# Scalability and Redundancy

## Chapter 5

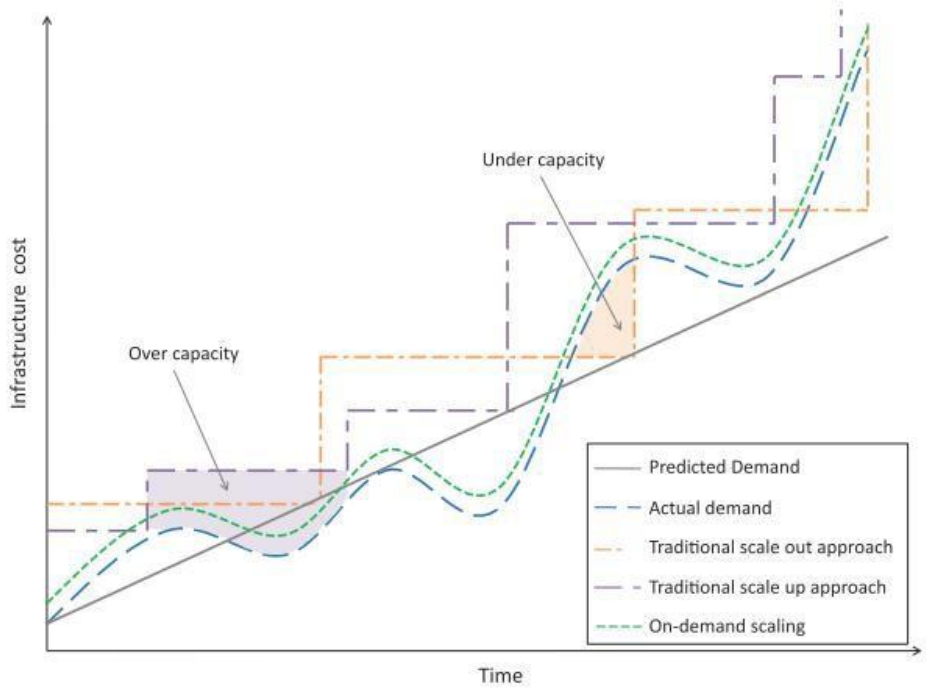
# Outline

# Scalability

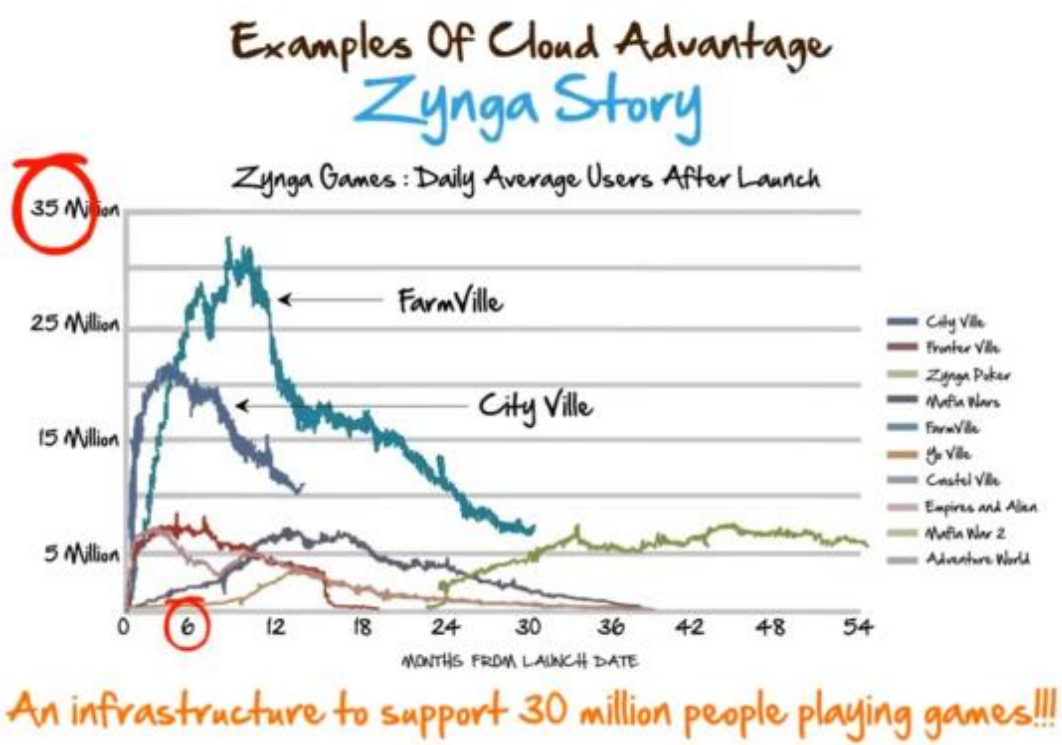
- The ability of a system to increase or decrease the quantity of resources as per the demand
- One of the biggest advantages of cloud computing
  - Users can increase the number of resources when there is workload and decrease the resources level when there is less or no demand
- For example: Ram is handling his business firm and Users can access his business through an app
- Ram uses cloud computing services to manage and maintain his business firm effectively
- Suddenly thousands of visitors start accessing his business supplies through the app
  - Increment the traffic on his app, leading to workload
  - As crashing or slow accessing of the app may make the user upset and create the wrong impression about its services

# Scalability

- Cloud scalability enables resources to grow as your business grows or traffic increases and shrinks when the load diminishes
- Scalability can apply to
  - CPU
  - Storage
  - Memory
  - Network I/O



# Case study: Zynga



# Case study: Zynga

# Case study: Friendster

- Lunch before mySpace, Facebook
- Had 3 million users within the first few months
- Problem: become slower and slower
  - took minutes to log in
  - more time to send out messages
- Give opportunities to myspace and Facebook
- Did not solve the first basic problem: site didn't work
  - solution: keep the servers up and running i.e: scaling

# Case study: Cuil

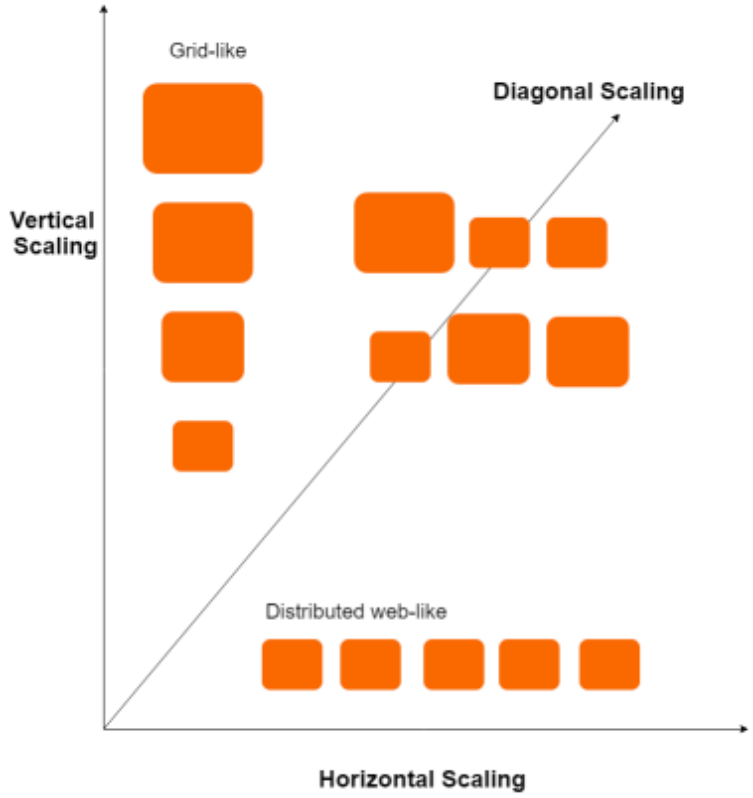
- Search Index competition between: google and yahoo
- 3 Ex Google employees announced to launch search engine: CUIL
  - Index: 3 times the size of google
  - considered as Google Killer
- Become popular but Cuil anticipated: decent traffic (especially tech community)
- As launched: server crashed due to excessive load
  - serious files corruption due to overloading
  - entire index get corrupted

# Key features of cloud scalability

1. **Growing or shrinking**
   - Scalability allows you to increase or decrease resources to meet the demands of the business
2. **Performance**
   - Handle traffic burst and huge workload, thereby increasing the performance of the business or organization
3. **Cost-efficient**
   - Pay only for those resources that you use
   - It gives the ability to scale out the resources that are of no use, thus saving money and making your system cost-efficient
4. **Capacity**
   - Automatically increases the storage capacity of the system as your business grows
   - No need to worry about additional space needed for handling the organization's continuously growing data

# Types of scalability

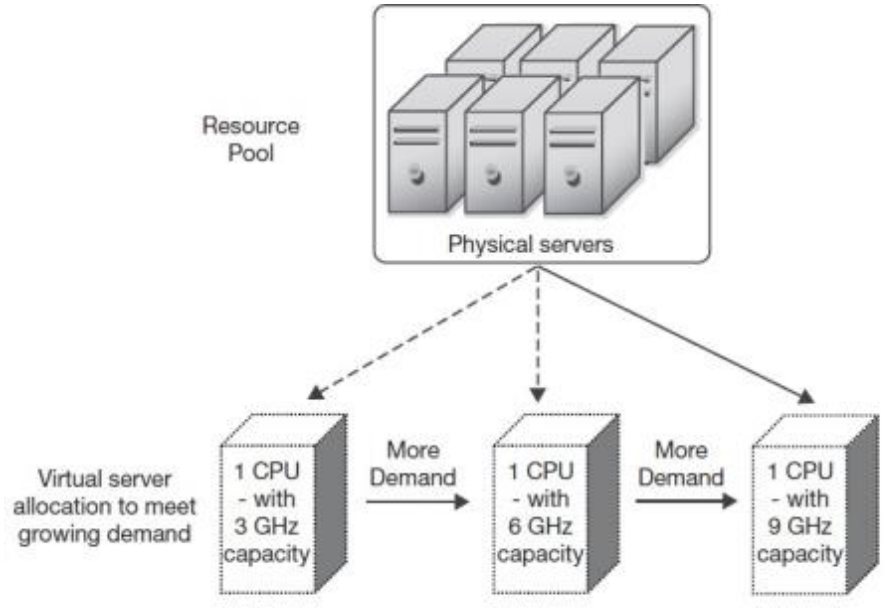
1. Vertical scaling (scale-up)
2. Horizontal scaling (scale-out)
3. Diagonal scaling



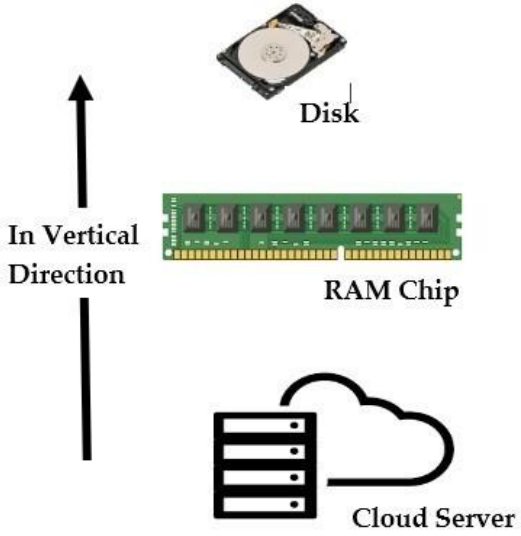
# Vertical Scaling

- Vertical scaling refers to adding more power to the existing infrastructure by adding more resources
- Adding a more powerful processor (CPU), more memory (RAM), faster storage such as solid-state drives (SSDs)
- Done to increase the capacity of existing hardware or software by adding resources
- Enhance server without manipulating your code
- Limited by the fact that you can only get as big as the size of the server

# Vertical Scaling




Computing resource scaled up by replacing it with more powerful resource as demand increases



# Advantages of Vertical Scaling

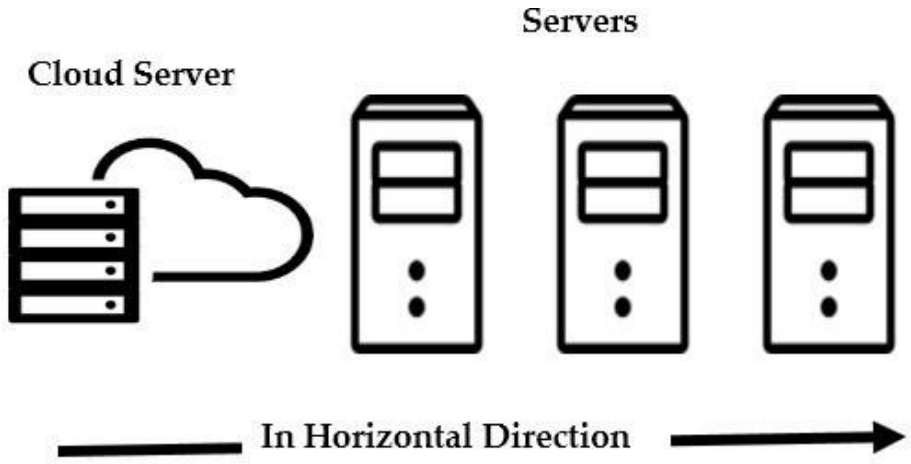
- Reduced software costs
- Easy Implementation
- Licensing fees is less
- Consumes less power
- Cooling costs are lesser than horizontal scaling
- Application compatibility is maintained

# Disadvantages of Vertical Scaling:

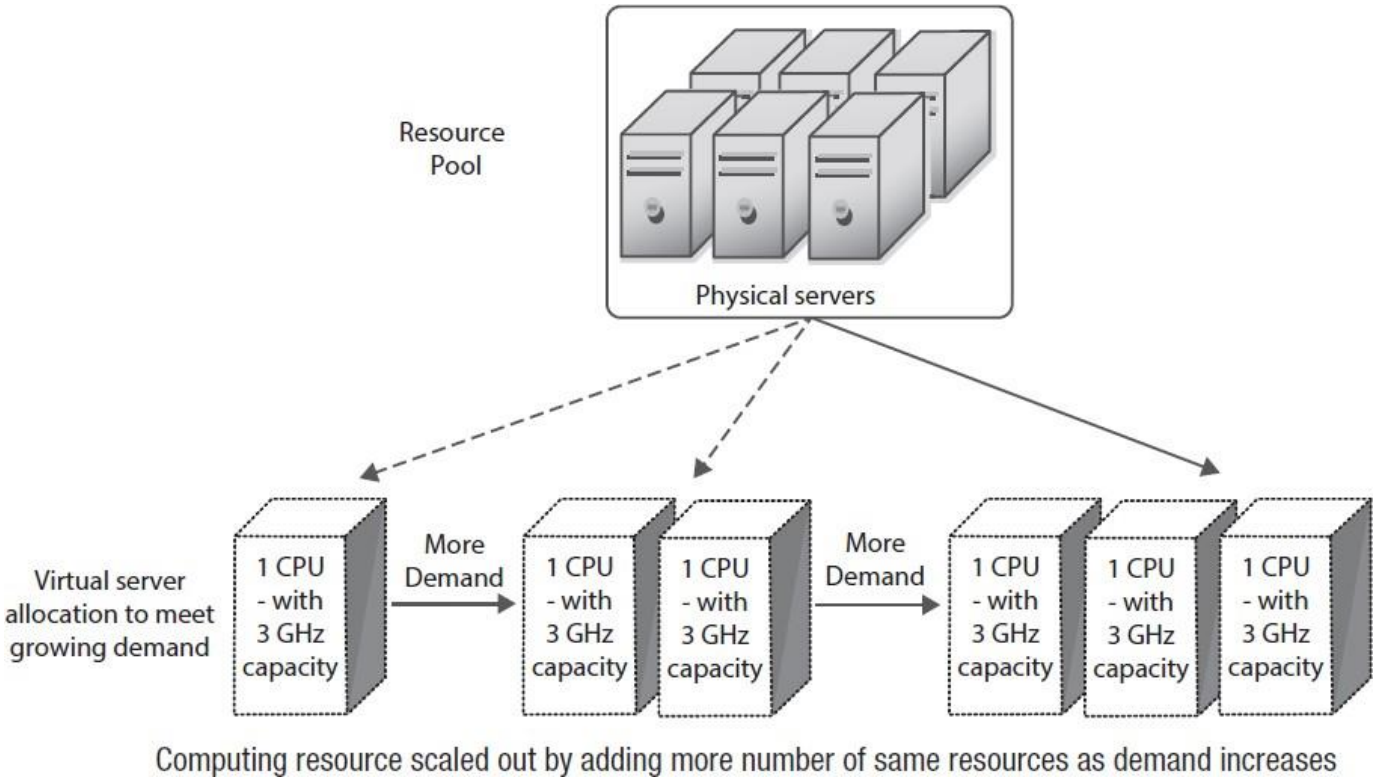
- Limited Scaling
- The risk for downtime is much higher than horizontal scaling.
- Greater risk of outages and hardware failures.
- Finite scope of upgradeability in the future.
- Severe vendor lock-in.
- The cost of implementing is expensive

# Horizontal scaling

- Improves the cloud throughput by adding new computing infrastructure
- Adding new computing nodes or machines to enhance the data processing and storage capabilities
- Harnesses the power of multiple physical machines, i.e. distributing computing
- Allocates processing and storage tasks across physical machines connected by a single network



# Horizontal scaling




Computing resource scaled out by adding more number of same resources as demand increases

# Advantages of Horizontal Scaling

- **Simple process:**
  - No need to analyze existing specifications and calculate which ones need to be upgraded
  - Add more compatible computing nodes or machines to the existing data center
- **Enhanced performance:**
  - Increased network traffic enhances performance better than scaling up (as vertical scaling is known)
  - Workload is shared among multiple machines, and new machines can be added as and when required
- **Reduced downtime:**
  - Adding a machine requires minimal, if any, downtime for the existing machines
- **Higher fault tolerance:**
  - No need to rely on one server or machine for all operations
  - By distributing the workload among many nodes, the risk of data loss is minimized in case of any fault

# Disadvantages of horizontal scaling

- Increase in the complexity of operations and maintenance
- Backup also becomes difficult as the synchronization and communication among numerous machines need to be addressed
- Increase cost compared to vertical scaling

# Horizontal Scaling is more Cloud-Native approach

- Cloud computing model promises to deliver infinite scalability
- Vertical scaling has limitations and it can only grow as per available resource components
- The infinite scalability promise cannot be fulfilled with a vertical scaling approach in the cloud computing environment where application load may reach the zenith on different occasions
- Horizontal scaling spreads the load across multiple resource nodes and can thus support load unrestricted by any resource component's ability
  - The more resource nodes are added, the more load can be supported and thus this scaling approach is rather pertinent to deliver infinite scalability

# Diagonal Scaling

- Mixture of both Horizontal and Vertical scalability
  - resources are added both vertically and horizontally
  - means increasing both the size and number of instances simultaneously
- Grow within the existing server until it hits the capacity
- This approach is often used when an application is resource-intensive and requires more capacity to handle the workload
- For example, suppose a web application is experiencing high traffic volume that is causing slow response times
  - To improve performance, diagonal scaling could be used to add more instances of the application, while also increasing the resources of each instance
  - This would increase the overall capacity of the application while also ensuring that each instance has enough resources to handle the workload

# Ways to scale cloud

1. Manual
2. Scheduled
3. Automatic/ Dynamic

## Manual

1.The concerned person should always have a look at the load and traffic of the server so that he does not miss any chance of scaling up the resources to meet the demands

2. S/he should never forget to scale down the resources when the demand is low

it may **lead** to the extra cost

prone to human error

# Scheduled

- Instead of keeping an eye on the market trends and manually scaling up and down the resources, you can schedule your scaling for the peak time
  - consider the timings when the application is most likely to suffer from traffic or workload
- Set up a scheduled time to scale up or in the resources during that particular period, and scale down or out during normal routine
  - Friday night to Monday morning (as most weekends correspond to more traffic)
  - you can schedule scaling up or in and for the rest of the time, it will remain at the normal scale
- Failure if unexpected traffic or peak time occurs

# Automatic

- The process of automatically scaling up or down and scaling in or out the resources to meet the demands
- Auto scaling monitors your business trend and automatically adjusts the scaling up or down to maintain consistent growth at a low cost
- Resources are automatically increased during demand spikes and decreased when demand drops

# Benefits of cloud scalability

- **Convenience:**
  - Add more Virtual Machines that are available without delay—and customized to the exact needs of an organization
- **Flexibility and speed:**
  - As business needs change and grow—including unexpected spikes in demand—cloud scalability allows IT to respond quickly
  - Even smaller businesses have access to high-powered resources that used to be cost-prohibitive
- **Cost savings:**
  - Businesses can avoid the upfront costs of purchasing expensive equipment that could become outdated in a few years
  - Using cloud providers, organization pay for only what they use and minimizes waste
- **Disaster recovery:**
  - With scalable cloud computing, you can reduce disaster recovery costs by eliminating the need for building and maintaining secondary data centers

# Which scaling method is best for your app?

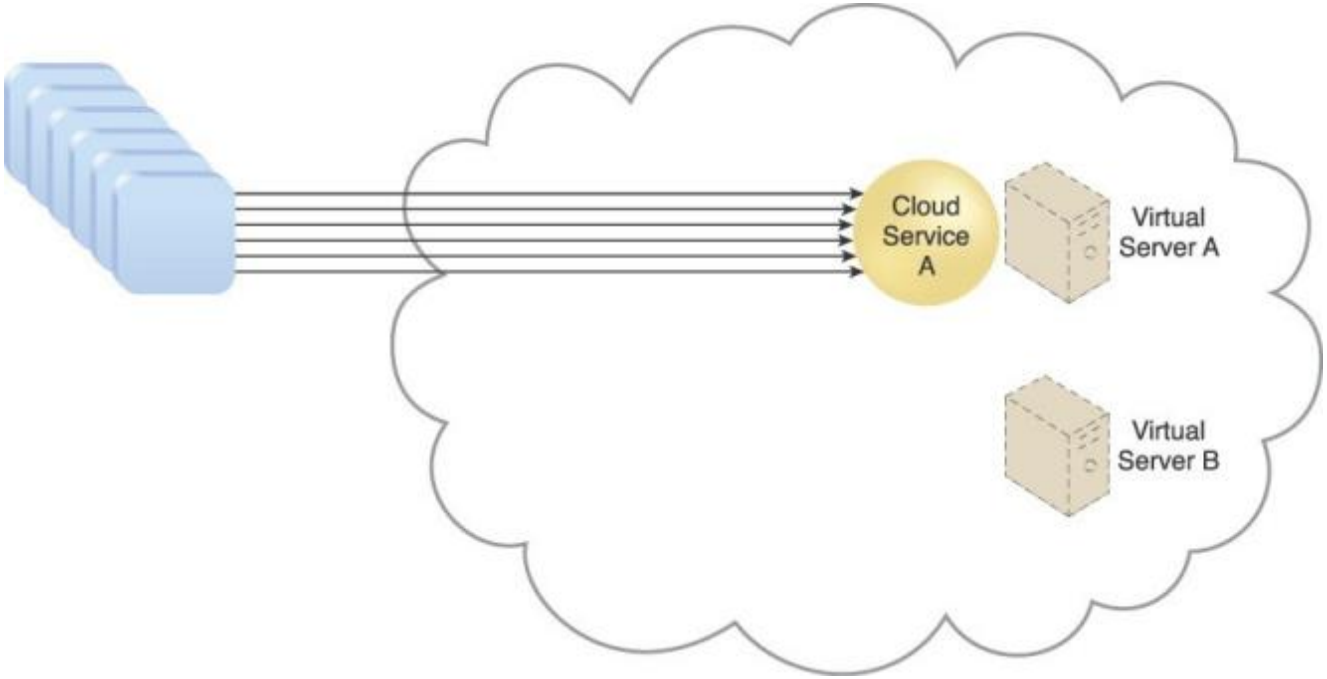
- **Load balancing**
  - Fundamental mechanism used to establish the base workload balancing logic in order to carry out the distribution of the workload
  - Vertical scaling system is best for balancing loads because there is no need to balance in the single server (vertical scaling)
  - Horizontal scaling requires you to balance the workload evenly
- **Point of failure**
  - The horizontal scaling system has more than one server, so when one server crashes, the next one picks up the slack
  - The vertical scaling system has only one server, so once the server crashes, everything goes offline
- **Speed**
  - The vertical scaling system is faster because it runs on one server, the vertical scaling system has interprocess communication, i.e., the server communicates within itself and it's fast
  - The horizontal scaling system has network calls between two or more servers
    - Remote Procedure Calls (RPC) are generally slow

# Which scaling method is best for your app?

- **Data consistency**
  - The vertical scaling system is data consistent because all information is on a single server
  - The horizontal scaling system is scaled out with multiple servers, so data consistency can be a huge issue
- **Hardware limitations**
  - The horizontal scaling system scales well because the number of servers is linear to the number of users in the database or server
  - The vertical scaling system has a limitation because everything runs on a single serve

# How can IT resource over-utilization be avoided?

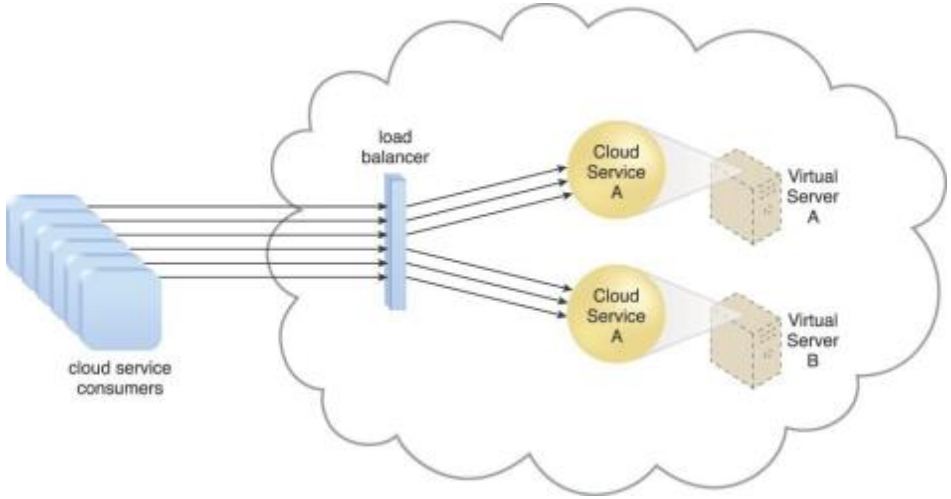
Problem: IT resources that are shared or are made available to consumers with unpredictable usage requirements can become over-utilized when usage demands near or exceed their capacities. This can result in runtime exceptions and failure conditions that cause the affected IT resources to reject consumer requests or shut down altogether.



# How can IT resource over-utilization be avoided?

Solution:

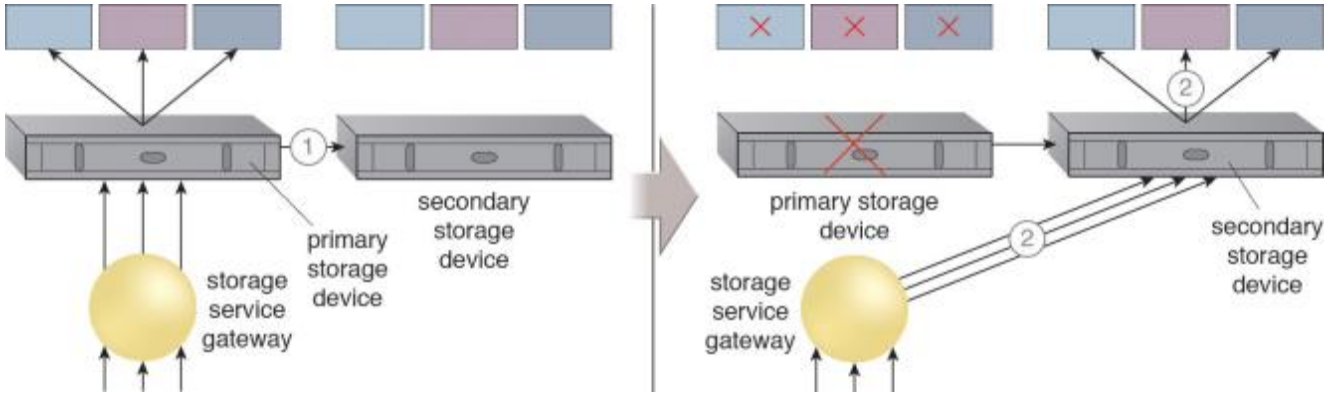
- The IT resource is horizontally scaled via the addition of one or more identical IT resources
- A load balancing system further extends the cloud architecture to provide runtime logic capable of evenly distributing the workload across all available IT resources
- This minimizes the chances that any one of the IT resources will be over-utilized (or under-utilized)



# Redundancy

- Redundancy means duplication
  - act of duplicating data or files so as to provide universal access and backup for clients
  - keep multiple copies of data on multiple servers
- Main motive is to prevent or recover from any failure
  - if one server fails, your data is safe and secured on other servers.
- Cloud computing saves your data on multiple cloud servers such that if any server becomes corrupted or is unavailable (maintenance or needs to be upgraded), you can still access your data from other servers
- Redundancy is not only limited to creating multiple copies of data
  - means to have a duplicate copy of server, system, and other devices so that if anyone out of these gets failed, you can easily access it from the redundant option
- Cloud storage allows companies to have a safe and secure copy of their important documents and databases, thereby ensuring data protection

# Redundancy


1
primary
storage
device
secondary
storage
device
storage
service
gateway
primary storage
device
2
2
secondary
storage
device
storage
service
gateway

# Benefits of redundancy

- **Automatic data backups:**
  - Cloud storage enables automatic data backup for all applications and data
  - For example Google Docs, the data is auto-saved on the cloud and if any disruption occurs
    - the next time that you use the application, it gives you the facility of resuming your work from that point only where the disruption occurred
- **Access from any device, any time, any place:**
  - cloud vendor automatically stores our data on multiple servers
  - This allows us to access our data from any device, any time anywhere
- **Data security:**
  - ensures the security of data by storing it on multiple servers
  - if one system fails, then the other can be used for accessing the data
- **Uptime guaranteed by service-level agreement (SLA):**
  - Data and applications are always available for use
  - Cloud vendors ensure a continual uptime for data and application performance