounts are sometimes offered for long-term contracts with cloud providers. If you are willing to charge a higher price and not be locked in, you get flexibility.

Let's look at some examples where we can use it.

Cloud Rapid Elasticity Example 1

Let us tell you that 10 servers are needed for a three-month project. The company can provide cloud services within minutes, pay a small monthly OpEx fee to run them, not a large upfront CapEx cost, and decommission them at the end of three months at no charge.

We can compare this to before cloud computing became available. Let's say a customer comes to us with the same opportunity, and we have to move to fulfill the opportunity. We have to buy 10 more servers as a huge capital cost.

When the project is complete at the end of three months, we'll have servers left when we don't need them anymore. It's not economical, which could mean we have to forgo the opportunity.

Because cloud services are much more cost-efficient, we are more likely to take this opportunity, giving us an advantage over our competitors.

Benefits and Limitations of Cloud Elasticity

Elasticity in the cloud has many powerful benefits.

1. **Elasticity balances performance with cost-effectiveness**

- An Elastic Cloud provider provides system monitoring tools that track resource usage. Then they automatically analyze resource allocation versus usage.
- The goal is always to ensure that these two metrics match to ensure that the system performs cost-effectively at its peak.

2. **It helps in providing smooth services.**
   - Cloud elasticity combines with cloud scalability to ensure that both the customer and the cloud platform meet changing computing needs when the need arises.
   - For a cloud platform, Elasticity helps keep customers happy.
   - While scalability helps it handle long-term growth, Elasticity currently ensures flawless service availability. It also helps prevent system overloading or runaway cloud costs due to over-provisioning.

The limits or disadvantages of cloud elasticity:

- Cloud elasticity may not be for everyone. Cloud scalability alone may be sufficient if you have a relatively stable demand for your products or services online.

1.8 MEASURED SERVICE

Measured services, also known as metered services or pay-as-you-go billing, is a fundamental characteristic of cloud computing. It refers to the concept of monitoring, measuring, and charging users based on their actual usage of computing resources and services provided by the cloud service provider.

This model allows users to pay only for the resources they consume, promoting cost efficiency and transparency. Here's a closer look at measured services in cloud computing:

1. **Usage Monitoring:**

- Cloud service providers continuously monitor the usage of various resources, including computing power, storage, bandwidth, and active user accounts.
- The monitoring is done in real-time, enabling accurate tracking of resource consumption.

2. **Resource Consumption Metrics:**

- Different cloud services have specific metrics for measuring usage.

For example:

- Compute resources (CPU, RAM): Measured in hours of usage or compute capacity utilized.

- Storage: Measured in gigabytes (GB) or terabytes (TB) of data stored.

- Bandwidth: Measured in the volume of data transferred in and out of the cloud environment.

3. Pay-as-You-Go Model:

- With measured services, users are billed on a pay-as-you-go basis. The billing is typically calculated per hour, per minute, or even per second, depending on the cloud provider and the service used.
- Users are only charged for the precise amount of resources they use during a given time frame.

4.Cost Transparency:

- Measured services offer cost transparency, as users can easily track their resource usage and corresponding costs through the cloud provider's management console or billing dashboard.
- This helps users understand and optimize their spending.

5.Flexibility and Cost Control:

- The measured services model allows users to adjust their resource allocation according to their needs.
- Users can scale resources up or down dynamically, and they will only pay for the resources used during specific periods, which can help optimize costs.

6.Automatic Scaling and Load Balancing:

- Measured services are often closely integrated with auto-scaling and load balancing features.
- These capabilities automatically adjust resource allocation based on demand, ensuring optimal performance and cost-effectiveness.

7.No Upfront Commitments:

- Measured services eliminate the need for upfront commitments or fixed contracts.
- Users can start using the services without any minimum usage commitments and can stop using them whenever they want.

Measured services are a critical component of cloud computing, providing a flexible and cost-effective approach to resource utilization. This model is particularly advantageous for businesses with varying workloads, as it allows them to respond dynamically to changing resource demands without overprovisioning or incurring unnecessary expenses.

Key aspects of measured services in cloud computing include:

1. **Pay-as-you-go:**

   Customers are billed based on the actual usage of resources, and they pay only for what they use. This provides cost-efficiency, as businesses do not have to invest in and maintain expensive infrastructure that might remain underutilized.

2. **Resource scaling:**

Measured services enable automatic scaling of resources to match the changing demands. This means that during periods of high demand, resources can be dynamically increased, and during low-demand periods, resources can be scaled down, ensuring optimal resource utilization and cost-effectiveness.

3. **Resource tracking and reporting:**

Cloud providers track resource usage for each customer and provide detailed usage reports. These reports help customers understand their consumption patterns, identify areas of optimization, and manage costs more effectively.

4. **Metering and monitoring:**

The cloud provider employs metering and monitoring tools to measure the consumption of various resources accurately. This data is used for billing purposes and to optimize the performance of the cloud infrastructure.

5. **Resource pooling:**

Cloud providers pool computing resources across multiple customers, creating a multi-tenant environment. The pooled resources are dynamically allocated to different users based on their needs, ensuring efficient resource utilization.

Overall, measured services in cloud computing promote flexibility, scalability, and cost-effectiveness by aligning resource allocation and billing with actual usage. This model allows businesses to adapt to changing demands and avoid overprovisioning, making cloud computing an attractive option for various organizations.