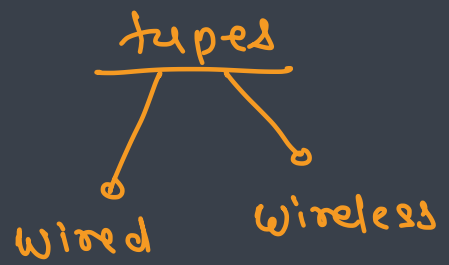
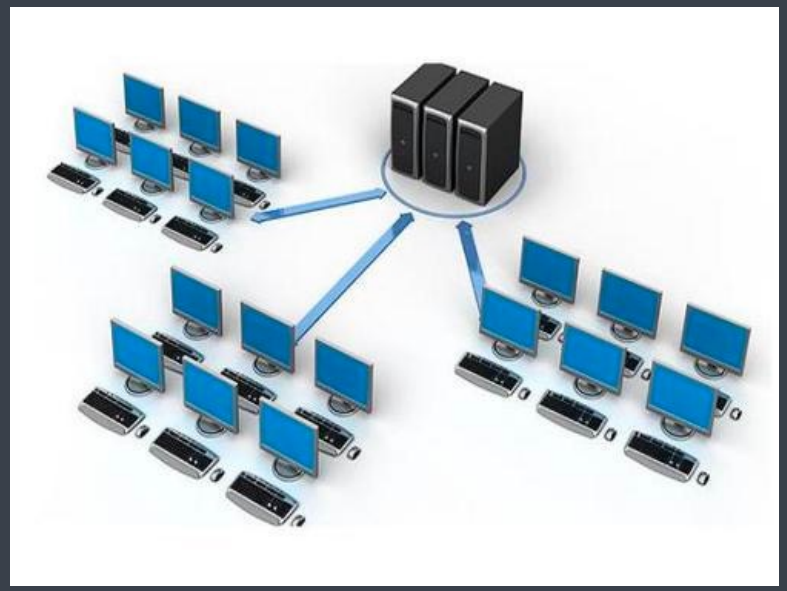
Networking

What is network ?

- It is the interconnection of multiple devices, generally termed as Hosts connected using multiple paths
- The purpose of network is
  - sending/receiving data or media
- It involves various devices like hubs, switches, routers etc.

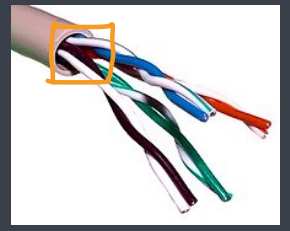




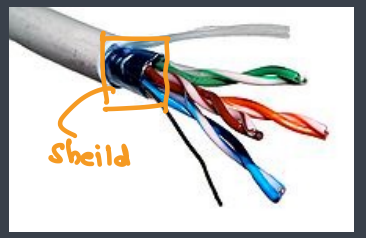
Wired network

- The network build by connecting devices together using wires/cables as a medium to transfer the data
- Cables
  - Coaxial cable
  - Twisted pairs cables
  - Fiber optics

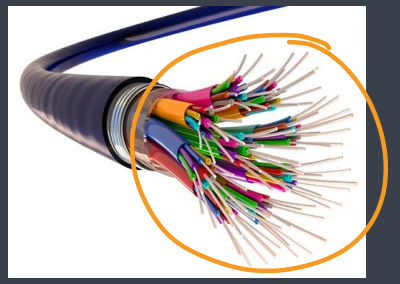
Cat → Category cable



UTP



STP



fo

Wireless network

EM waves

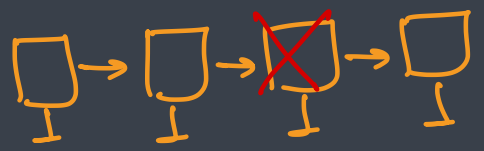
- The network build by connecting devices together using air as a medium to transfer the data
- EM Waves are used to transfer data from sender to receiver

Network Types

- Personal Area Network (wireless)
  - Smallest network which is very personal to the user (meters)
  - E.g. BlueTooth, NFC
- Local Area Network → LAN, WLAN
  - Spans across building(s) and operated under single administrative system
  - E.g. company, school network
  - Technologies: ~~TokenRing~~ or Ethernet ✓
- Metropolitan Area Network
  - Spans across cities
  - E.g. cable network
  - Technologies: high speed fiber optics
- Wide Area Network → (Internet)
  - Spans across countries
  - Technologies: ATM, Frame Relay

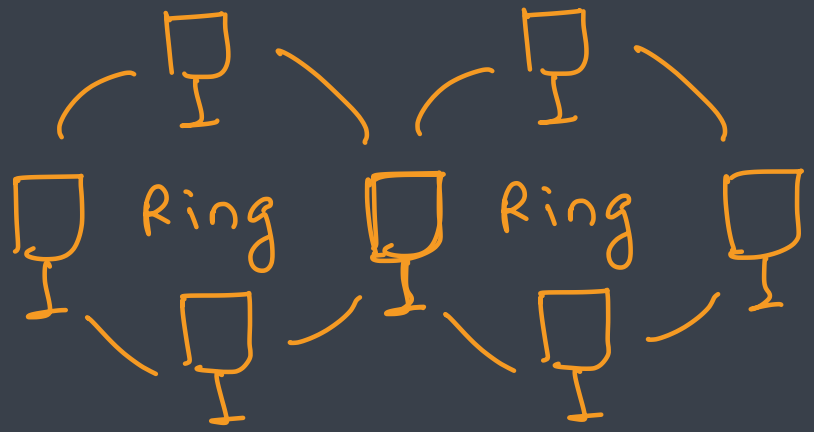
What is a network topology ?

- Physical arrangement of computers is known as topology
- Famous topologies
  - Bus
  - Ring
  - Token Ring
  - Star
  - Mesh



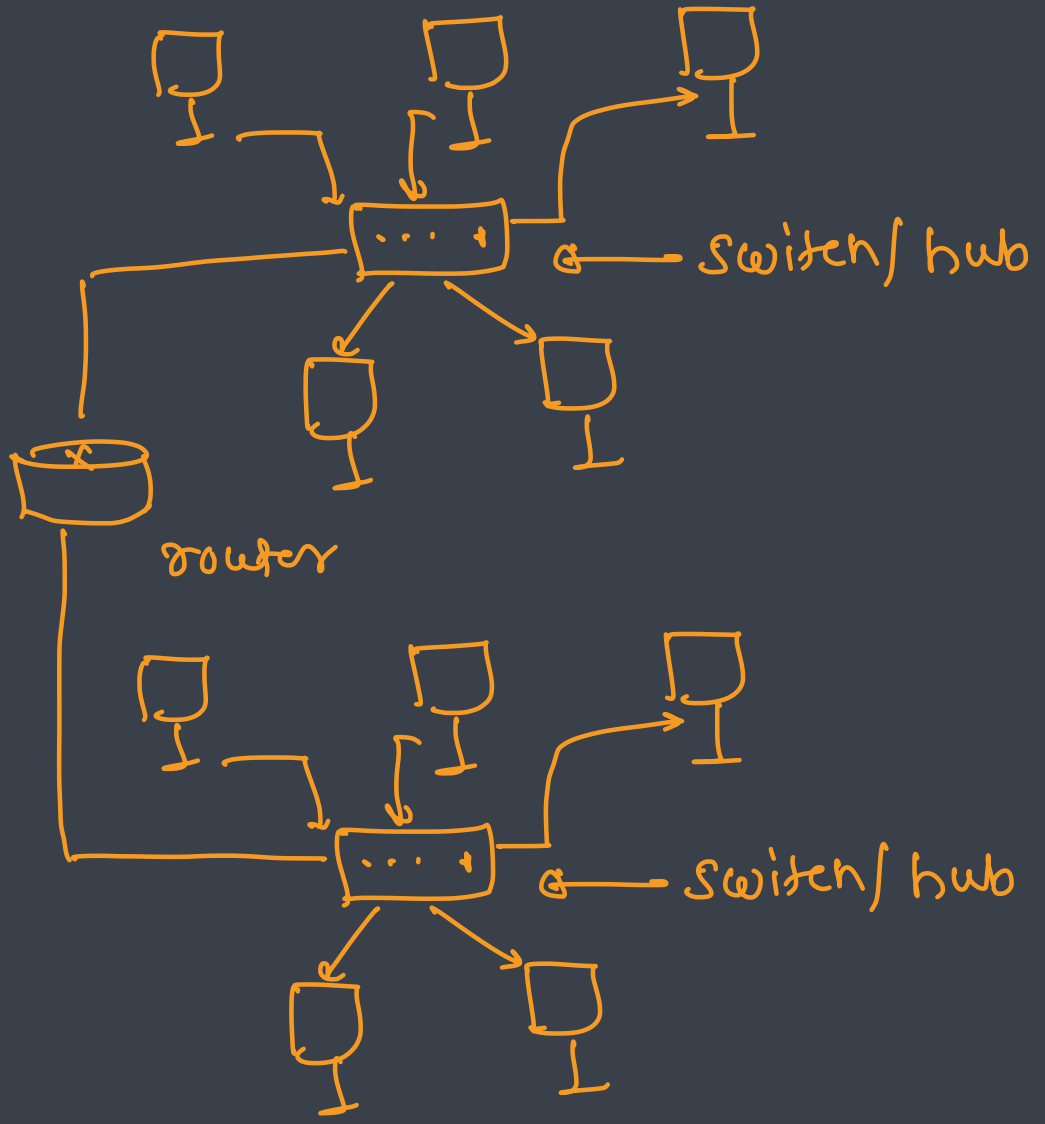
Bus

Mesh



connected Ring

Star



ISO OSI model

- Conceptual model that characterizes and standardizes the communication functions of a telecommunication or computing system without regard to its underlying internal structure and technology
- Goal is the interoperability of diverse communication systems with standard communication protocols
- Layered architecture having 7 layers
  - Application → application layer PDU
  - Presentation → presentation layer PDU
  - Session → session layer PDU
  - Transport → segment
  - Network → packet
  - Data Link → frame
  - Physical → binary

PDU → protocol Data Unit

program → process
↳ process control block (PCB)
→ pid
→ port
↳ test ...

Server

→ software / app which serves

→ types

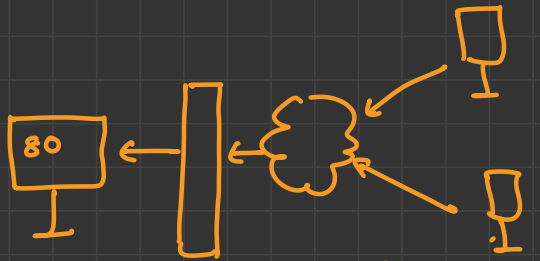
- web server → apache, nginx, flask
- db server
- file server

port - number used to communicate with the server process

→ standard → 1 - 1024

→ Ephemeral
→ > 1024
↳ random

↳ Reserved



firewall (SG)

80 - X

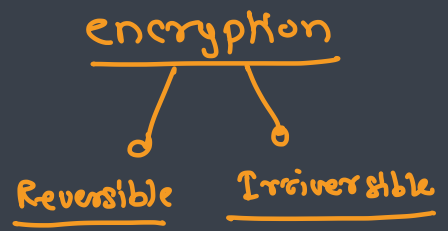
Application Layer

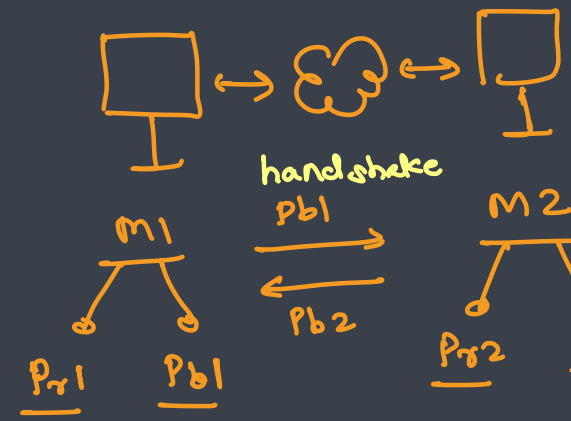
- Specifies interface methods used by hosts in a communications network
- Contains communication protocols
  - **HTTP [80]**: Hyper Text Transfer Protocol
  - **HTTPs [443]**: Secure Hyper Text Transfer Protocol
  - **FTP [20, 21]**: File Transfer Protocol
  - **SFTP [115]**: Simple FTP
  - **DNS [53]**: Domain Name Service
  - **NFS [1023]**: Network File System
  - **POP3 [110]**: Post Office Protocol
  - **SMTP [25]**: Simple Mail Transfer Protocol
  - **SSH [22]**: Secure Shell
  - **LDAP [389]**: Lightweight Directory Access Protocol

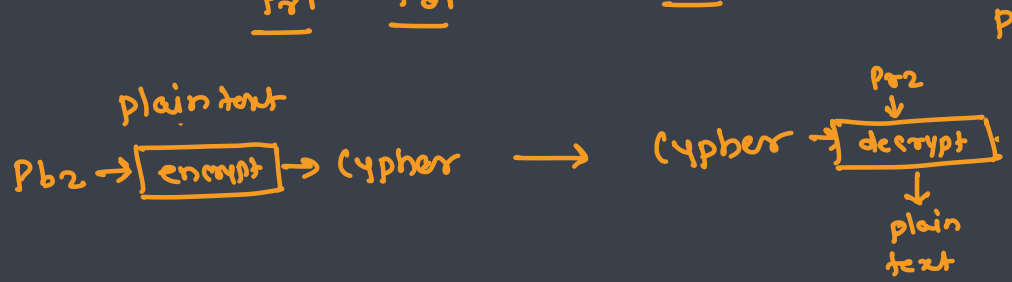
protocol → set of rules

(plain)

(encrypt) → digital certificate → private key / public key







Presentation Layer

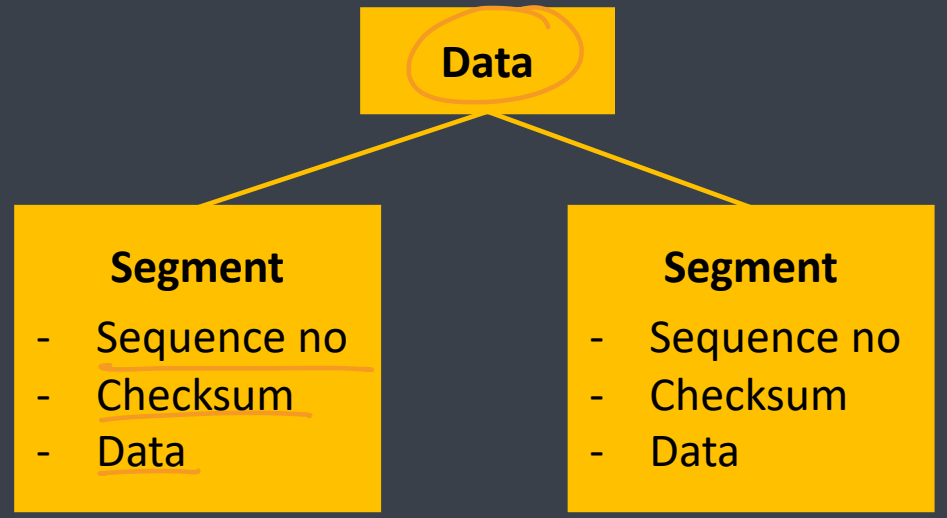
- Serves as the data translator for the network
- Also known as syntax layer
- Responsible for
  - Translation
  - Compression/Decompression
  - Encoding/Decoding
  - Encryption/Decryption

Session Layer

- Provides mechanism for opening, closing and managing session between processes
- Communication sessions consist of requests and responses that occur between applications
- Protocols
  - **ASP**: AppleTalk Session Protocol
  - **ADSP**: AppleTalk Data Stream Protocol
  - **NetBIOS**: Network BIOS
  - **PAP**: Password Authentication Protocol
  - **PPTP**: Point to Point Tunnelling Protocol
  - **RPC**: Remote Procedure Call
  - **SCP**: Session Control Protocol
  - **SDP**: Socket Direct Protocol

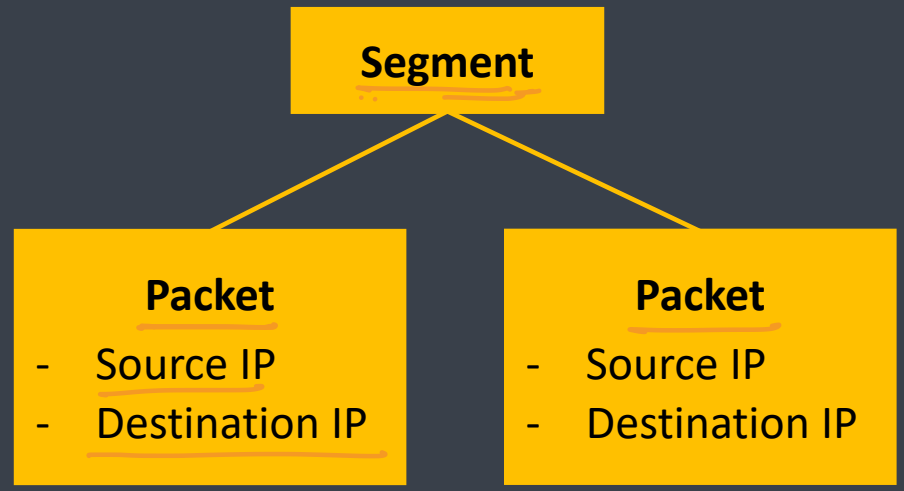
Transport Layer

- Provide host-to-host communication services for applications
- Creates Segment (data unit) containing
  - Sequence number
  - Checksum
  - Port number
- Protocols
  - TCP
    - Connection oriented protocol
    - Provides: Flow Control, Error checking
    - Guarantees data delivery
    - Slower than UDP
    - E.g. WWW, HTTP
  - UDP
    - Connectionless protocol
    - Does not provide flow control
    - Does not guarantee data delivery
    - Faster than TCP
    - E.g. streaming, online games



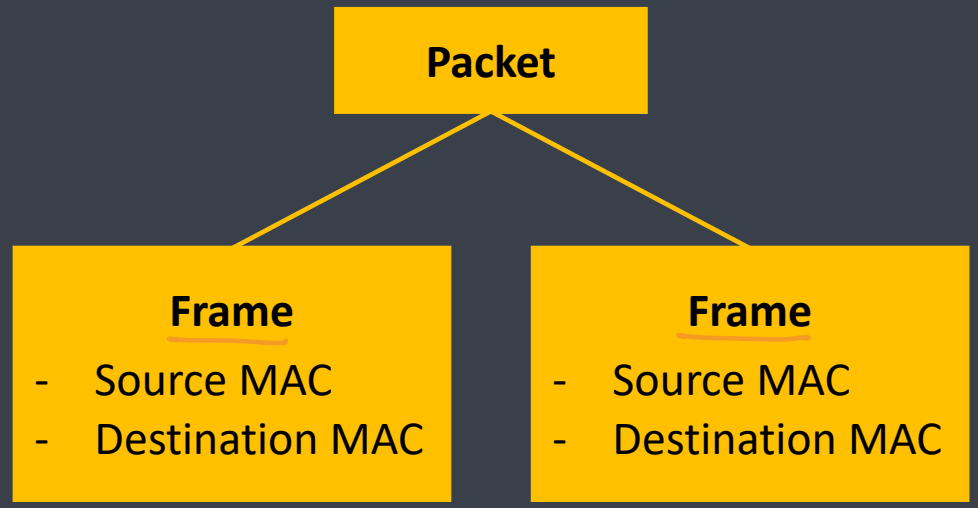
Network Layer

- Responsible for packet forwarding including routing through intermediate routers
- Responsible for splitting segment into packets containing
  - Source IP address
  - Destination IP address
- Protocols
  - **IP**: Internet Protocol
  - **IPX**: Internetwork Packet Exchange
  - **IPSec**: Internet Protocol Security
  - **EGP**: Exterior Gateway Protocol



Data Link Layer

- Transfers data between
  - adjacent network nodes in a wide area network (WAN) or
  - between nodes on the same local area network (LAN) segment
- Encapsulates packet into Frames containing
  - Source MAC Address
  - Destination MAC Address
- Sublayers
  - Logical Link Layer
  - Media Access Control Layer



Data Link Layer: Logical Link Layer

- The uppermost sublayer multiplexes protocols running at the top of data link layer, and optionally provides flow control, acknowledgment, and error notification
- Provides addressing and control of the data link
- Services
  - Error control (automatic repeat request, ARQ)
  - Flow control [Data-link-layer flow control is not used in LAN protocols such as Ethernet, but in modems and wireless networks]

Data Link Layer: Media Access Control Layer

- Refers to the sublayer that determines who is allowed to access the media at any one time (CSMA/CD)
- Determines where one frame of data ends and the next one starts (frame synchronization)
- Frame synchronization uses: time based, character counting, byte stuffing and bit stuffing.
- Services
  - Multiple access protocols for channel-access control,
    - CSMA/CD protocols for collision detection and re-transmission in Ethernet networks
    - CSMA/CA protocol for collision avoidance in wireless networks
  - Physical addressing (MAC addressing)
  - LAN switching (packet switching), including MAC filtering, Spanning Tree Protocol (STP) and Shortest Path Bridging (SPB)
  - Data packet queuing or scheduling

Physical Layer

- Consists of the electronic circuit transmission technologies of a network
- Fundamental layer underlying the higher level functions in a network which provides means of transmitting raw bits rather than logical packets or segments
- The bitstream may be grouped into code words or symbols and converted to a physical signal that is transmitted over a transmission medium
- Translates logical communications requests from the data link layer into hardware-specific operations to cause transmission or reception of electronic signals
- Services
  - Modulation/Demodulation
  - Multiplexing
- Consists of
  - Cables/wires
  - Devices like hub, repeaters etc.

Addressing Modes: MAC Address

physical address

- Used to identify NIC uniquely
- Consists of 6 bytes [48 bits]
- First 3 bytes represents manufacturer
- Next 3 bytes represents NIC's unique address

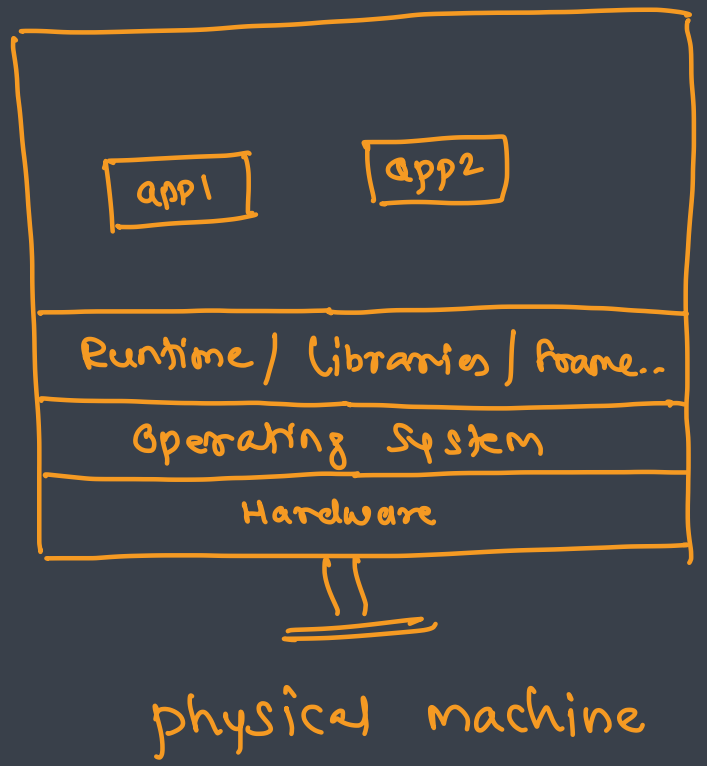
Addressing Modes: IP Address

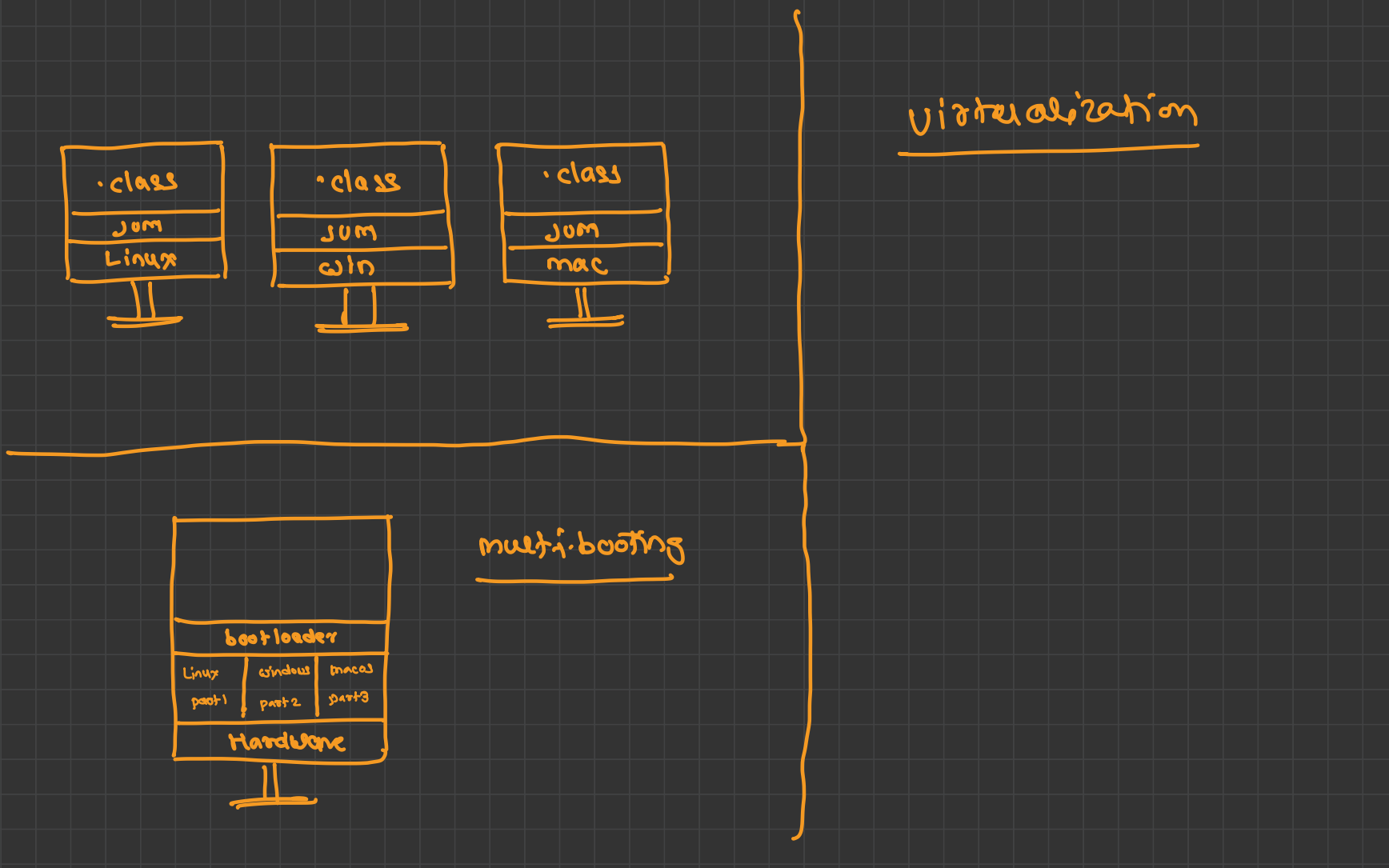
- Used to identify every device uniquely
- Set by operating system running on the device
- Can be written in
  - Decimal: 192.168.100.10
  - Binary: 11000000.10101000.01100100.00001010
- Versions
  - IPv4
    - 32 bit [4 bytes] address
    - Classful and Classless addressing
  - IPv6
    - 128 bit address
- Types
  - Private: used to communicate with other devices in local network
  - Public: used to communicate with other devices over internet

Virtualization

Traditional Deployment (physical machines)

- Early on, organizations ran applications on physical servers
- There was no way to define resource boundaries for applications in a physical server, and this caused resource allocation issues
- For example, if multiple applications run on a physical server, there can be instances where one application would take up most of the resources, and as a result, the other applications would underperform
- A solution for this would be to run each application on a different physical server
- But this did not scale as resources were underutilized, and it was expensive for organizations to maintain many physical servers




.class
JVM
Linux
.class
JVM
win
.class
JVM
mac
virtualization
bootloader
Linux
part1
window
part2
macos
part3
Hardware
multi-booting

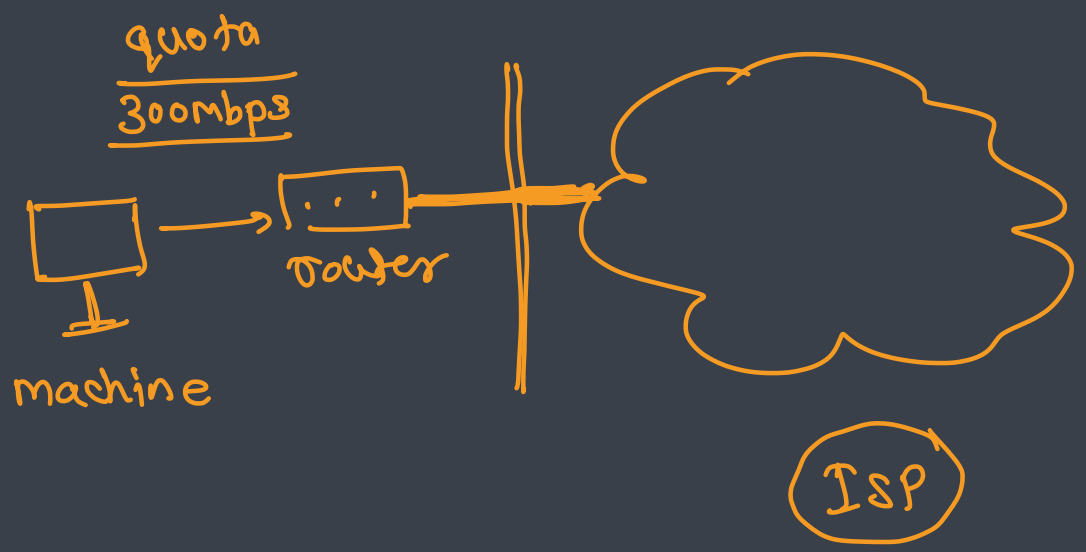
What is virtualization

- Virtualization is the creation of a virtual -- rather than actual -- version of something, such as an operating system (OS), a server, a storage device or network resources
- Virtualization uses software that simulates hardware functionality in order to create a virtual system
- This practice allows IT organizations to operate multiple operating systems, more than one virtual system and various applications on a single server
- Types
  1. Network virtualization
  2. Storage virtualization
  3. Data virtualization
  4. Desktop virtualization
  5. Application virtualization
  6. Hardware virtualization
  7. OS virtualization

     Containerization

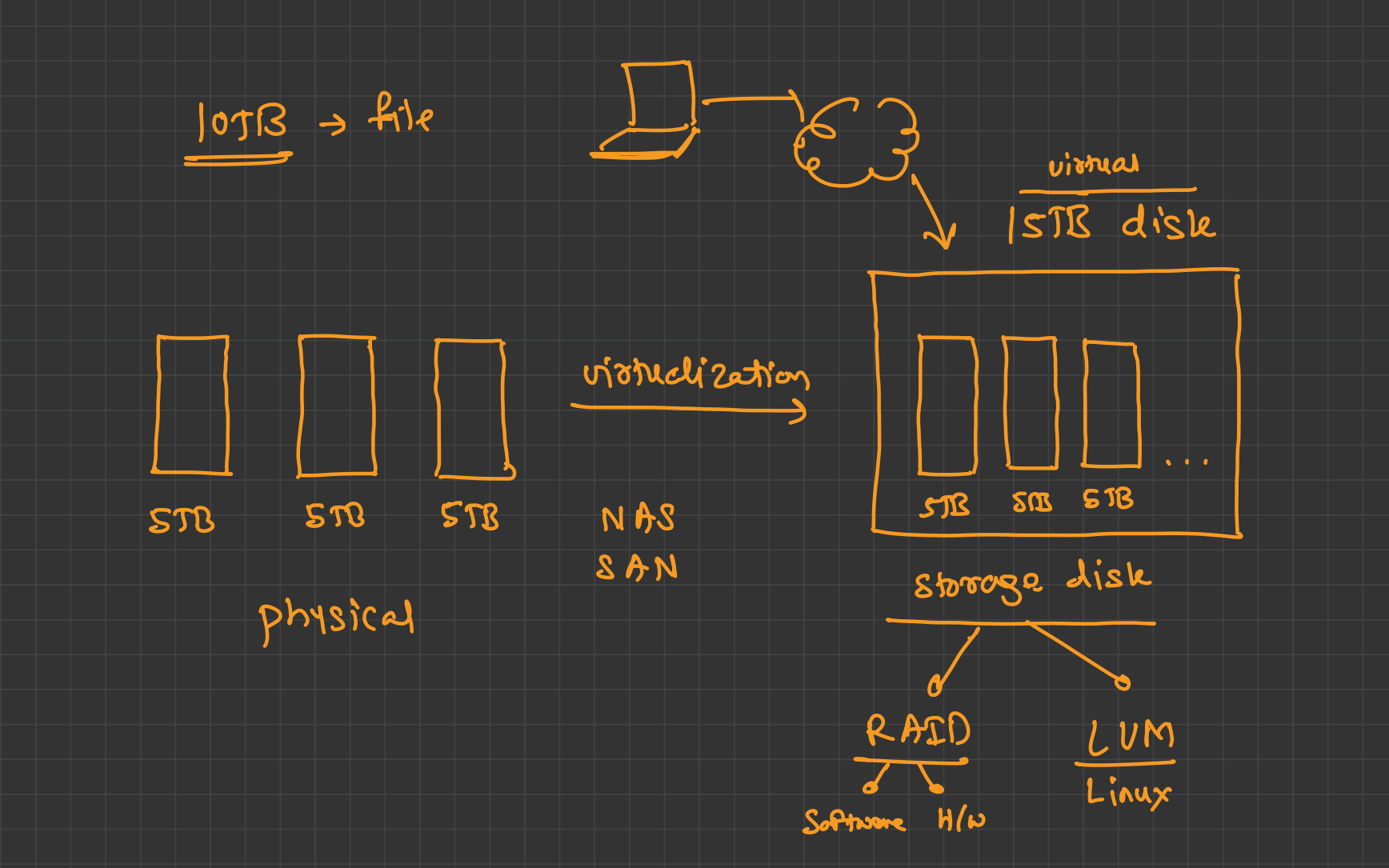
Network Virtualization

- Network virtualization takes the available resources on a network and breaks the bandwidth into discrete channels
- Admins can secure each channel separately, and they can assign and reassign channels to specific devices in real time
- The promise of network virtualization is to improve networks' speed, availability and security, and it's particularly useful for networks that must support unpredictable usage bursts



Storage Virtualization

- Storage virtualization is the pooling of physical storage from multiple network storage devices into what appears to be a single storage device that is managed from a central console
- Storage virtualization is commonly used in storage area networks
- Applications can use storage without having any concern for where it resides, what technical interface it provides, how it has been implemented, which platform it uses and how much of it is available
- Benefits
  - Makes the remote storage devices appear local
  - Multiple smaller volumes appear as a single large volume
  - Data is spread over multiple physical disks to improve reliability and performance
  - All operating systems use the same storage device
  - Provided high availability, disaster recovery, improved performance and sharing


10TB → file
virtual
1STB disk
virtualization
5TB
5TB
5TB
...
5TB
5TB
5TB
NAS
SAN
physical
storage disk
RAID
LVM
Linux
Software
H/w

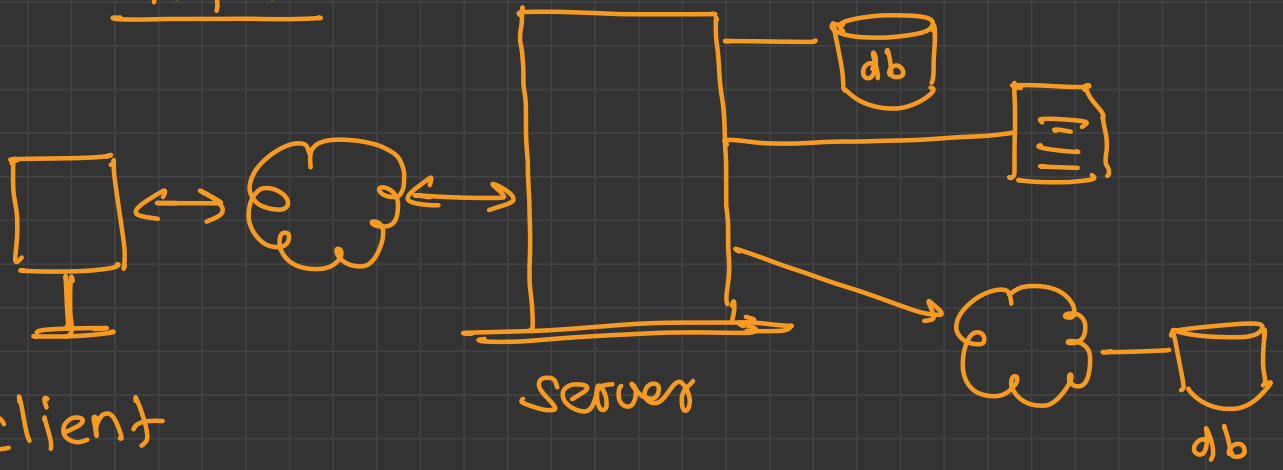
Data virtualization

- Data virtualization is the process of aggregating data from different sources of information to develop a single, logical and virtual view of information so that it can be accessed by front-end solutions such as applications, dashboards and portals without having to know the data's exact storage location
- The process of data virtualization involves abstracting, transforming, federating and delivering data from disparate sources
- The main goal of data virtualization technology is to provide a single point of access to the data by aggregating it from a wide range of data sources
- Benefits
  - Abstraction of technical aspects of stored data like APIs, Language, Location, Storage structure
  - Provides an ability to connect multiple data sources from a single location
  - Provides an ability to combine the data result sets across multiple sources (also known as data federation)
  - Provides an ability to deliver the data as requested by users

RPC → protocol

REST
GraphQL
} design pattern

Data Aggregation
federation

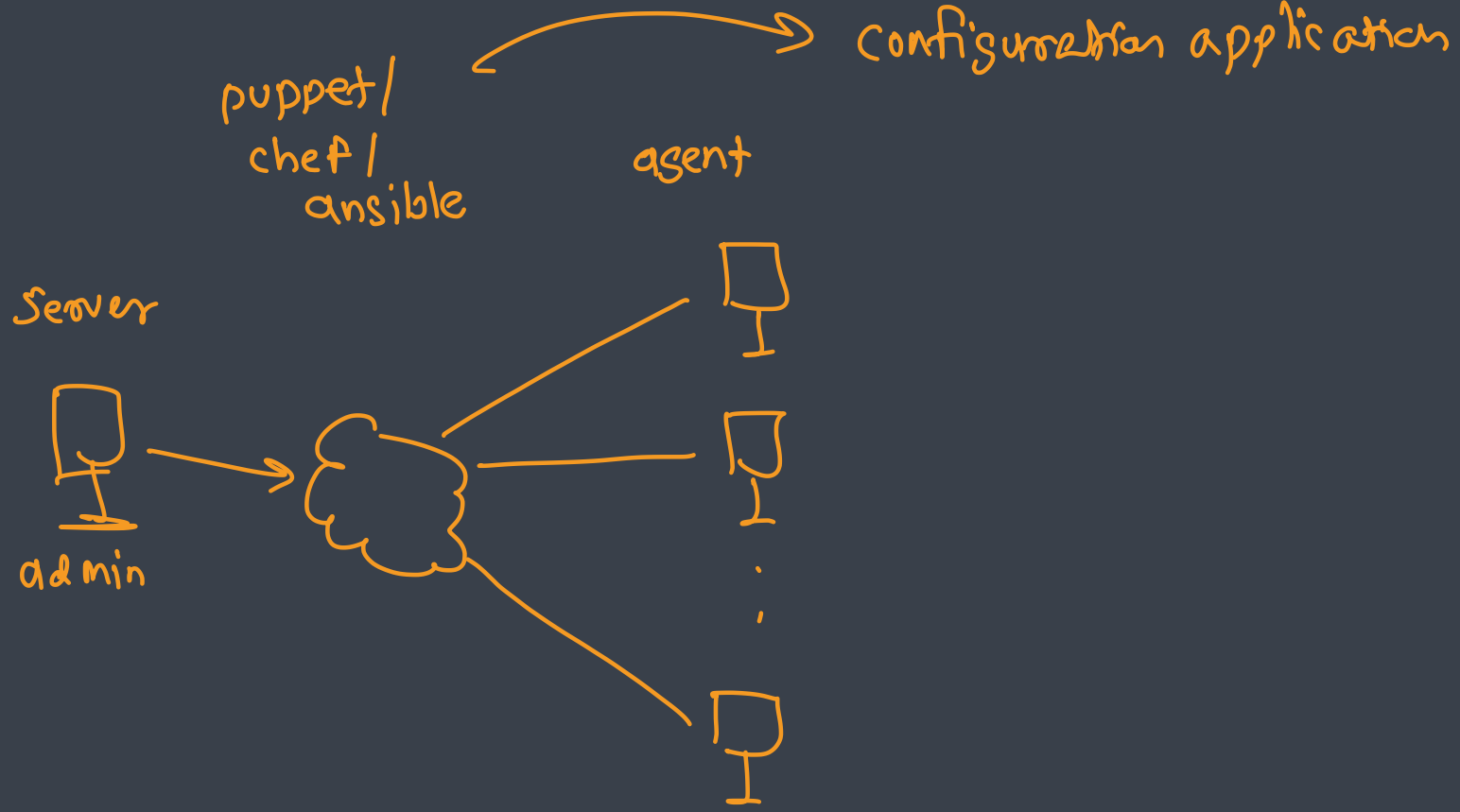


Desktop virtualization

- With desktop virtualization, the goal is to isolate a desktop OS from the endpoint that employees use to access it
- It provides an ability to connect to the desktop from remote site
- When multiple users connect to a shared desktop, as is the case with Microsoft Remote Desktop Services, it's known as shared hosted desktop virtualization

Application Virtualization

- With application virtualization, an app runs separately from the device that accesses it
- Application virtualization makes it possible for IT admins to install, patch and update only one version of an app rather than performing the same management tasks multiple times

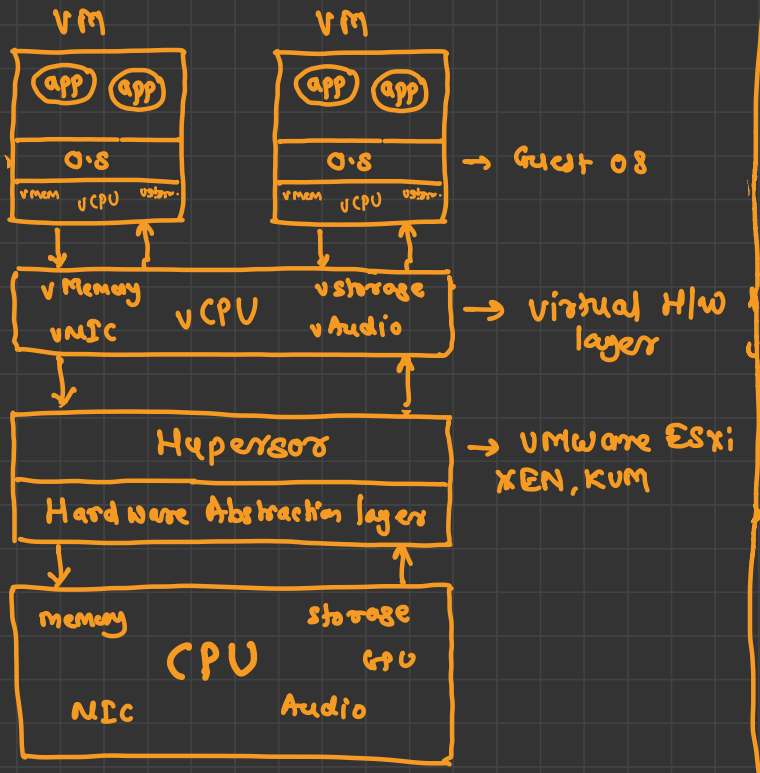


Hardware Virtualization

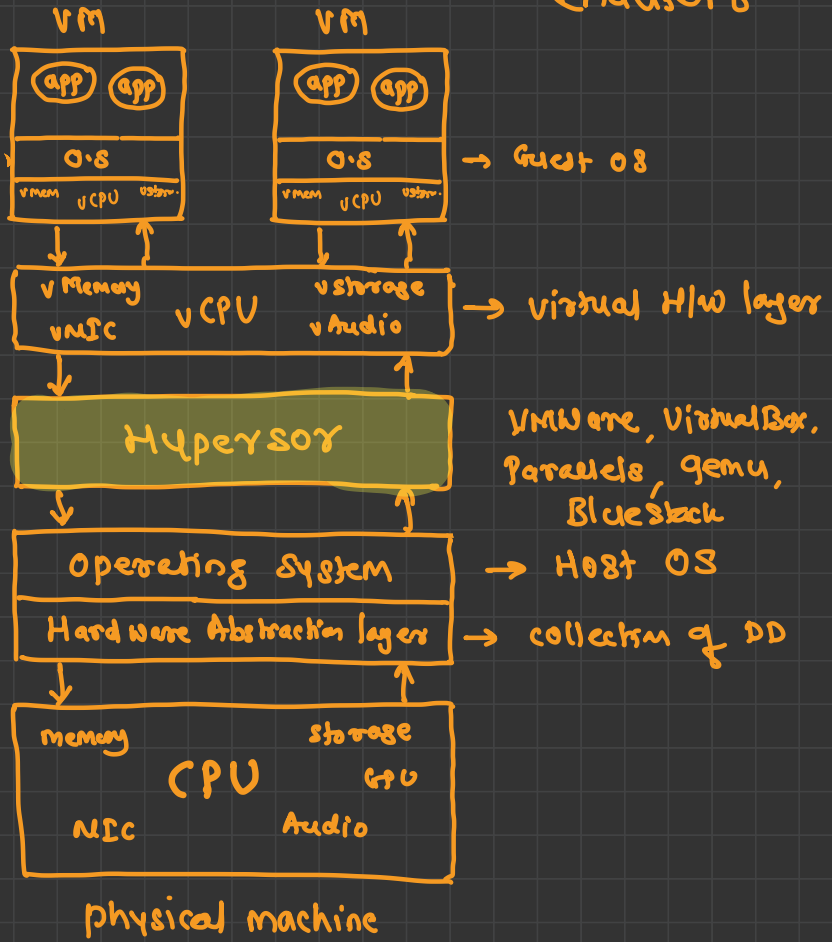
- Hardware virtualization or platform virtualization refers to the creation of a virtual machine that acts like a real computer with an operating system
- The process of masking the hardware resources like
  - CPU
  - Storage
  - Memory
- For example, a computer that is running Microsoft Windows may host a virtual machine that looks like a computer with the Ubuntu Linux operating system; Ubuntu-based software can be run on the virtual machine
- The process of creating Machines

Type I - (Bare metal)

↳ cloud providers



Type II - (Hosted) → testers / devs / endusers



Virtual Machine

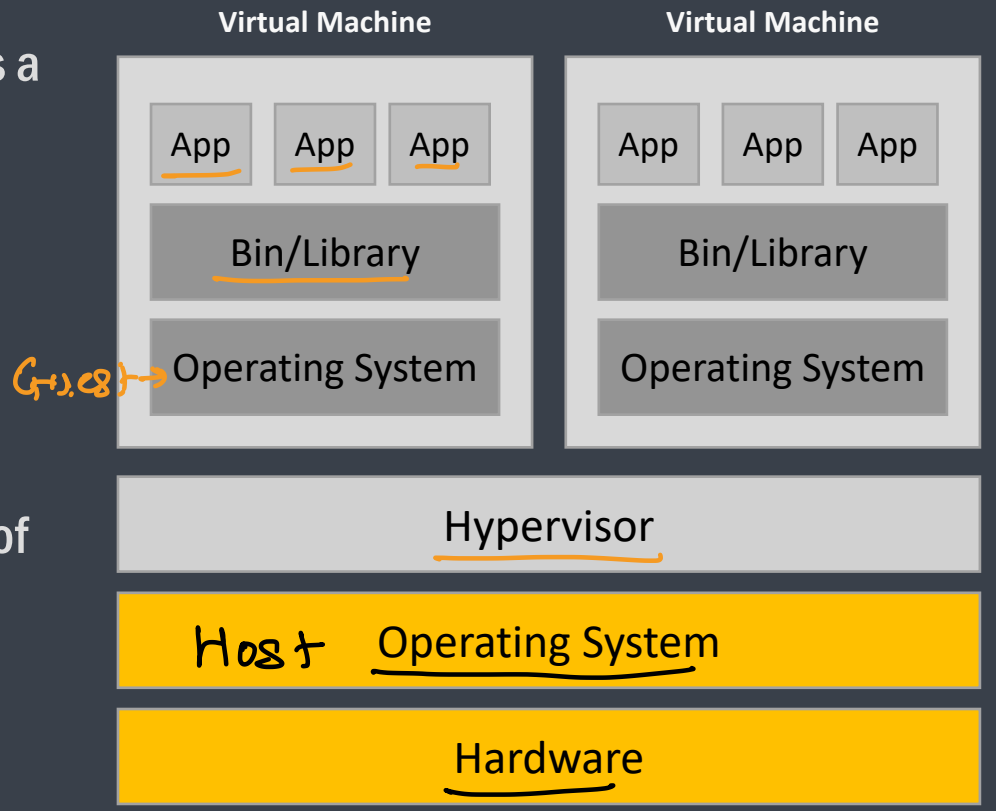
- A virtual machine is the emulated equivalent of a computer system that runs on top of another system
- Virtual machines may have access to any number of resources
  - Computing power - through hardware-assisted but limited access to the host machine's CPU
  - Memory - one or more physical or virtual disk devices for storage
  - A virtual or real network interfaces
  - Any devices such as
    - video cards,
    - USB devices,
    - other hardware that are shared with the virtual machine
- If the virtual machine is stored on a virtual disk, this is often referred to as a disk image

Types of hardware virtualization

- Type I
  - A Type 1 hypervisor runs directly on the host machine's physical hardware, and it's referred to as a bare-metal hypervisor
  - It doesn't have to load an underlying OS first
  - With direct access to the underlying hardware and no other software, it is more efficient and provides better performance
  - It is best suited for enterprise computing or data centers
  - E.g. VMware ESXi, Microsoft Hyper-V server and open source KVM
- Type II
  - A Type 2 hypervisor is typically installed on top of an existing OS, and it's called a hosted hypervisor
  - It relies on the host machine's pre-existing OS to manage calls to CPU, memory, storage and network resources
  - E.g. VMware Fusion, Oracle VM VirtualBox, Oracle VM Server for x86, Oracle Solaris Zones, Parallels and VMware Workstation

Virtualized Deployment

- It allows you to run multiple Virtual Machines (VMs) on a single physical server's CPU
- Virtualization allows applications to be isolated between VMs and provides a level of security as the information of one application cannot be freely accessed by another application
- Virtualization allows better utilization of resources in a physical server and allows better scalability because
  - an application can be added or updated easily
  - reduces hardware costs
- With virtualization you can present a set of physical resources as a cluster of disposable virtual machines
- Each VM is a full machine running all the components, including its own operating system, on top of the virtualized hardware



Advantages of virtualization

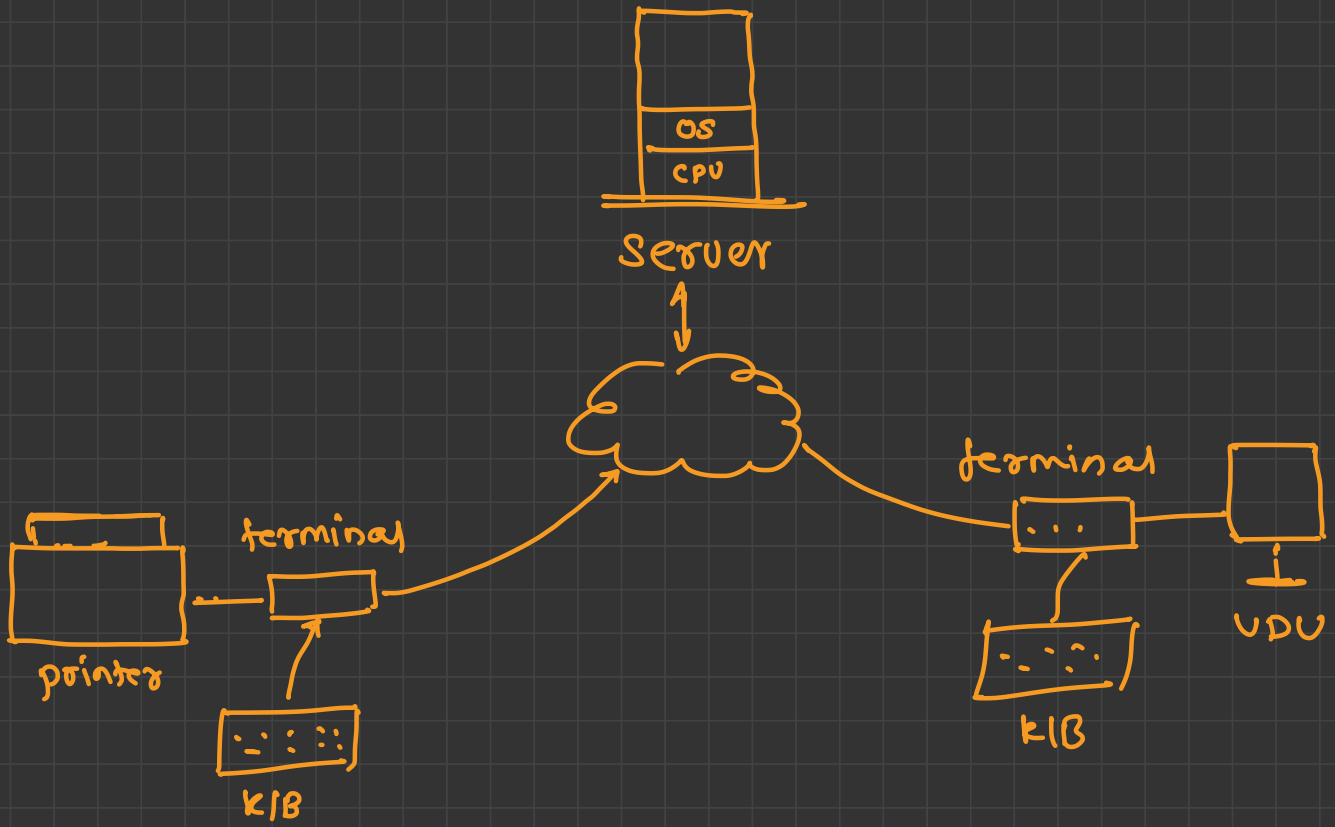
- **Lower costs**
  - Virtualization reduces the amount of hardware servers necessary within a company and data center
  - This lowers the overall cost of buying and maintaining large amounts of hardware
- **Easier disaster recovery**
  - Disaster recovery is very simple in a virtualized environment
  - Regular snapshots provide up-to-date data, allowing virtual machines to be feasibly backed up and recovered
  - Even in an emergency, a virtual machine can be migrated to a new location within minutes
- **Easier testing**
  - Testing is less complicated in a virtual environment
  - Even if a large mistake is made, the test does not need to stop and go back to the beginning
  - It can simply return to the previous snapshot and proceed with the test
- **Quicker backups**
  - Backups can be taken of both the virtual server and the virtual machine
  - Automatic snapshots are taken throughout the day to guarantee that all data is up-to-date
  - Furthermore, the virtual machines can be easily migrated between each other and efficiently redeployed
- **Improved productivity**
  - Fewer physical resources results in less time spent managing and maintaining the servers
  - Tasks that can take days or weeks in a physical environment can be done in minutes
  - This allows staff members to spend majority of their time on more productive tasks, like raising revenue and fostering business initiatives

Cloud

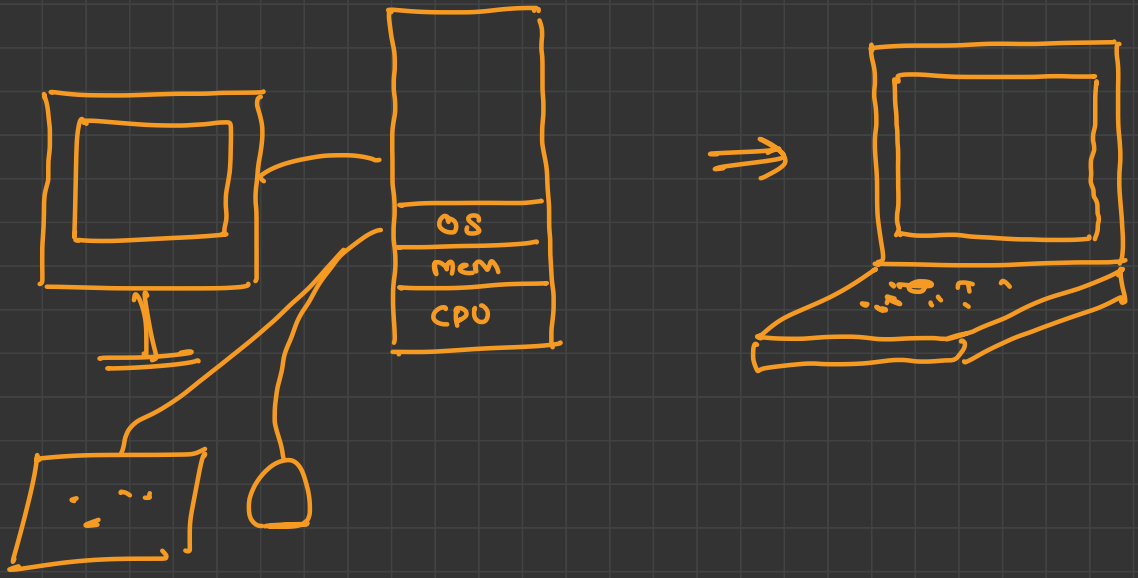
Computing Model

- Desktop computing
- Client-Server computing
- Cluster computing
- Cloud Computing

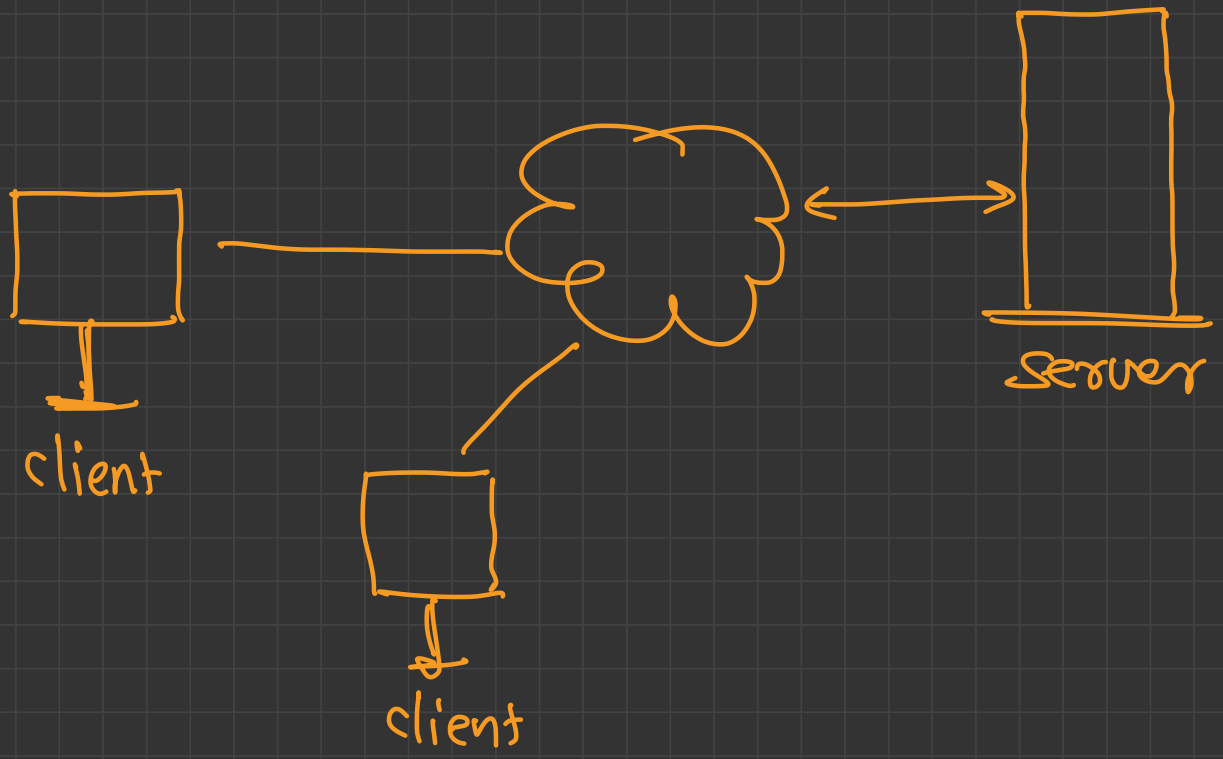
0] No-computing → pre-computing → pre-CPU



1) Desktop Computing

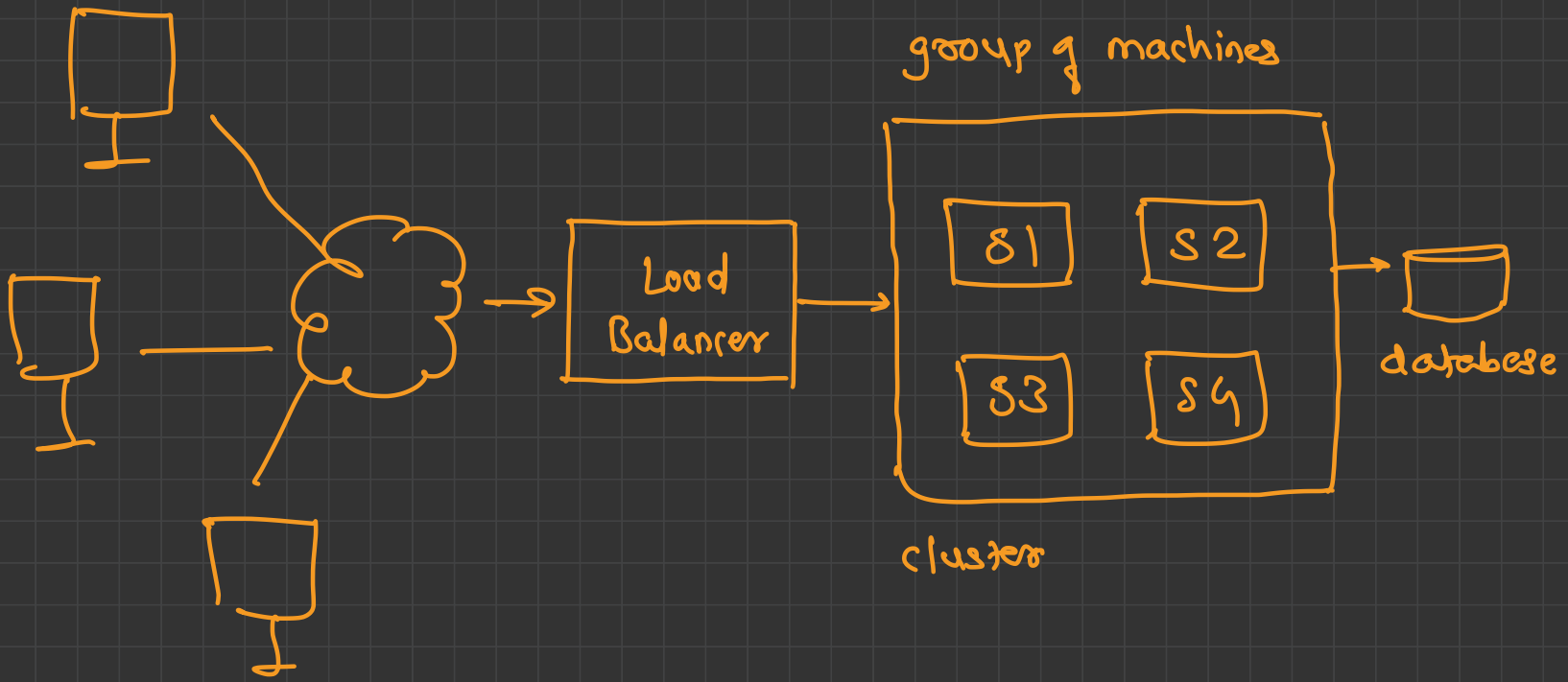


2] client-server computing



3] cluster computing

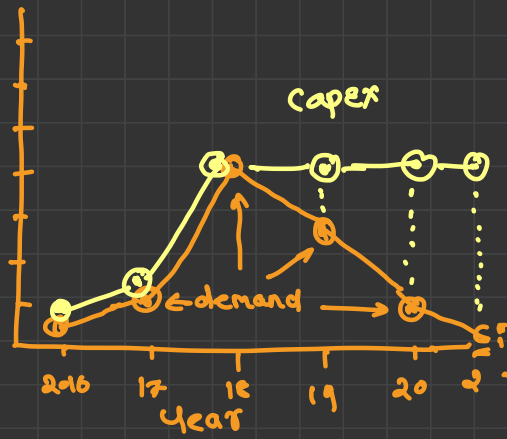
↳ Horizontal Scaling



Capex
Capital Expenditure

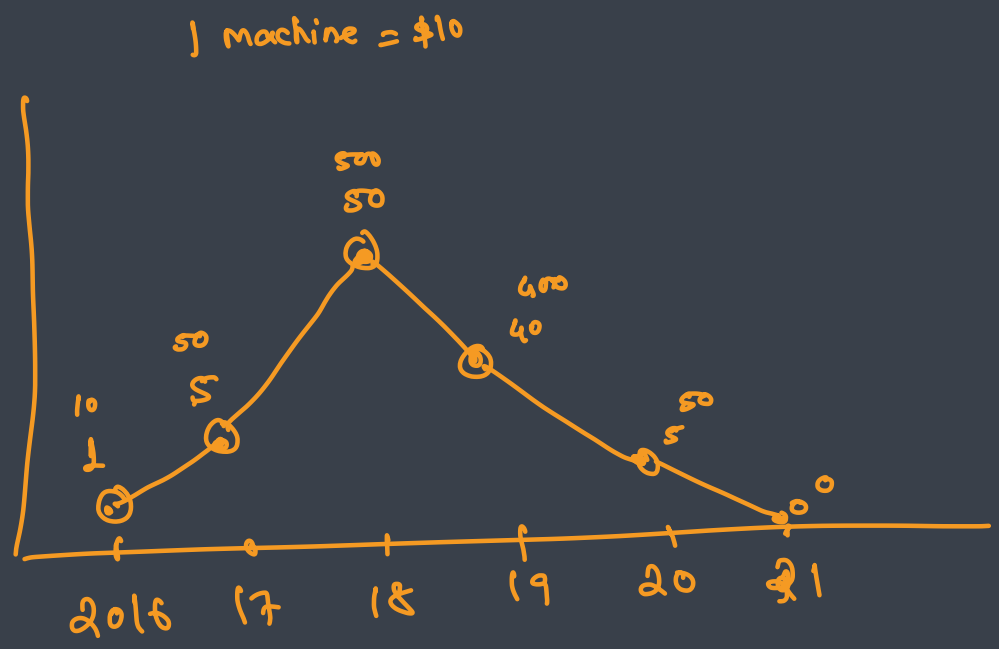
Opex
operational expenditure

		invested ↓		↓	actual
2016 →	1000 →	1 machine →	2L	10K	1
2017 →	10000 →	10 machines →	18L	10K	10
2018 →	50,000 →	50 machines →	80L	10K	50
2019 →	40,000 →	50 machines →	1cr	10K	40
2020 →	5000 →	50 machines →	1cr	10K	5
2021 →	↓ 0 →	50 machines →	1cr	-	0



What is cloud computing ?

- The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
- Is the delivery of on-demand computing resources – everything from data centers over the internet on a pay for use basis
- Cloud computing is an umbrella term used to refer to Internet based development and services

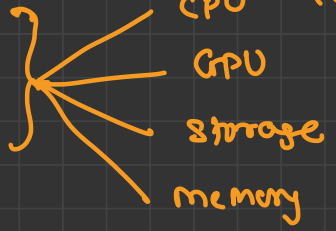


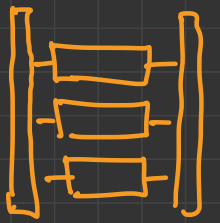
What is Data Center ?

- Where your IT devices and applications are located
- For a non-technical person it is the cloud where the user's files/data is stored
- Components
  - Servers
  - Security
  - WAN
  - Storage
  - File Sharing

servers

- rack server
- tower server





1 rack server = 100 TB

1 rack = 10 rack = 10 x 100 TB = 1 PB

1 rack tow = 20 racks = 20 x 1 PB = 20 PB

1 room = 10 rows = 10 x 20 = 200 PB

1 floor = 10 rooms = 10 x 200 PB = 2 EB

1 building = 10 floors = 10 x 2 = 20 EB

1 data center = 5 building = 20 x 5 = 100 EB

1 AZ = 3 DC = 100 x 3 = 300 EB

1 Region = 3 AZ = 300 x 3 = 900 EB

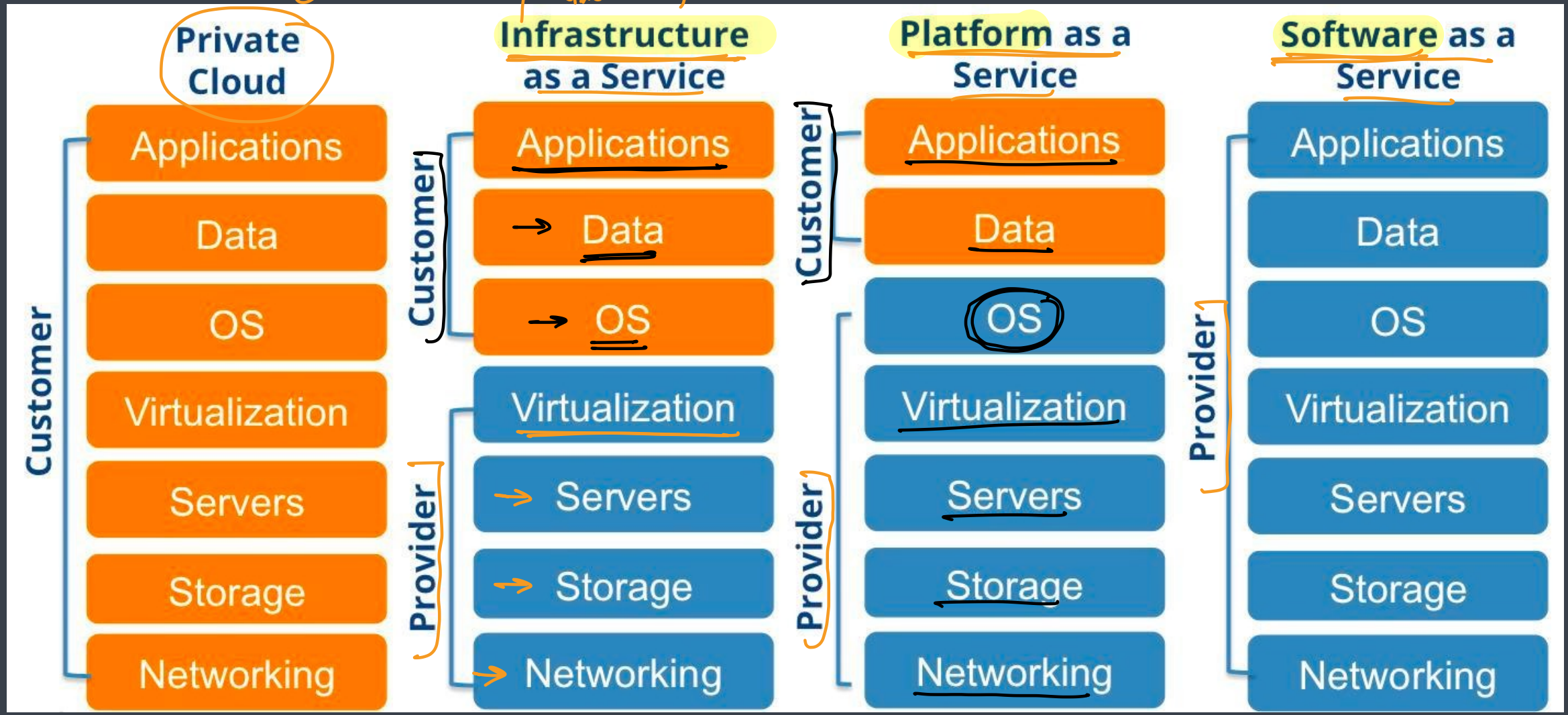
cloud provider: 30 regions = 900 x 30 = 27 GB

∞ resources

Terminologies

- **Scalability**
  - refers to the idea of a system in which every application or piece of infrastructure can be expanded to handle increased load
- **Elasticity**
  - the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible
- **Availability**
  - refers to the ability of a user to access information or resources in a specified location and in the correct format
- **Information Assurance**
  - availability, integrity, authentication, confidentiality and nonrepudiation
- **On-demand service**
  - A model by which a customer can purchase cloud services as needed

Service Models



Service Models

End users

Software as a Service (SaaS)

Google Apps, Facebook, YouTube, Dropbox, Google Photos

developers

Platform as a Service (PaaS)

AWS Elastic Bean, Google App Engine, Amazon Simple DB, S3, Microsoft Azure

operations team

Infrastructure as a Service (IaaS)

Amazon EC2, Google Compute VM, Azure VM

Data Center provider

Service models

- → IaaS – Infrastructure – VM, Storage, VPC
- → PaaS – platform → Elastic Beans, S3
- → SaaS – software → netflix
- → DaaS → database → RDS → MySQL, SQL Server, Oracle etc
  NoSQL → OpenShift, Redis (cache), etc.
- → FaaS → function → Lamba (Python, JS, Java)
- → EaaS → Everything

Service Models: IaaS

- Infrastructure as a Service
- Allocates virtualized computing resources to the user through the internet
- IaaS is completely provisioned and managed over the internet
- helps the users to avoid the cost and complexity of purchasing and managing their own physical servers
- Every resource of IaaS is offered as an individual service component and the users only have to use the particular one they need
- The cloud service provider manages the IaaS infrastructure while the users can concentrate on installing, configuring and managing their software
- Generally meant for operations team to setup the required infrastructure
- Benefits
  - Time and cost savings: more installation and maintenance of IT hardware in-house,
  - Better flexibility: On-demand hardware resources that can be tailored to your needs,
  - Remote access and resource management.

Service Models: PaaS

- Provides a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app
- Generally meant for developers
- Benefits
  - Mastering the installation and development of software applications
  - Time saving and flexibility for development projects: no need to manage the implementation of the platform, instant production
  - Data security: You control the distribution, protection, and backup of your business data

Service Models: SaaS

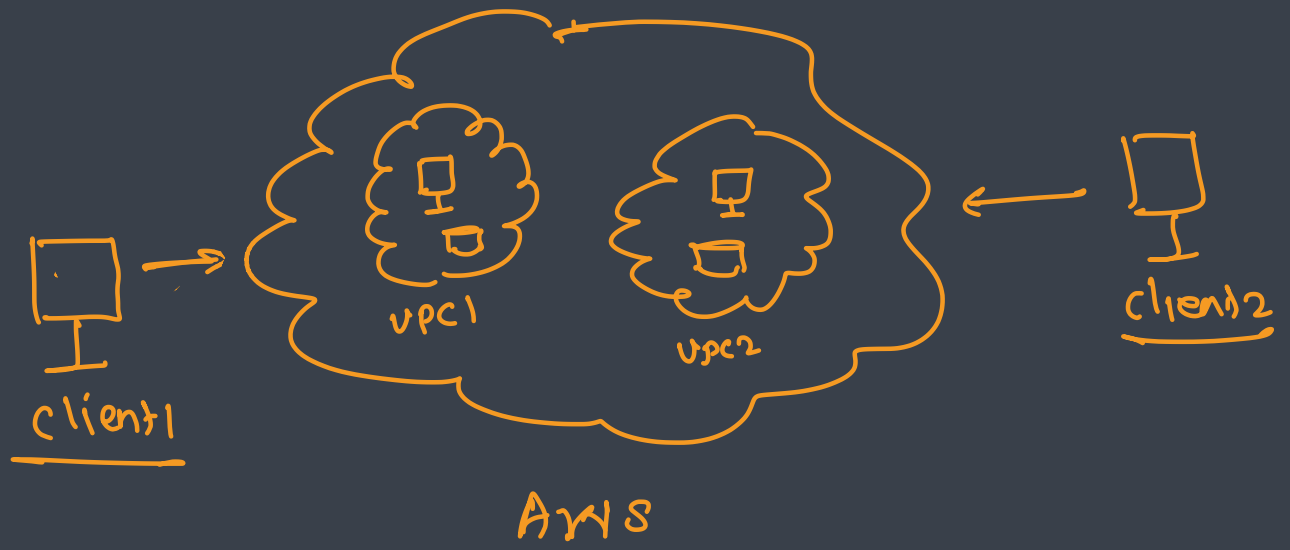
- Software as a Service
- Software distribution model in which a third-party provider hosts applications and makes them available to customers over the Internet
- User wont know which computer or operating system or infrastructure is used to host the software
- Generally meant for end user
- Benefits
  - You are entirely free from the infrastructure management and aligning software environment: no installation or software maintenance
  - You benefit from automatic updates with the guarantee that all users have the same software version
  - It enables easy and quicker testing of new software solutions.

Cloud Computing Characteristics

- Rapid Elasticity
- On Demand Self Service
- Broad Network Access
- Location Independent Resource Sharing
- Measured Services

Cloud Deployment Models: Public

- Supports all users who want to make use of a computing resource, such as hardware (OS, CPU, memory, storage) or software (application server, database) on a subscription basis
- Most common uses of public clouds are for application development and testing, tasks such as file-sharing, and e-mail service
- Requires internet to access the resources



Cloud Deployment Models: Private

- Typically infrastructure used by a single organization
- Such infrastructure may be managed by the organization itself to support various user groups, or it could be managed by a service provider that takes care of it either on-site or off-site
- Private clouds are more expensive than public clouds due to the capital expenditure involved in acquiring and maintaining them
- However, private clouds are better able to address the security and privacy concerns of organizations

Cloud Deployment Models: Hybrid

- Organization makes use of interconnected private and public cloud infrastructure
- Many organizations make use of this model when they need to scale up their IT infrastructure rapidly, such as when leveraging public clouds to supplement the capacity available within a private cloud
- For example, if an online retailer needs more computing resources to run its Web applications during the holiday season it may attain those resources via public clouds.

Cloud Services

- Compute: used to create the Virtual Machine (EC2)
- Storage: used to provide the storage
- Database: RDBMS + No SQL
- Security and Identity Management - Iam
- Media Services
- Machine Learning *
- Cost Management
- Application Integration - SQS, SNS, SES

Storage

- File storage - EFS
- block storage - EBS
- object storage - S3

Advantages

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

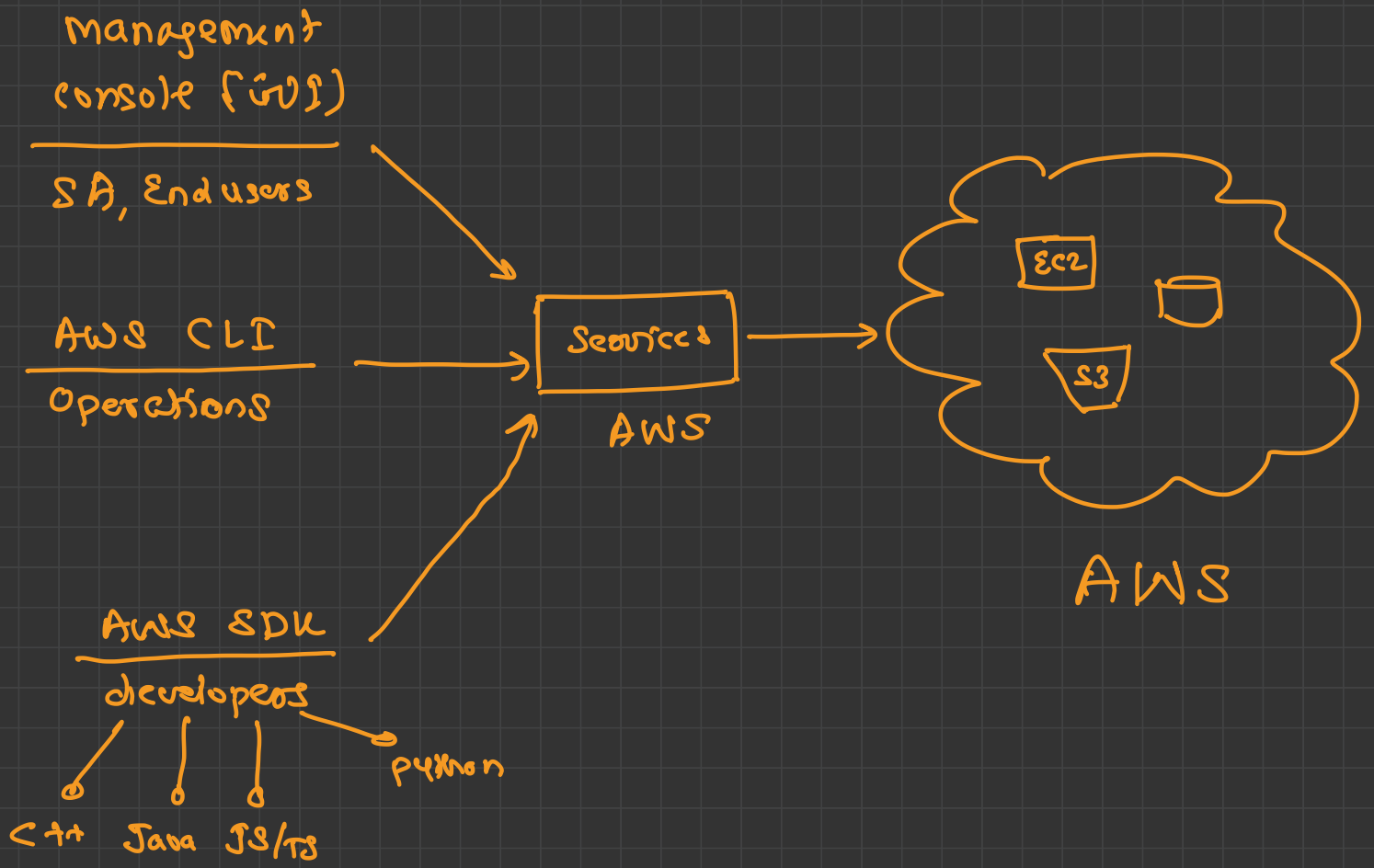
Disadvantages

- Requires a constant Internet connection
- Does not work well with low-speed connections
- Features might be limited
- Stored data might not be secure
- Stored data can be lost
- Each cloud systems uses different protocols and different APIs

Cloud Providers

- Amazon Web Services
- Google Cloud Platform
- Microsoft Azure
- Rackspace
- DigitalOcean
- Alibaba Cloud
- Oracle Cloud
- IBM Cloud

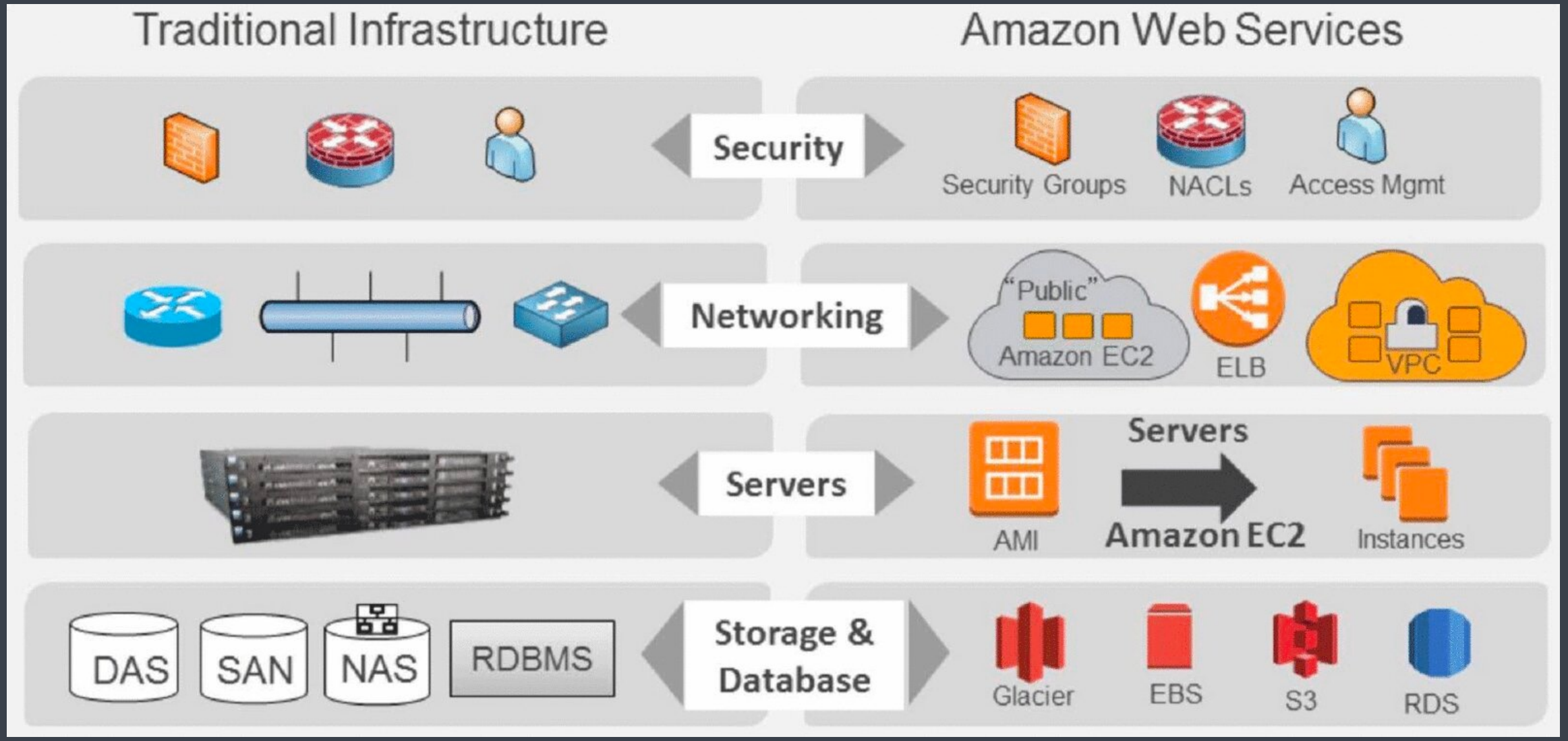
AWS


Management
console (GUI)
SA, End users
AWS CLI
Operations
AWS SDK
developers
C++
Java
JS/TS
python
Services
AWS
EC2
S3
AWS

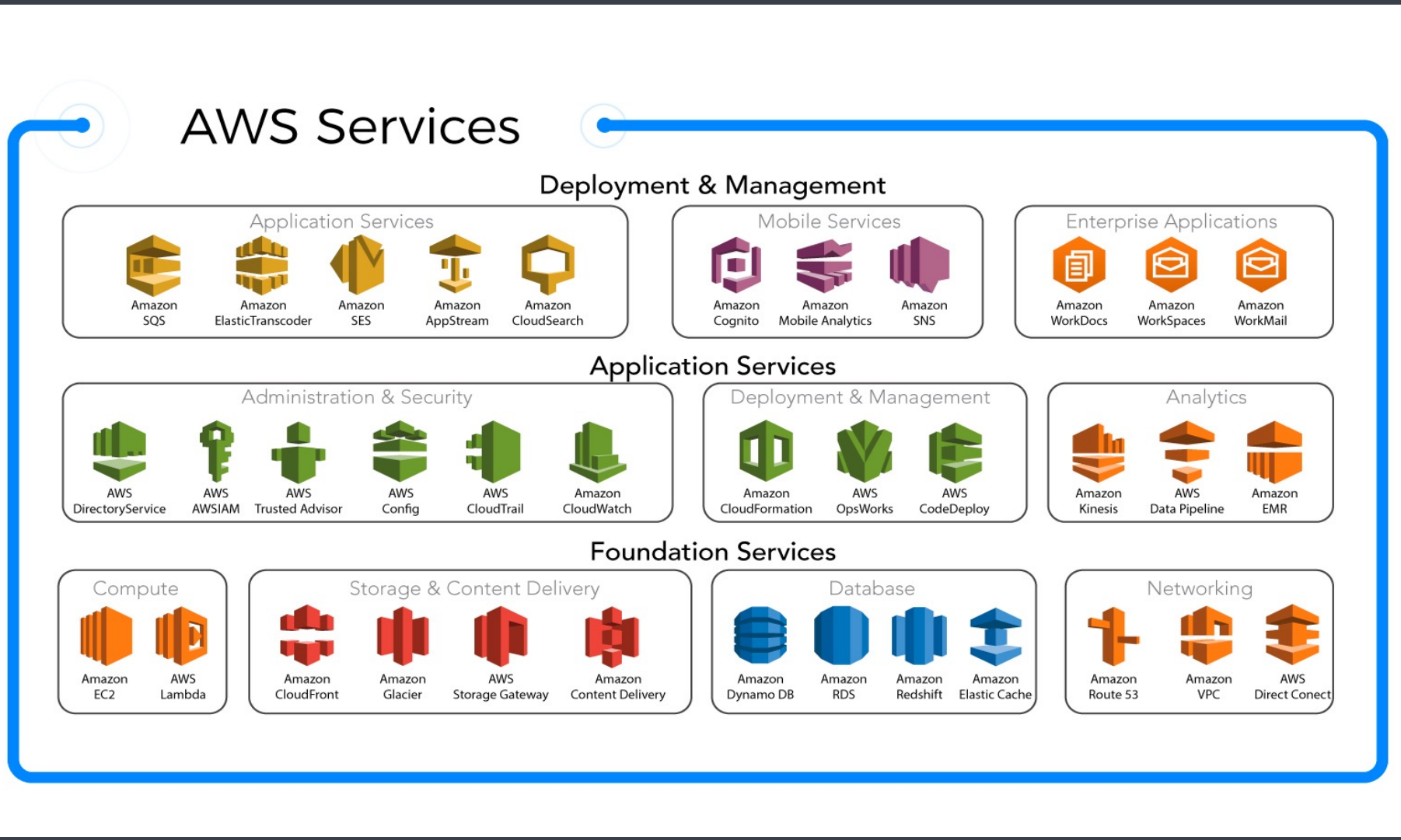
What is AWS ?

- AWS stands for Amazon Web Services
- Platform that offers flexible, reliable, scalable, easy-to-use and cost-effective cloud computing solutions
- Amazon’s cloud implementation
- It’s a combination of IaaS, PaaS and SaaS offerings

Traditional vs AWS



AWS Services



Global Infrastructure: Region

- Geographic area having availability zone(s)
- Collection of availability zones that are geographically located close to one other
- Every Region will act independently of the others, and each will contain at least two Availability Zones
- E.g.
  - US East: N. Virginia, Ohio
  - US West: N. California, Oregon
  - Asia Pacific: Mumbai, Seoul, Singapore, Sydney, Tokyo

Global Infrastructure: Availability Zone

- Essentially the physical data centers of AWS
- This is where the actual compute, storage, network, and database resources are hosted that we as consumers provision within our Virtual Private Clouds (VPCs)
- Availability Zones are always referenced by their Code Name, which is defined by the AZs Region Code Name that the AZ belongs to, followed by a letter
- E.g.
  - the AZs within the eu-west-1 region (EU Ireland), are
    - eu-west-1a
    - eu-west-1b
    - eu-west-1c

Global Infrastructure: Edge Locations

- Edge Locations are AWS sites deployed in major cities and highly populated areas across the globe
- Generally used to cache data and reduce latency for end-user access by using the Edge Locations as a global Content Delivery Network (CDN)
- Edge Locations are primarily used by end users who are accessing and using your services
- E.g.
  - Route 53: DNS Lookup
  - CloudFront
    - Content Delivery Network (CDN)
    - Cached contents, streaming distribution, acceleration

EC2

EC2

- Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud
- It is a virtual machine you will be building in the cloud
- EC2 instances are designed to mimic traditional on-premise servers, but with the ability to be commissioned and decommissioned on-demand for easy scalability and elasticity
- EC2 supports variety of operating systems:
  - Linux: Amazon Linux, Ubuntu, Red Hat Enterprise, SUSE Linux Enterprise Server, Fedora, Debian, CentOS, Gentoo Linux, Oracle Linux, FreeBSD
  - Windows: Windows Server, Windows
- Every instance comprised of
  - Amazon Machine Image (AMI)
  - Instance type
  - Network Interface
  - Storage

Amazon Machine Image (AMI)

- Operating System used to create virtual machine (EC2 instance)
- AMI are built for a specific region
- You can copy an AMI from one region to another
- You can also create a custom AMI with required applications/configuration
- AMI contains
  - Template for root volume
  - Launch permissions that control which account can use the AMI
  - EBS mapping that specifies the volume(s) to attach the instance when its launched
- AMI comes into two types
  - Instance store backed AMI
  - EBS backed AMI

Instance Type

- Used to decide the EC2 instance configuration
- AWS provides various instance types [https://aws.amazon.com/ec2/instance-types/]
  - General purpose: A (ARM), T (Cheapest), M (Main)
  - Compute optimized: C (Compute)
  - Memory optimized: R (RAM), X (Extreme RAM), Z (High compute and memory)
  - Accelerated computing: P (Picture-GPU), G (Graphics), F (Fast)
  - Storage optimized: I (IOPS), D (Data), H (High Disk Throughput)

Instance Types

Type	Category	Description	Use Cases
M5	General Purpose	Balance of compute, memory and network resources	Mid-sized databases
C5	Compute Optimized	Advanced CPUs	Modelling, Analytics
H1	Storage Optimized	Local HDD Storage	Map Reduce
R4	Memory Optimized	More RAM for $	In-memory caching
X1	Memory Optimized	Terabytes of RAM and SSD	In-memory database
I3	IO Optimized	Local SSD storage, high IOPS	NoSQL databases
G3	GPU Graphics	GPUs with video encoders	3d rendering
P3	GPU Compute	GPUs with tensor cores	Machine Learning
F1	Accelerated Computing	FPGA, custom hardware accelerations	Genomics
T2	Burstable	Shared CPUs, lowest cost	Web servers

Security Group

- Acts as a virtual firewall for your instance to control inbound and outbound traffic
- Controls the ports and protocols that can reach the front-end listener
- Every EC2 instance must have at least one security group attached
- Up to 5 security groups can be attached to an EC2 instance
- Security groups act at the instance level, not the subnet level
- Security group contains rules
  - You can specify allow rules, but not deny rules
  - You can specify separate rules for inbound and outbound traffic
  - When you create a security group, it has no inbound rules
  - By default, a security group includes an outbound rule that allows all outbound traffic
  - Security groups are stateful
  - Instances associated with a security group can't talk to each other unless you add rules allowing it
  - Security groups are associated with network interfaces

EC2 Key Pairs

- Uses PEM format (Privacy Enhanced Mail)
- Used to authenticate a client when logging into EC2 instance
- Each key pair consists of a public key and a private key
- AWS stores the public key on the instance and your are responsible for storing the private key
- To log into the instance you must create and authenticate with key pair
  - Linux instances have no password and you use a key pair to log in
  - With windows you use a key pair to obtain the administrator password and then log into the instance with RDP
- During the creation process of an EC2 instance you are required to either create a new key pair or use existing pair
- The private key is available for download and stored on your local drive
- NOTE: it will be available only once in the form of .pem file