Learning Outcomes:-

- **Data Center Fundamentals**
  - Historical perspective and evolution
  - Key components of a data center
- **Data Center Networking**
  - Data center network topologies
  - SDN (Software-Defined Networking) in data center
- **Data Center Automation and Scaling**
  - Automation in Data Centers
  - Infrastructure as Code (IaC) and automation tools
  - Scalability and elasticity in cloud data centers

INTRODUTION TO Data Center

- A data center is a physical space that is environmentally controlled with clean electrical power and network connectivity that is optimized for hosting servers.
- The temperature and humidity of data center environment are controlled to enable proper operation of the equipment and the facility is physically secured to prevent deliberate or accidental damage to the physical equipment.
- The data center is a place where can accommodate many computing resources that collect, store, share, manage, and distribute a large volume of data. It consists of all necessary data center facility elements (space, power, and cooling) and IT infrastructure elements (server, storage, and network) based on business requirements.
- It is a facility made up of networked computers, storage systems, and computing infrastructure that businesses and other organizations use to organize, process, store large amounts of data and to broadcast this data.
- A business typically relies heavily on applications, services, and data within a data center, making it a focal point and critical asset for everyday operations.

INTRODUTION TO Data Center

- Enterprise data centers increasingly incorporate cloud computing resources and facilities to secure and protect in-house, onsite resources.
- As enterprises increasingly turn to cloud computing, the boundaries between cloud providers' data centers and enterprise data centers become less clear.

INTRODUTION TO Data Center

- This facility will have one or more connections to the public Internet, often via redundant and physically separated cables into redundant routers.
- Behind the routers will be security applications, like firewalls or deep packet inspection elements, to enforce a security perimeter protecting servers in the data centre.
- Behind the security appliances are often load balancers which distribute traffic across front end servers like web servers.
- Often there is one or two tiers of server behind the application front end like second tier servers implementing application or business logic and a third tier of database servers.
- Establishing and operating a traditional data center facility including IP routers and infrastructure, security applications, load balancers, servers' storage and supporting systems requires a large capital outlay and substantial operation expenses, all to support application software that often has widely varying load so that much of the resource capacity is often underutilized.

Working of a Data Center

- A data center facility enables an organization to assemble its resources and infrastructure for data processing, storage, and communication, including:
- **systems for storing, sharing, accessing, and processing data across the organization;**
- **physical infrastructure to support data processing and data communication;**
- **Utilities such as cooling, electricity, network access, and uninterruptible power supplies (UPS).**
- Gathering all these resources in one data center enables the organization to:
- **protect proprietary systems and data;**
- **Centralizing IT and data processing employees, contractors, and vendors;**
- **Enforcing information security controls on proprietary systems and data;**
- **Realize economies of scale by integrating sensitive systems in one place.**

Importance of a Data Center

- Data centers support almost all enterprise computing, storage, and business applications.
- To the extent that the business of a modern enterprise runs on computers, the data center is business.
- Data centers enable organizations to concentrate their processing power, which in turn enables the organization to focus its attention on:
- **IT and data processing personnel;**
- **computing and network connectivity infrastructure; And**
- **Computing Facility Security.**

Need of a Data Center

- Almost all business activities, including data storage and computing, are handled by data centers.
- The data center is the business of a modern firm to the extent that it is run on computers.
- In the world of business IT, data centers are built to support company applications and activities such as:
- **Use email and sharing to communicate**
- **The computer software that increases productivity**
- **Customer relationship management (CRM)**
- **Enterprise resource planning (ERP) and databases**
- **Artificial intelligence, machine learning, and big data are all terms that refer to the use of computers.**
- **Virtual desktop infrastructure, collaboration, and communication services**

Evolution of Data Centers

- The origins of the first data centers can be traced back to the 1940s and the existence of early computer systems such as the Electronic Numerical Integrator and Computer (ENIAC).
- These early machines were complicated to maintain and operate and had cables connecting all the necessary components.
- They were also in use by the military - meaning special computer rooms with racks, cable trays, cooling mechanisms, and access restrictions were necessary to accommodate all equipment and implement appropriate safety measures.
- However, it was not until the 1990s, when IT operations began to gain complexity and cheap networking equipment became available, that the term data center first came into use.
- It became possible to store all the necessary servers in one room within the company. These specialized computer rooms gained traction, dubbed data centers within organizations.

Evolution of Data Centers

- At the time of the dot-com bubble in the late 1990s, the need for Internet speed and a constant Internet presence for companies required large amounts of networking equipment required large facilities.
- At this point, data centers became popular and began to look similar to those described above.
- In the history of computing, as computers get smaller and networks get bigger, the data center has evolved and shifted to accommodate the necessary technology of the day.

Key Components of Data Centers

- There are mainly Three components of data centers.
- **storage**
- **Network**
- **Compute resource**
- A **modern data center** concentrates an organization's data systems in a well-protected physical infrastructure, which includes:
- Server;
- storage subsystems;
- networking switches, routers, and firewalls;
- cabling;
- Physical racks for organizing and interconnecting IT equipment.

Key Components of Data Centers

- **Datacenter Resources typically include:**
- power distribution and supplementary power subsystems;
- electrical switching;
- UPS;
- backup generator;
- ventilation and data center cooling systems, such as in-row cooling configurations and computer room air conditioners; And
- Adequate provision for network carrier (telecom) connectivity.
- It demands a physical facility with physical security access controls and sufficient square footage to hold the entire collection of infrastructure and equipment.

Key Components of Data Centers

Network infrastructure:-

- Datacenters require physical components between servers, switches, routers, and firewalls to connect them to the outside world.
- They can handle a lot of traffic without slowing down when set up correctly and structurally.
- The core layer is connected to the access layer through core switches at the edge, while the data center's Internet connection is accessed through a middle aggregate layer.
- Routers, switches, and firewalls facilitate communication between servers, devices, and external networks.
- Load balancers distribute network traffic across multiple servers to optimize performance and ensure redundancy.
- Advances such as cloud-level agility and scalability are now available in on-premises networks through advancements like hyper scale network data center security and software-defined networking.

Key Components of Data Centers

Storage infrastructure:-

- Sensitive data is housed in data center equipment used for both organizational and customer needs.
- Increasing the amount of storage accessible by backing up data in multiple formats increases storage capacity.
- Storage systems are responsible for storing and retrieving data.
- They can include traditional hard disk drives (HDDs), solid-state drives (SSDs), and network-attached storage (NAS) or storage area network (SAN) solutions.
- In addition, non-volatile storage technologies have improved data access speeds. Software-defined storage systems also enhance employee efficiency in managing a storage system, just as software-defined networking does.

Key Components of Data Centers

Computing resources:-

- The data center's engines are its servers.
- Servers may use a variety of mechanisms to process and memory, depending on the platform: physical processing and memory, virtualized processing and memory, distributed across containers, or distributed among remote nodes in an edge-computing architecture.
- Because large data centers must execute tasks that are most appropriate for them, specialized processors such as those specializing in artificial intelligence (AI) and machine learning (ML) may not be the best option.

Key Components of Data Centers

Support Infrastructure:-

Power Infrastructure:

- Uninterruptible Power Supply (UPS): Provides a temporary power source during outages, ensuring continuous operation.
- Power Distribution Units (PDUs): Distribute power to servers and networking equipment.

Cooling Systems:

- Cooling systems maintain an optimal temperature within the data center to prevent equipment overheating.
- They include air conditioning units, precision air conditioners, and liquid cooling solutions.

Security Systems:

- Physical Security: Includes measures such as access controls, surveillance cameras, and biometric authentication to secure the physical premises.
- Digital Security: Involves firewalls, intrusion detection/prevention systems, and encryption to protect data and networks.

Key Components of Data Centers

Fire Suppression Systems:

- Specialized systems, such as gas-based fire suppression or water mist systems, are designed to quickly extinguish fires without damaging sensitive equipment.

Backup and Disaster Recovery Systems:

- Backup systems create copies of critical data to prevent data loss in case of hardware failure or other disasters.
- Disaster recovery solutions ensure the ability to quickly recover operations after a catastrophic event.

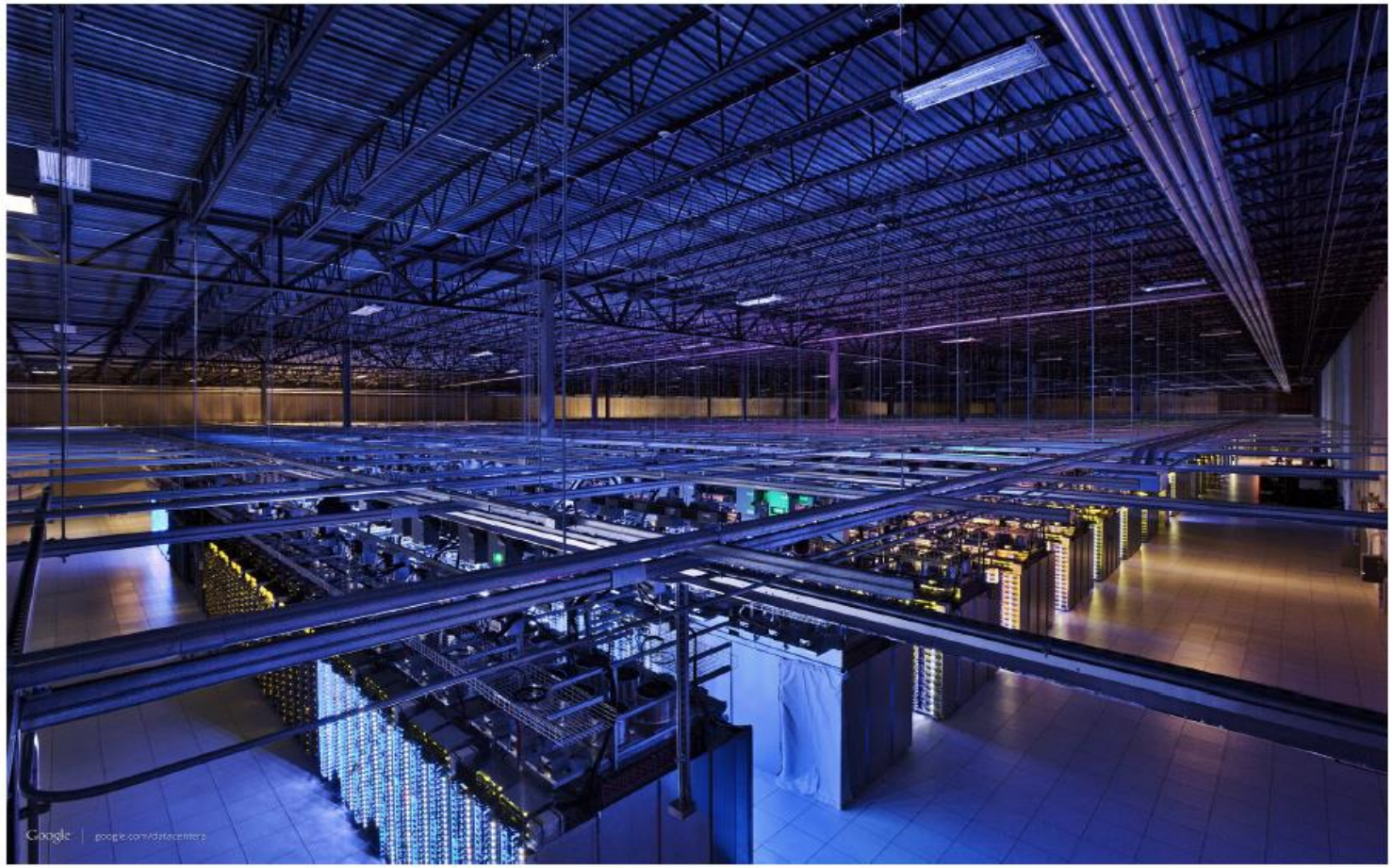
Physical Infrastructure:

- Raised Flooring: Provides space for cabling and airflow management.
- Server Racks: House and organize servers and networking equipment.
- Cable Management: Ensures organized and efficient routing of cables within the data center.

Monitoring and Analytics Tools:

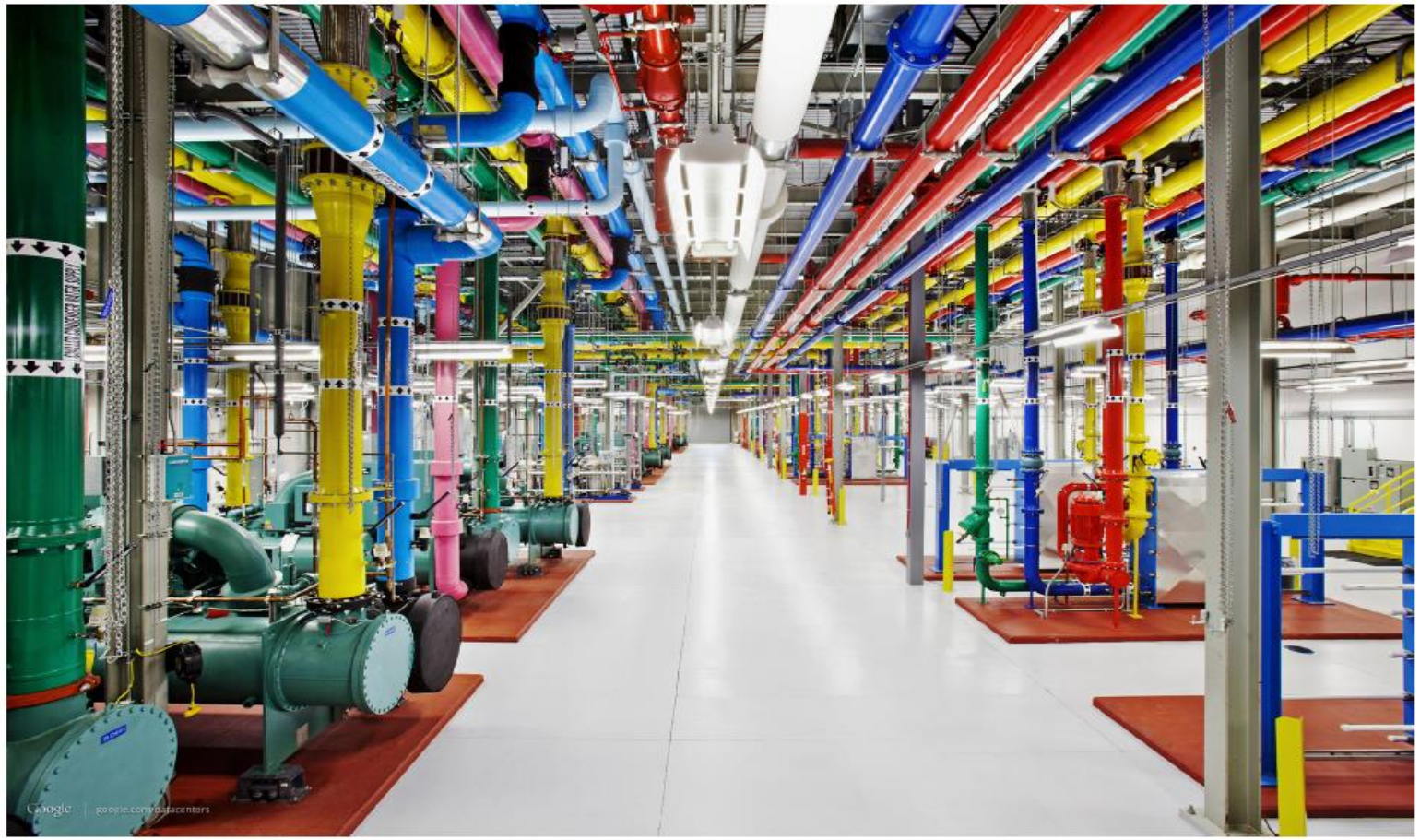
- Tools for real-time monitoring, performance analysis, and predictive analytics help optimize resource usage and identify potential issues before they impact operations.

Google's Data Center



Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Google's Data Center Cooling Plant



Source: http://webodysseum.com/technologyscience/visit-the-googles-data-centers/

Data Center Equipment Cabinets

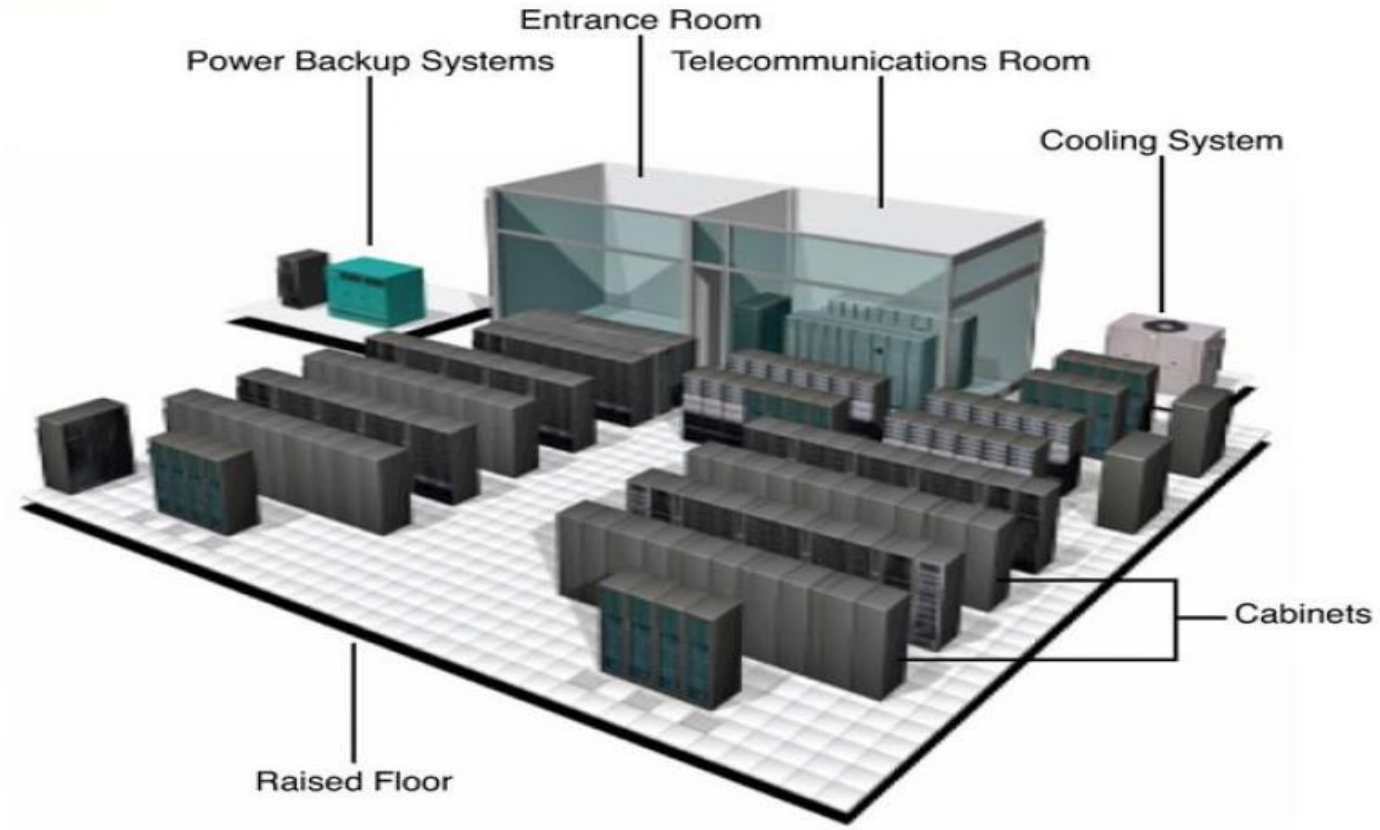
Three Layers: Bottom: Signal, Middle: Power, Top: Fiber

Minimize patching between cabinets and racks

Cabling under raised floors provides better appearance and cooling

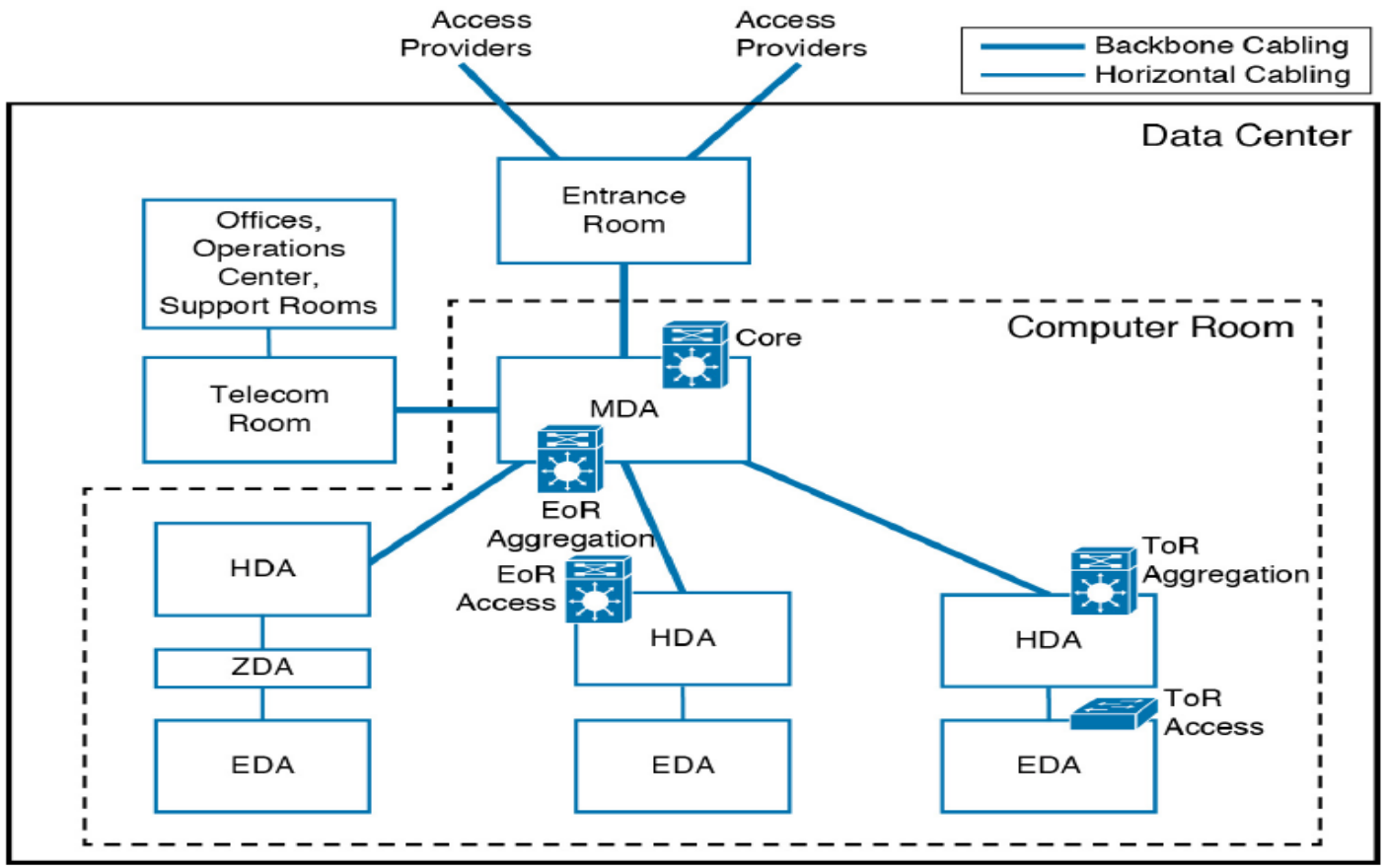
Ref: Ref: C. DiMinico, "Telecommunications Infrastructure Standard for Data Centers," IEEE 802.3 HSSG Meeting, Nov. 2006, http://www.ieee802.org/3/hssg/public/nov06/diminico_01_1106.pdf

Data Center Physical Layout



ANSI/TIA-942-2005 Standard

- Main Distribution Area (MDA)
- Horizontal Distribution Area (HDA)
- Equipment Distribution Area (EDA)
- Zone Distribution Area (ZDA)



Source: Santana 2014

ANSI/TIA-942-2005 Standard

- Computer Room: Main servers
- Entrance Room: Data Center to external cabling
- Cross-Connect: Enables termination of cables
- Main Distribution Area (MDA): Main cross connect. Central Point of Structured Cabling. Core network devices
- Horizontal Distribution Area (HDA): Connections to active equipment.
- Equipment Distribution Area (EDA): Active Servers+Switches. Alternate hot and cold aisle.
- Zone Distribution Area (ZDA): Optionally between HDA and EDA. ZDA allows easy
- Backbone Cabling: Connections between MDA, HDA, and Entrance room

Data Center Support Infrastructure

- Data centers are a critical asset that is protected with a robust and reliable support infrastructure made up of power subsystems, uninterruptible power supplies (UPS), backup generators, ventilation and cooling equipment, fire suppression systems and building security systems.
- Industry standards exist from organizations like the Telecommunications Industry Association (TIA) and the Uptime Institute to assist in the design, construction and maintenance of data center facilities. For instance, Uptime Institute defines these four tiers:
- **Tier I: Basic capacity, must include a UPS.**
- **Tier II: Redundant capacity and adds redundant power and cooling.**
- **Tier III: Concurrently maintainable and ensures that any component can be taken out of service without affecting production.**
- **Tier IV: Fault tolerant, allowing any production capacity to be insulated from ANY type of failure.**

Data Center Support Infrastructure

Tier 1: Basic site infrastructure:-

- These are the most basic data center designs and include a UPS. Tier I data centers do not have redundant systems, but they should ensure at least 99.671 percent availability.

Tier 2: Redundant-capacity component site infrastructure:-

- These data centers are designed for maximum uptime and include system, power, and cooling redundancy of 99.741%.

Tier 3: Concurrently maintainable site infrastructure:-

- Concurrently maintainable ensures that no component can be taken offline without disrupting operations. The data centers in this section are only partially fault-tolerant. Complete redundancy with a 99.98 percent uptime guarantee ensures that your system will run uninterruptedly even in the event of an outage.

Tier 4: Fault-tolerant site infrastructure:-

- It's fault-tolerant, so any production capacity can be insulated from any outage. As a result, the data centers have 99.995 percent availability, or no more than 26.3 minutes of downtime per year, total fault tolerance, system redundancy, and 96-hour disaster recovery.

Types of Data Center Facilities

- There are many types of data centers and service models to choose from in these facilities with various different data center infrastructure.
- The method by which data centers are classified differs based on whether organizations control them, how they match (if they fit) into the topology of other data centers, what computing and storage technologies they employ, and even their energy efficiency.
- The following are the four most frequent types of data centers:
- **Enterprise data centers**
- **Managed services data centers**
- **Colocation data centers**
- **Cloud data centers**

Types of Data Center Facilities

Enterprise data centers

- These are developed, owned, and operated by businesses, and they're designed to appeal to their consumers.
- They are generally located on the corporate campus.
- this is a type of data center built and owned by a company that may or may not be onsite.

Managed services data centers

- The data centers themselves are owned and maintained by a third party (or managed data center services provider) on behalf of a business.
- The firm purchases the equipment and infrastructure rather than renting them.
- Facilities where third-party providers offer managed IT services, including infrastructure management, security, and support.

Types of Data Center Facilities

Colocation data centers

- In colocation data centers, a firm leases space in a data center energy efficiency owned by others and located off the company's premises.
- The data center physical infrastructure, such as the facility, cooling systems, bandwidth, security, and other features, is housed by a colocation provider.
- The company provides and manages the components, such as servers, storage, and firewalls.

Cloud data centers

- A cloud managed services providers such as Amazon Web Services (AWS), Microsoft (Azure), IBM Cloud, or another public cloud computing provider hosts data and applications in this off-premises data center facility model.
- Deliver on-demand computing resources, storage, and services over the internet.

Data Center Networking

- Data center networking is the integration of a constellation of networking resources — switching, routing, load balancing, analytics, etc. — to facilitate the storage and processing of applications and data.
- Modern data center networking architectures leverage full-stack networking and security virtualization platforms that support a rich set of data services connecting everything from VMs, containers, and bare metal applications while enabling centralized management and granular security controls.
- This model of data center networking represents a significant shift from the standard networking model in data centers not long ago.
- From on-premises physical servers, to virtualized infrastructure, to an integrated edge-to-cloud model of networking and security that is present wherever apps and data live, data center networking has evolved greatly in a short time.

Need of Data Center Networking

- **Automation:-**Achieving speed and agility in modern data centers depends greatly on automated provisioning of networking services for applications. Far faster and more reliable than a human administrator, modern networking platforms not only find the most efficient way to program a network, balance workloads, and automate time-consuming tasks, they also respond dynamically to changes in usage.
- **Consistent policies:-**With modern data center networking responsible for integrating resources from edge to cloud, consistent application of policies is essential.
- **Single pane of glass:-**Typically connecting resources located both on-premises, in the cloud, and at the edge, modern data center networking platforms offer centralized management from a single console.
- **Granular security:-**Today's data center networking platforms often feature integrated security controls that can include micro-segmentation and IDS/IPS.
- **Global visibility:-** Most data center networking platforms can display a visual representation of the network and its interconnections, which makes troubleshooting network issues much easier.

Working of Data Center Networking

- A modern data center networking platform runs all network services required to support traditional enterprise applications entirely in software, enabling the automation of what were previously manual and error-prone provisioning tasks.
- It also makes possible capacity planning, security policy planning, and network troubleshooting. When applications are decommissioned, the networking platform handles de-provisioning policies associated with that application, which prevents the sprawl of stale policies that would otherwise degrade manageability, security, connectivity, and compliance.

Data Center Networking topologies

- Centralized
- Hierarchical (Three-Tier) Topology
- End of Rack(EOR-Zoned)
- Tope of Rack(TOR) (Leaf-Spine)
- Clos Topology
- Tree Topology
- Fat Tree Topology

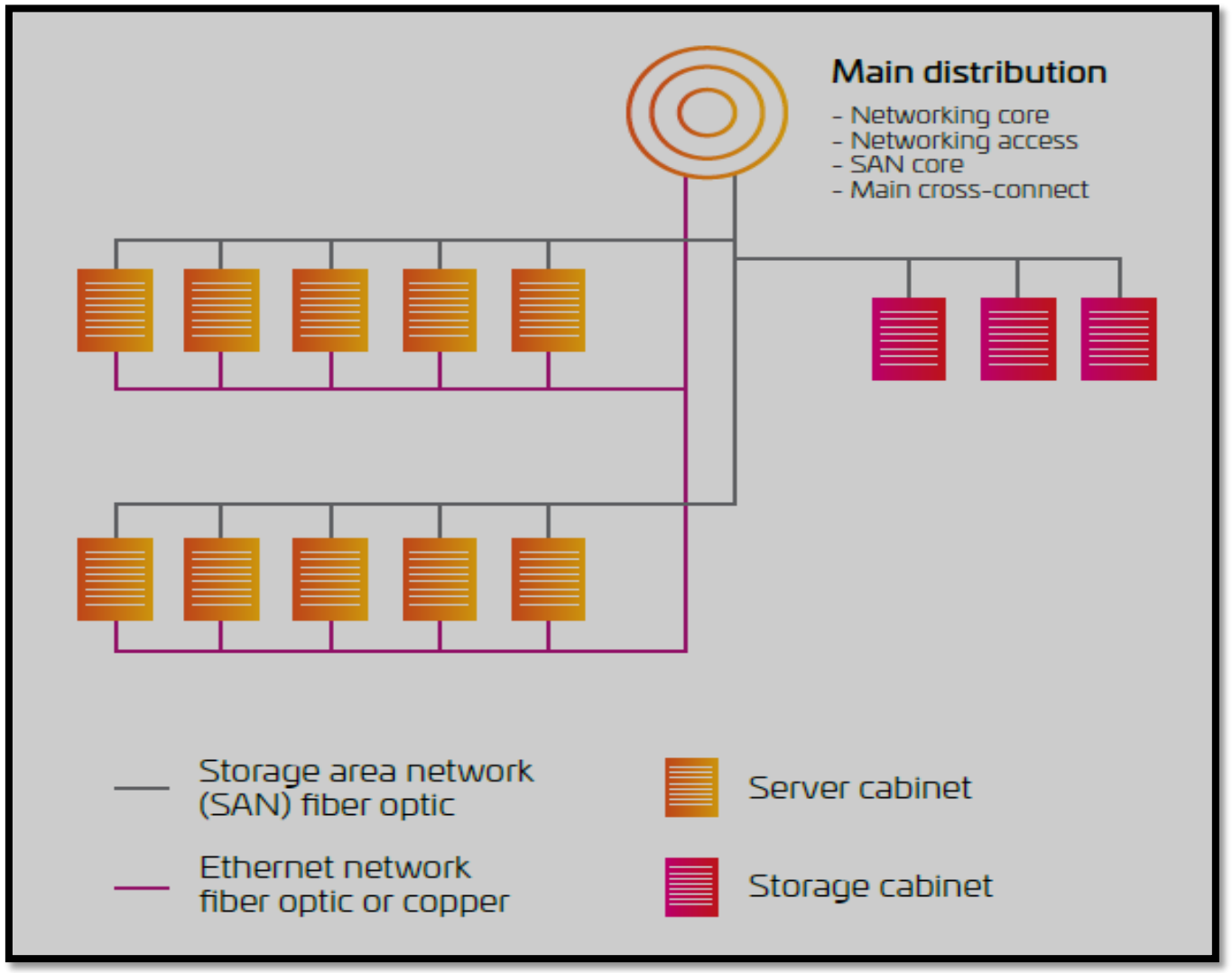
Data Center Networking topologies

Centralized Topology:-

- The centralized model is an appropriate topologies for smaller data centers (under 5,000 square feet).
- As shown, there are separate local area network (LAN)/ storage area network (SAN) environments and each one has home run cabling that goes to each of the server cabinets and zones.
- Each server is effectively cabled back to the core switches, which are centralized in the main distribution area.
- This provides very efficient utilization of port switches and makes it easier to manage and add components.
- The centralized topology works well for smaller data centers but does not scale up well, which makes it difficult to support expansions.
- In larger data centers, the high number of extended-length cable runs required causes congestion in the cable pathways and cabinets, and increases cost.

Data Center Networking topologies

Centralized:-



Data Center Networking topologies

Centralized Topology :-

- While some larger data centers use zoned or top-of-rack topologies for LAN traffic, they may also utilize a centralized architecture for the SAN environments.
- This is especially true where the cost of SAN switch ports is high and port utilization is important.

Advantages:-

- Resource Consolidation
- Cost Savings
- Improved Security

Disadvantages:-

- Single Point of Failure
- Network Bottlenecks

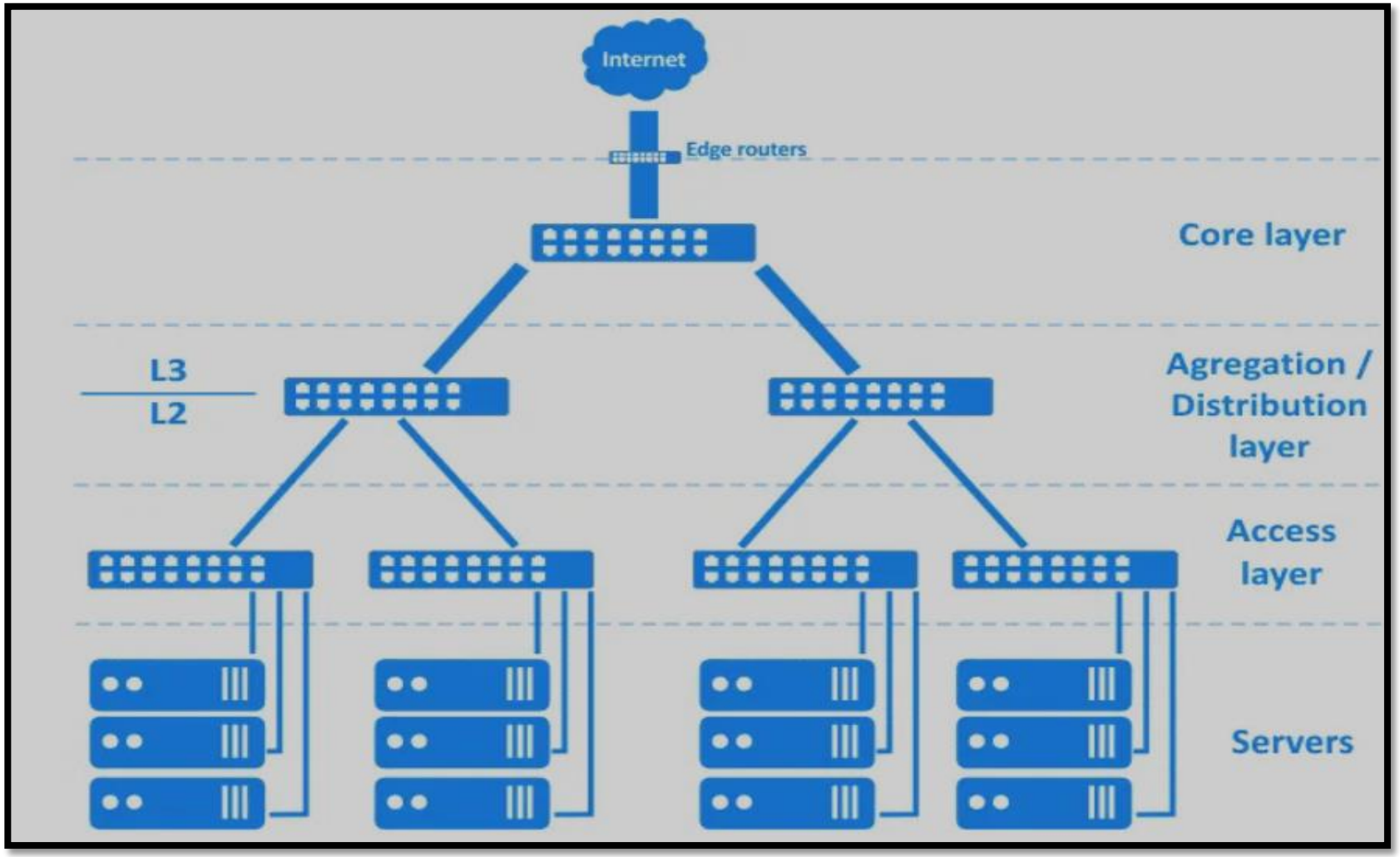
Data Center Networking topologies

Hierarchical (Three-Tier) Topology:-

- This topology breaks down into three primary layers:
- the core layer,
- the aggregation or distribution layer,
- the access layer.
- In this topology, servers are connected to switches in the access layer.
- Edge routers are connected to the core to provide access from/to WAN and the internet.

Data Center Networking topologies

Hierarchical (Three-Tier) Topology:-



Data Center Networking topologies

Hierarchical (Three-Tier) Topology:-

Core Layer:

- The core layer is the backbone of the data center network, providing high-speed, non-blocking connectivity between distribution switches.
- Its primary function is to facilitate fast and efficient data transfer between different parts of the network, without introducing bottlenecks.
- Core layer switches are typically high-performance devices capable of handling large volumes of traffic with minimal latency.

Distribution Layer:

- The distribution layer aggregates traffic from access switches and routes it to the core layer.
- It provides segmentation, policy enforcement, and access control within the network.
- Distribution layer switches often implement features such as VLANs (Virtual Local Area Networks), Quality of Service (QoS), and routing protocols to optimize traffic flow and ensure efficient network operation.

Data Center Networking topologies

Hierarchical (Three-Tier) Topology:-

Access Layer:

- The access layer connects end devices such as servers, storage systems, and user devices to the network.
- Its primary function is to provide connectivity and access to network resources for devices within the data center.
- Access layer switches typically have a high port density to accommodate numerous end devices and often support features such as Power over Ethernet (PoE) for powering devices like IP phones and wireless access points.

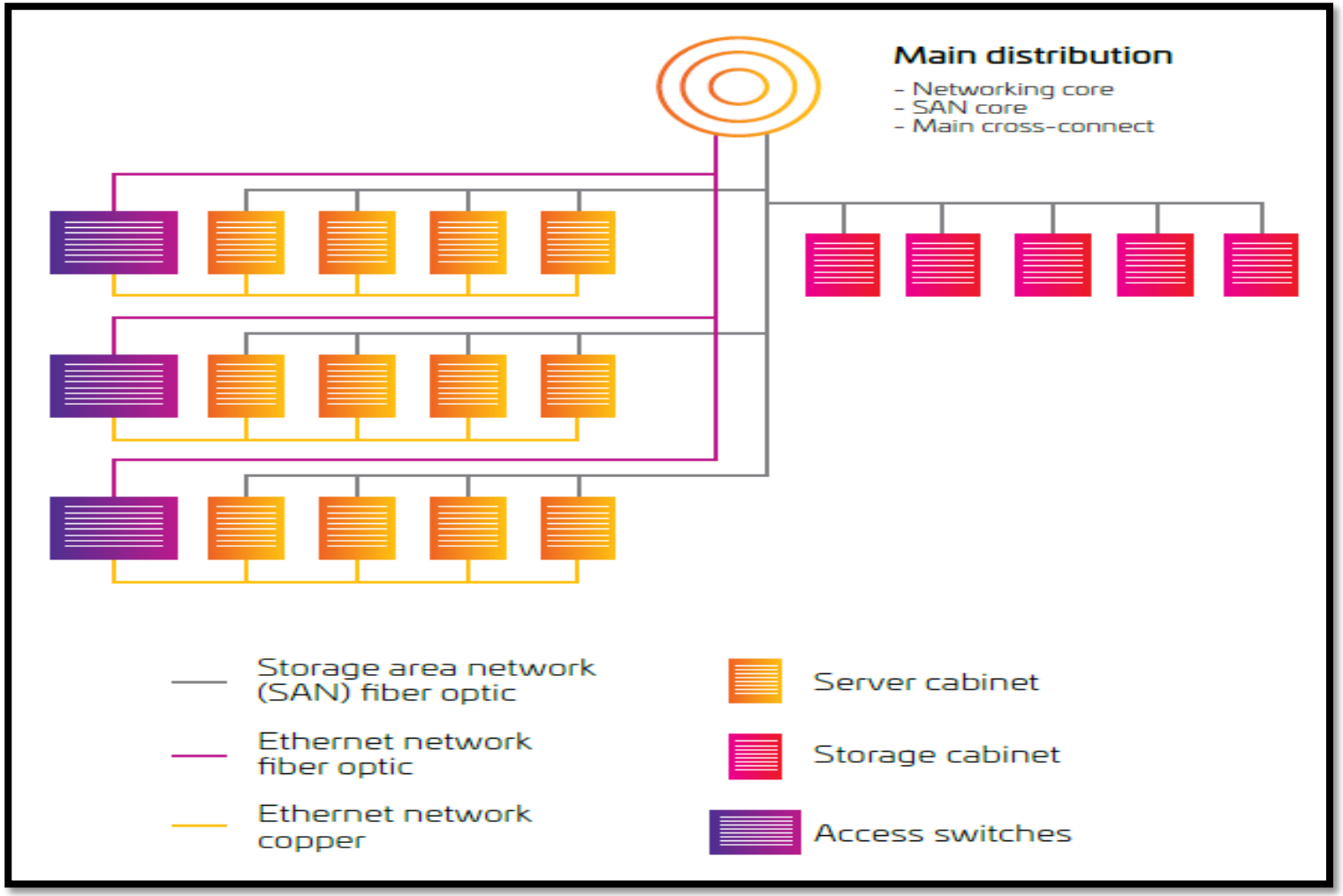
Data Center Networking topologies

End of Rack (Zoned)Topology:-

- Zoned topology consists of distributed switching resources.
- The switches can be distributed among end-of-row (EoR) or middle of- row (MoR) locations, with chassis-based switches typically used to support multiple server cabinets.
- This solution is recommended by the ANS/TIA-942 Data Center Standards and is very scalable, repeatable, and predictable.
- Zoned architecture is usually the most cost-effective design, providing the highest level of switch and port utilization while minimizing cabling costs.

Data Center Networking topologies

End of Rack (Zoned)Topology:-



Data Center Networking topologies

End of Rack (Zoned)Topology:-

- In certain scenarios, end-of-row switching provides performance advantages.
- For example, the local area network (LAN) ports of two servers (that exchange large volumes of information) can be placed on the same end-of-row switch, for low-latency port-to port switching.
- A potential disadvantage of end of- row switching is the need to run cable back to the end-of-row switch.
- Assuming every server is connected to redundant switches, this cabling can exceed what is required in top-of-rack architecture.

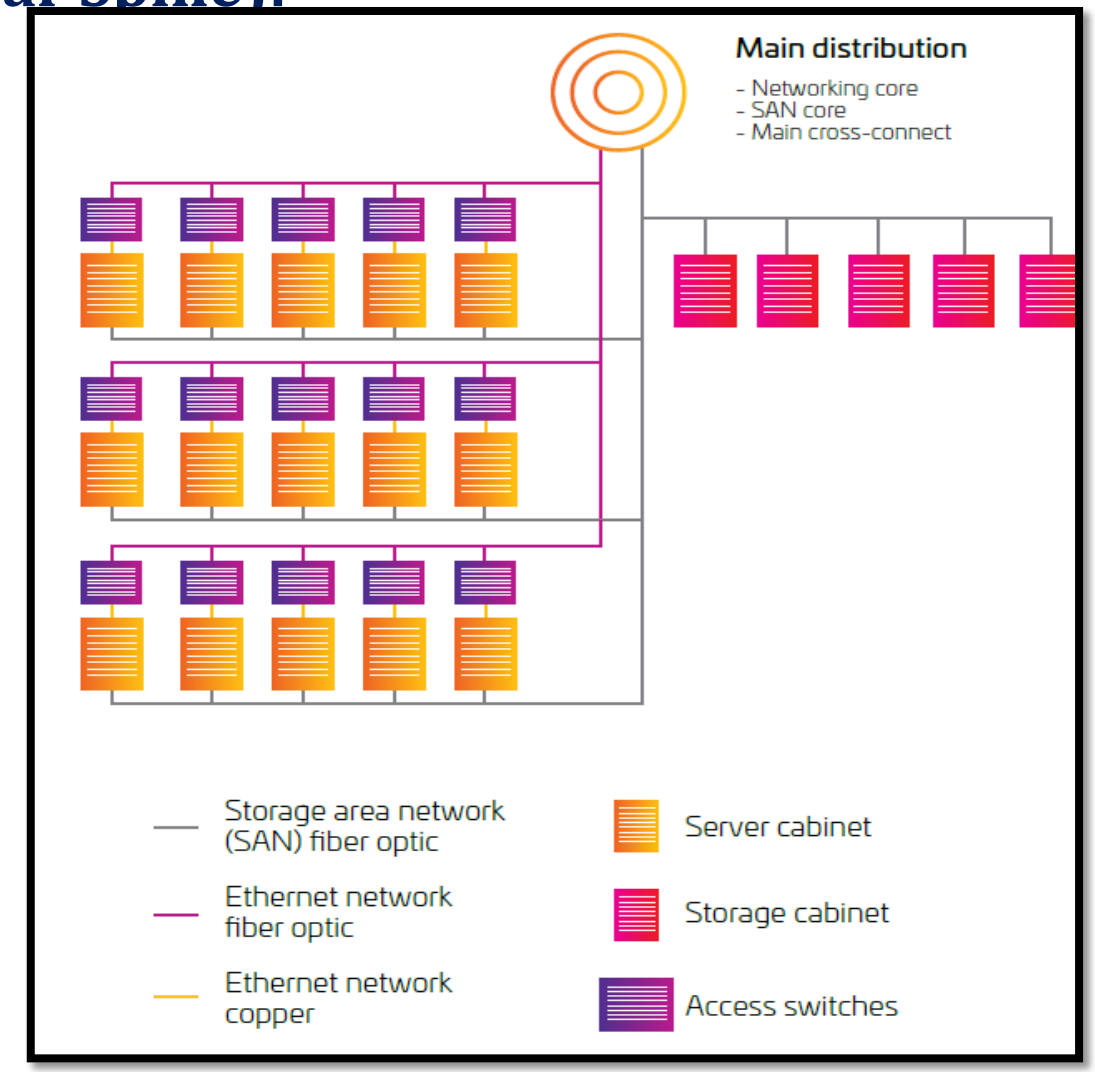
Data Center Networking topologies

Top of Rack Topology(Leaf-Spine):-

- Top-of-rack (ToR) switching typically consists of two or more switches placed at the top of the rack in each server cabinet, as shown below.
- This topology can be a good choice for dense one rack-unit (1RU) server environments.
- All servers in the rack are cabled to both switches for redundancy. The top-of-rack switches have uplinks to the next layer of switching.
- Top of rack significantly simplifies cable management and minimizes cable containment requirements.
- This approach also provides fast port-to-port switching for servers within the rack and predictable oversubscription of the uplink.

Data Center Networking topologies

Top of Rack Topology(Leaf-Spine):-



Data Center Networking topologies

Top of Rack Topology(Leaf-Spine):-

- A top-of-rack design utilizes cabling more efficiently.
- The tradeoffs are often an increase in the cost of switches and the high cost for under-utilization of ports.
- Top-of-rack switching may be difficult to manage in large deployments, and there is also the potential for overheating of local area network (LAN) switch gear in server racks.
- As a result, some data centers deploy top-of-rack switches in a middle-of-row or end-of-row architecture to better utilize switch ports and reduce the overall number of switches used.

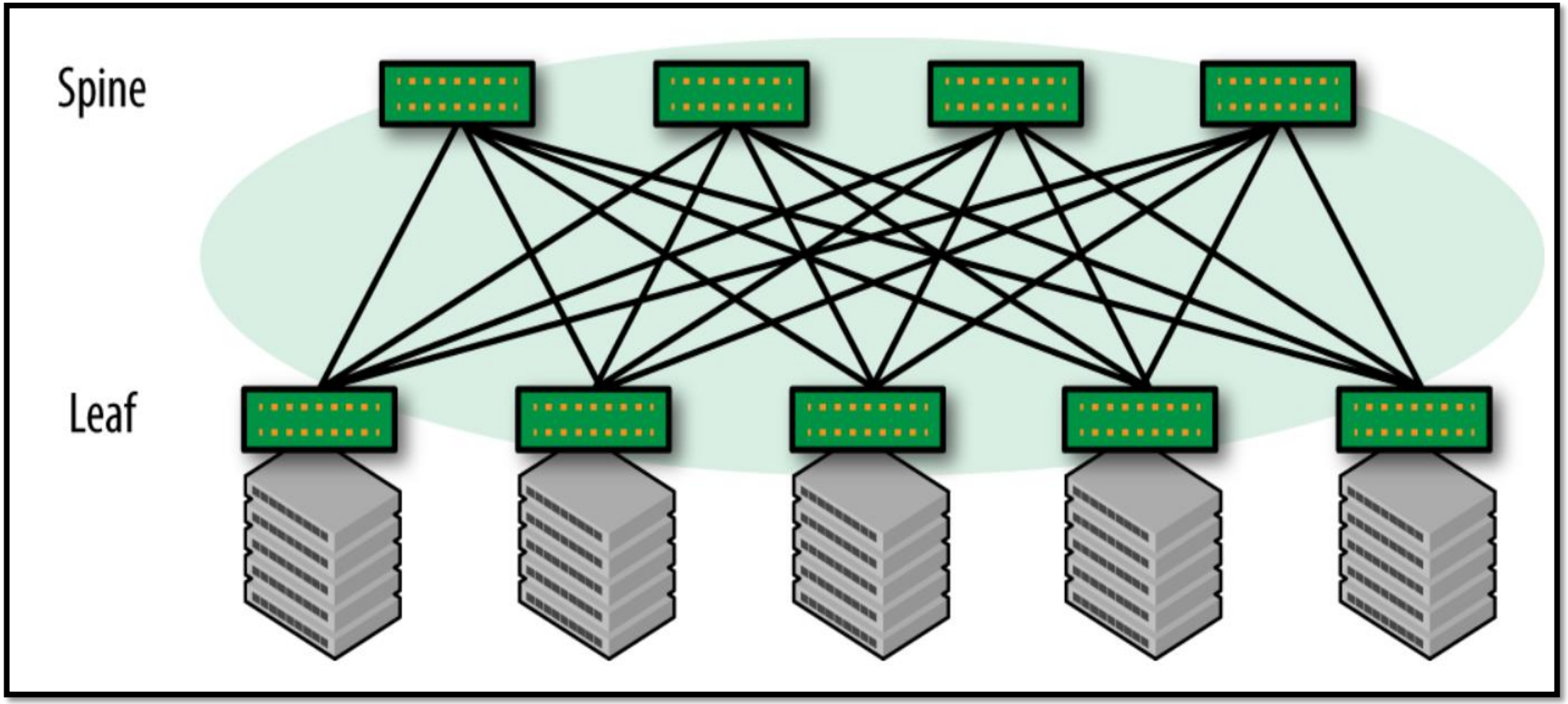
Data Center Networking topologies

Clos Topology(Leaf-Spine):-

- The Clos data center topology, also known as a Clos network or Clos fabric, is a highly scalable and resilient network architecture commonly used in large-scale data center environments.
- Named after its creator, Charles Clos, a French engineer, this topology is based on the principles of non-blocking switching and multi-stage switching networks.
- The Clos topology is often implemented using spine-and-leaf architecture, where spine switches provide connectivity between leaf switches.

Data Center Networking topologies

Clos Topology(Leaf-Spine):-



Data Center Networking topologies

Clos Topology(Leaf-Spine):-

Spine-and-Leaf Architecture:

- The Clos topology typically employs a spine-and-leaf architecture, consisting of spine switches and leaf switches.
- Spine switches form the core of the network and provide high-speed connectivity between leaf switches.
- Leaf switches connect servers, storage systems, and other network devices, providing access to the network fabric.

Non-Blocking Design:

- A key feature of the Clos topology is its non-blocking design, which ensures that there are enough paths available to handle any potential traffic flow without causing congestion.
- Non-blocking switching is achieved by ensuring that the number of spine switches and ports on each spine switch is sufficient to accommodate the traffic between leaf switches.

Data Center Networking topologies

Clos Topology(Leaf-Spine):-

Multi-Stage Switching Network:

- The Clos topology is based on a multi-stage switching network, where multiple layers of switches are interconnected in a hierarchical fashion.
- Each stage consists of a set of switches interconnected in such a way that every input/output port of each switch is connected to another switch at the next stage, forming a full-mesh topology.

Advantages:

- Scalability
- Redundancy and Resilience

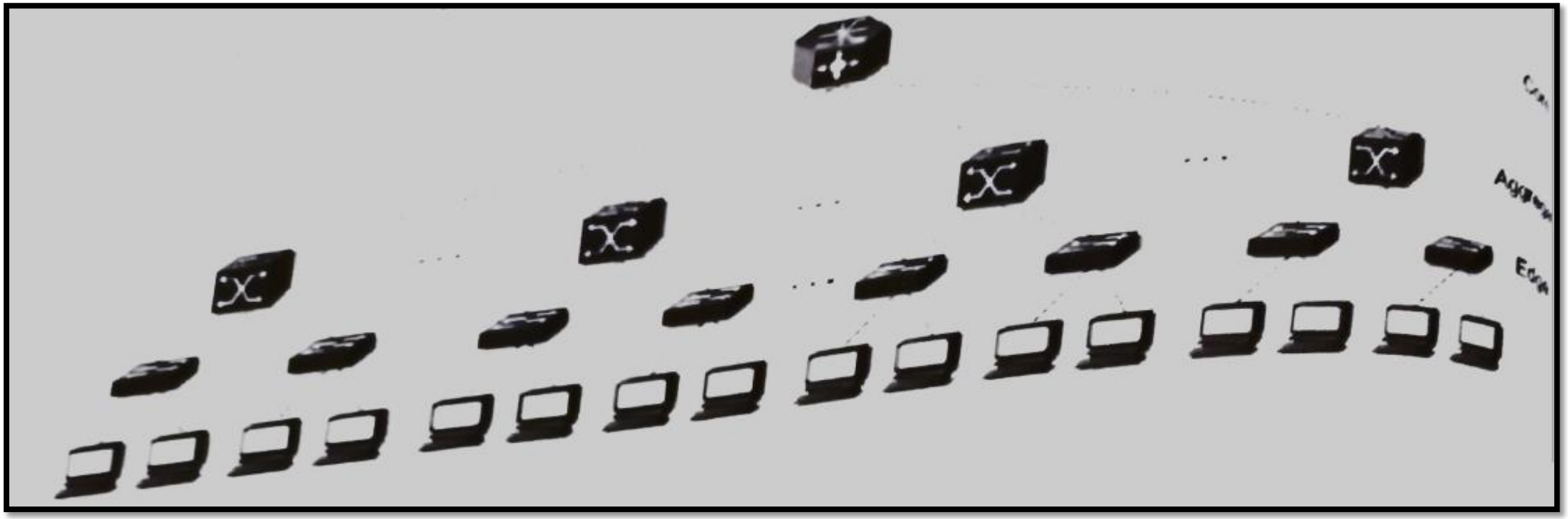
Data Center Networking topologies

Tree Topology:-

- In a tree topology, switches and routers are organized in a hierarchical tree-like structure.
- This topology is scalable and provides redundancy, as there are multiple paths for data transmission.
- It is commonly used in traditional data center architectures but may lack some of the flexibility required in highly dynamic cloud environments.

Data Center Networking topologies

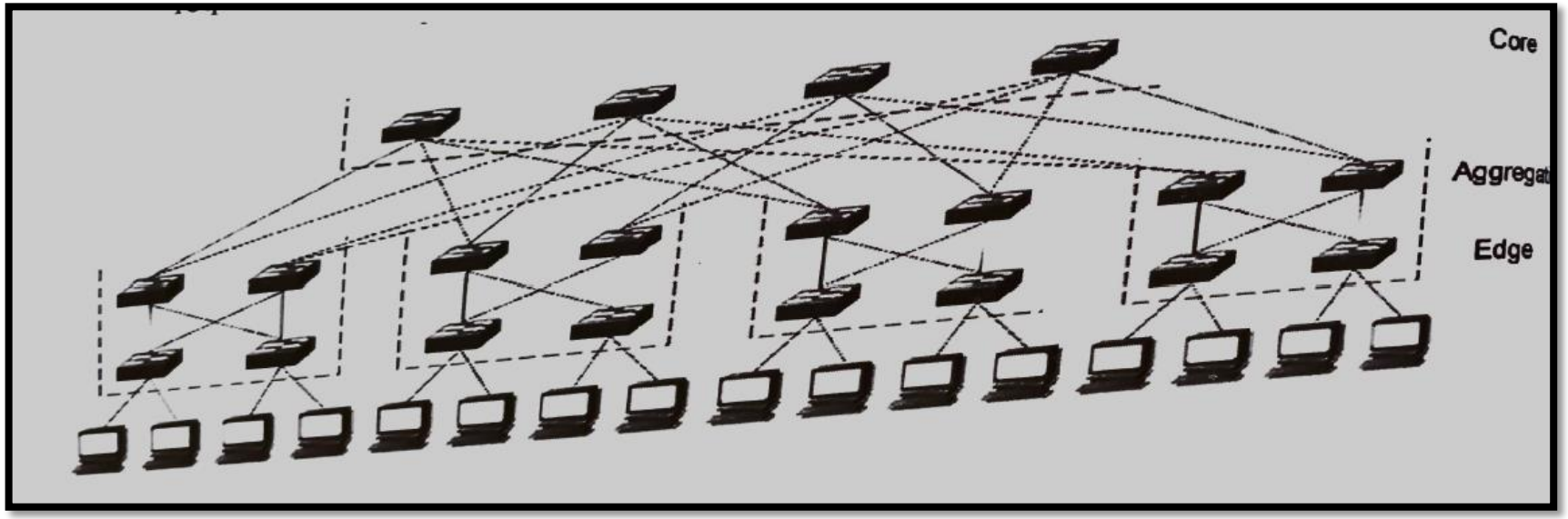
Tree Topology:-



Data Center Networking topologies

Fat-Tree Topology:-

- Fat-Tree is a type of Clos network, which is a non-blocking network architecture.
- It is highly scalable and provides low-latency, high-bandwidth connectivity.
- Fat-tree topologies are popular in large-scale cloud data centers due to their efficiency and fault-tolerance.



Software Defined Network

- SDN stands for Software Defined Network which is a networking architecture approach.
- It enables the control and management of the network using software applications.
- Through Software Defined Network (SDN) networking behavior of the entire network and its devices are programmed in a centrally controlled manner through software applications using open APIs.
- To understand software-defined networks, we need to understand the various planes involved in networking.
- **Data Plane**
- **Control Plane**

Software Defined Network

- **Data plane:**
- All the activities involving as well as resulting from data packets sent by the end-user belong to this plane.
- In computer networking, the data plane is the part of a network device responsible for forwarding data packets from one interface to another.
- It is also referred to as the forwarding plane or the user plane.
- This includes:
- Forwarding of packets.
- Segmentation and reassembly of data.
- Replication of packets for multicasting.

Software Defined Network

- **Control plane:**
- All activities necessary to perform data plane activities but do not involve end-user data packets belong to this plane.
- In computer networking, the control plane is part of a network device or system that is responsible for managing and controlling the flow of network traffic.
- In other words, this is the brain of the network.
- The activities of the control plane include:
- Making routing tables.
- Setting packet handling policies.

Software Defined Network in data center

- **Traditional Data Center Networking:**
- In a traditional data center network, the control plane (which makes decisions about where traffic is sent) and data plane (which forwards traffic to destinations) are coupled together on the network switches and routers.
- Configuring the network requires managing each network device individually via CLI or SNMP. This is cumbersome and error-prone.
- The network is fairly static and hard to change dynamically and require configuring each network device individually.

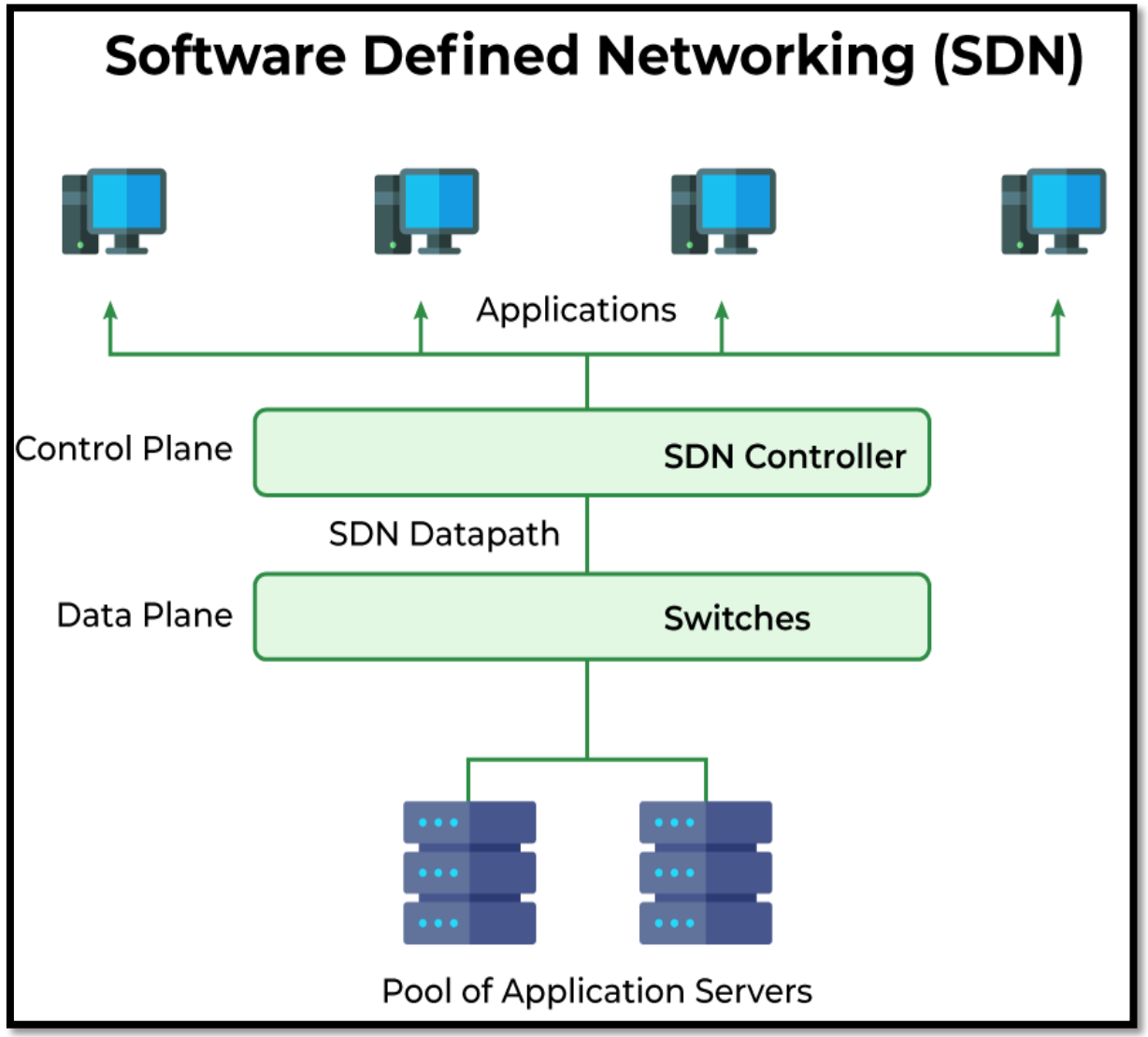
Software Defined Network in data center

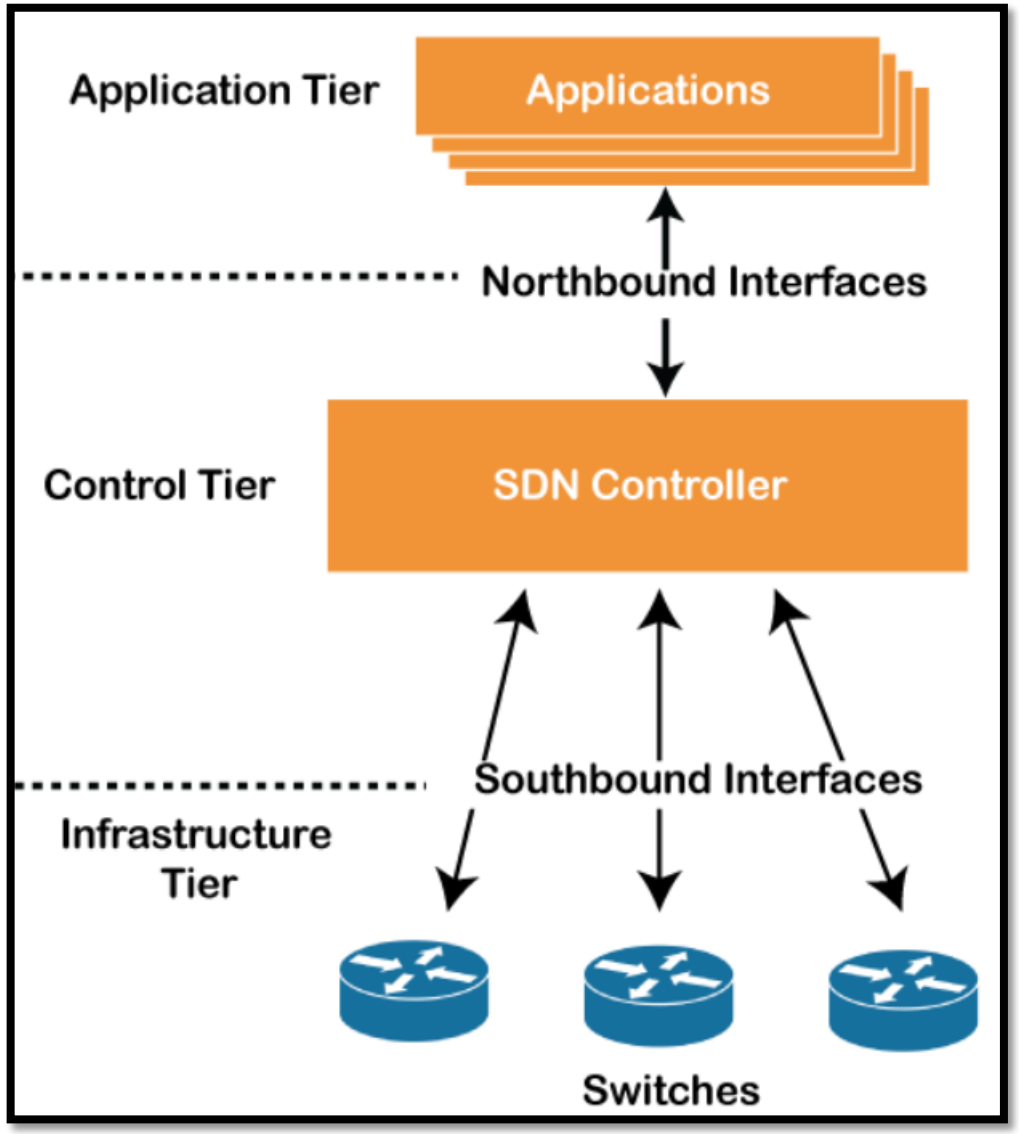
- **Software Defined Networking:**
- Decouples the control plane from the data plane.
- Control plane is centralized in software-based SDN controllers.
- SDN controllers have a global view of the entire network.
- SDN controllers communicate via APIs to simple forwarding devices and virtual network components.
- Forwarding devices simply forward packets based on rules set by the controllers.
- Allows programmable, dynamic configuration of network from a centralized point.
- SDN transforms the network into a programmable software-defined infrastructure to provide agility, flexibility and automation for cloud data centers. It is a key enabler of elasticity and efficiency in the cloud.

Software Defined Network

- **Key Benefits of SDN in data center:-**
- **Agile configuration:** Network can be dynamically programmed via software instead of needing manual device-by-device configuration. This allows the network to adapt quickly to changing requirements.
- **Centralized control:** Network control is centralized in SDN controllers instead of distributed across devices. This provides a unified view and point of automation.
- **Improved management:** Centralized controllers simplify network management by abstracting underlying network complexities.
- **Flexible traffic control:** Traffic can be dynamically routed based on application needs, load balancing, etc rather than static routing tables.
- **Better scalability:** Simpler forwarding devices are easier to scale out than fully featured routers and switches.
- **Automation:** SDN enables tight integration between network operations and automation/orchestration systems.
- **Reduced costs:** SDN's centralized control, automation and open standards hardware reduces operational costs.

Software Defined Network





Software Defined Network

SDN Architecture :-

- The architecture of software-defined networking (SDN) consists of three main layers:
- **The application layer**
- **The control layer**
- **The infrastructure layer**
- Each layer has a specific role and interacts with the other layers to manage and control the network.

Software Defined Network

- **Infrastructure Layer:**
- The infrastructure layer is the bottom layer of the SDN architecture, also known as the data plane.
- It consists of physical and virtual network devices such as switches, routers, and firewalls that are responsible for forwarding network traffic based on the instructions received from the control plane.

Software Defined Network

- **Control Layer:**
- The control layer is the middle layer of the SDN architecture, also known as the control plane.
- It consists of a centralized controller that communicates with the infrastructure layer devices and is responsible for managing and configuring the network.
- The controller interacts with the devices in the infrastructure layer using protocols such as OpenFlow to program the forwarding behaviour of the switches and routers.
- The controller uses network policies and rules to make decisions about how traffic should be forwarded based on factors such as network topology, traffic patterns, and quality of service requirements.

Software Defined Network

- **Application Layer:**
- The application layer is the top layer of the SDN architecture and is responsible for providing network services and applications to end-users.
- This layer consists of various network applications that interact with the control layer to manage the network.

Software Defined Network

- **Real time scenario: Example Library:**
- Traditional Networking(Non-SDN)
- In Traditional library , the librarian is responsible for managing and directing all activities. If you need a book , you ask the librarian, who then physically goes to the shelves , finds the book and given to you.
- SDN Cloud Computing (With an Analogy) :
- Imagine an SDN enabled library Instead of relying solely on the librarian, there's a smart system that controls everything.
- This system knows where every book is located and can instantly guide you to the right shelf.
- You don't need to wait for the librarian to fetch the book, you can go directly to the shelf and pick it up.

Software Defined Network

- **Applying the Analogy to SDN :**
- Control and Data Planes :
- Traditional: The librarian (human) acts as both the controller (making decisions) and the one executing tasks (finding books).
- SDN: The Smart system has a controller (making decisions) separate from the actual execution (finding books). This separation is like having a librarian and a computer system working together.
- Flexibility and Automation :
- Traditional: If you want to change how the library is organized, you need to ask the librarian to physically rearrange the shelves.
- SDN: With SDN, you can change how the library is organized (configure the network) using a computer interface without physically moving any books. It's like rearranging virtual shelves.

Software Defined Network

- **Applying the Analogy to SDN :**
- Adaptability (Example - New Book Arrival) :
- Traditional : When a new book arrives, the librarian needs to manually update the catalog and inform everyone about its location.
- SDN : The smart system automatically updates the catalog and informs everyone about the new book's location. It adapts to changes without human intervention.
- Efficiency (Example - Finding a Book) :
- Traditional : You wait for the librarian to find your book, potentially leading to delays.
- SDN: You go directly to the shelf and find your book instantly. SDN allows for faster and more direct access to resources.

Software Defined Network

- **Applying the Analogy to SDN :**
- SDN in a data center uses a centralized controller to make decisions about how data traffic should be directed.
- This controller communicates with switches and routers to manage the flow of data, providing a more flexible, automated, and efficient network infrastructure.
- Example in Cloud Computing :
- In a cloud data center, SDN enables dynamic allocation of resources. If a certain application needs more network bandwidth, the SDN controller can instantly adjust the network settings to meet that demand without manual configuration.
- In essence, SDN simplifies network management, improves efficiency, and allows for quick adaptation to changing conditions, much like the efficient organization and access of books in our smart library analogy.

Software Defined Network

- **Advantages of SDN:-**
- **Centralized Network Control:**
- One of the key benefits of SDN is that it centralizes the control of the network in a single controller, making it easier to manage and configure the network.
- This allows network administrators to define and enforce network policies in a more granular way, resulting in better network security, performance, and reliability.
- **Programmable Network:**
- In an SDN environment, network devices are programmable and can be reconfigured on the fly to meet changing network requirements.
- This allows network administrators to quickly adapt the network to changing traffic patterns and demands, resulting in better network performance and efficiency.
- **Cost Savings:**
- With SDN, network administrators can use commodity hardware to build a network, reducing the cost of proprietary network hardware.

Software Defined Network

- **Advantages of SDN:-**
- **Enhanced Network Security:** The centralized control of the network in SDN makes it easier to detect and respond to security threats. The use of network policies and rules allows administrators to implement fine-grained security controls that can mitigate security risks.
- **Scalability**: SDN makes it easier to scale the network to meet changing traffic demands. With the ability to programmatically control the network, administrators can quickly adjust the network to handle more traffic without the need for manual intervention.
- **Simplified Network Management:** SDN can simplify network management by abstracting the underlying network hardware and presenting a logical view of the network to administrators. This makes it easier to manage and troubleshoot the network, resulting in better network uptime and reliability.

Software Defined Network

- **Disadvantages of SDN:-**
- **Complexity:** SDN can be more complex than traditional networking because it involves a more sophisticated set of technologies and requires specialized skills to manage.
- For example, the use of a centralized controller to manage the network requires a deep understanding of the SDN architecture and protocols.
- **Dependency on the Controller:** The centralized controller is a critical component of SDN, and if it fails, the entire network could go down.
- This means that organizations need to ensure that the controller is highly available and that they have a robust backup and disaster recovery plan in place.
- **Compatibility:** Some legacy network devices may not be compatible with SDN, which means that organizations may need to replace or upgrade these devices to take full advantage of the benefits of SDN.

Software Defined Network

- **Disadvantages of SDN:-**
- **Security:** While SDN can enhance network security, it can also introduce new security risks. For example, a single point of control could be an attractive target for attackers, and the programmability of the network could make it easier for attackers to manipulate traffic.
- **Vendor Lock-In:** SDN solutions from different vendors may not be interoperable, which could lead to vendor lock-in. This means that organizations may be limited in their ability to switch to another vendor or integrate new solutions into their existing network.
- **Performance:** The centralized control of the network in SDN can introduce latency, which could impact network performance in certain situations. Additionally, the overhead of the SDN controller could impact the performance of the network as the network scales.

Data Center Automation and Scaling

- **Automation in Data Center:-**
- Data center automation, as the name implies, is the **process of managing and automating the data center workflows without human interaction or administration.**
- Data center automation is not about a single task or process in one single stage. It can be carried out in many different directions.
- Servers, storage, networking and other management tasks can be automated in data center automation.
- If there is a need to handle the entire data center operation with automation, different operational frameworks based on various tasks will be required.
- **Data center automation is the process in which the routine processes of data center operations are completed without any manual effort.**

Data Center Automation and Scaling

- **Automation in Data Center:-**
- Data center automation is the process by which routine workflows and processes of a data center—scheduling, monitoring, maintenance, application delivery, and so on—are managed and executed without human administration.
- Data center automation increases agility and operational efficiency. It reduces the time IT needs to perform routine tasks and enables them to deliver services on demand in a repeatable, automated manner.
- These services can then be rapidly consumed by end users.

Data Center Automation and Scaling

- **Importance of Automation in Data Center:-**
- The massive growth in data and the speed at which businesses operate today mean that manual monitoring, troubleshooting, and remediation is too slow to be effective and can put businesses at risk.
- Automation can make day-two operations almost autonomous.
- Ideally, the data center provider would have API access to the infrastructure, enabling it to inter-operate with public clouds so that customers could migrate data or workloads from cloud to cloud.
- Data center automation is predominantly delivered through software solutions that grant centralized access to all or most data center resources.
- Traditionally, this access enables the automation of storage, servers, network, and other data center management tasks.

Data Center Automation and Scaling

- **Importance of Automation in Data Center:-**
- Data center automation is immensely valuable because it frees up human computational time and:
- Delivers insight into server nodes and configurations
- Automates routine procedures like patching, updating, and reporting
- Produces and programs all data center scheduling and monitoring tasks
- Enforces data center processes and controls in agreement with standards and policies

Data Center Automation and Scaling

- **Key aspects of Automation in Data Center:-**
- **Provisioning and Configuration Management**
- Automated Provisioning: The process of deploying and configuring IT resources, such as servers, storage, and networking equipment, without manual intervention. This includes server provisioning, storage allocation, and network configuration.
- Configuration Management: Automation tools manage and maintain consistent configurations across a large number of devices, ensuring that they adhere to predefined standards and policies.
- **Orchestration and Workflow Automation**
- Orchestration: Involves the coordination and management of multiple automated tasks and processes to achieve specific business objectives. This includes the integration of various systems and tools to automate complex workflows, such as application deployment, disaster recovery, and scaling of resources.
- Workflow Automation: Automation of repetitive tasks and processes, such as backup scheduling, patch management, and compliance checks, to reduce manual effort and minimize errors.

Data Center Automation and Scaling

- **Key aspects of Automation in Data Center:-**
- **Monitoring and Remediation**
- Automated Monitoring: Continuous monitoring of infrastructure and applications to detect performance issues, security threats, and compliance violations. Automation tools can trigger alerts and responses based on predefined thresholds and conditions.
- Remediation: Automated actions taken to resolve identified issues, such as restarting services, reallocating resources, or applying corrective configurations to mitigate performance or security issues.
- **Self-Service and Lifecycle Management**
- Self-Service Portals: Provisioning of IT resources through user-friendly interfaces, allowing stakeholders to request and manage resources on-demand without administrative intervention.
- Lifecycle Management: Automated management of the entire lifecycle of IT resources, including provisioning, maintenance, updates, and retirement, to optimize resource utilization and minimize downtime.

Data Center Automation and Scaling

Infrastructure as Code (IaC):-

- **Infrastructure as Code (IaC) is the process of managing your IT infrastructure using automatic scripts instead of manually.**
- One of the crucial elements of the **DevOps software** development approach, it allows you to fully automate deployment and configuration, thus making continuous delivery possible.
- Infrastructure on the whole is a combination of components required to support the operations of your application.
- It consists of hardware such as servers, data centers, desktop computers and software including operating systems, web servers, etc.
- **Infrastructure as Code (IaC) is a practice that involves managing and provisioning computing infrastructure through machine-readable script files, rather than through traditional interactive configuration tools or physical hardware configuration.**
- It enables the automation of infrastructure deployment, configuration, and management, allowing organizations to treat their infrastructure as code and apply software development best practices to infrastructure management.

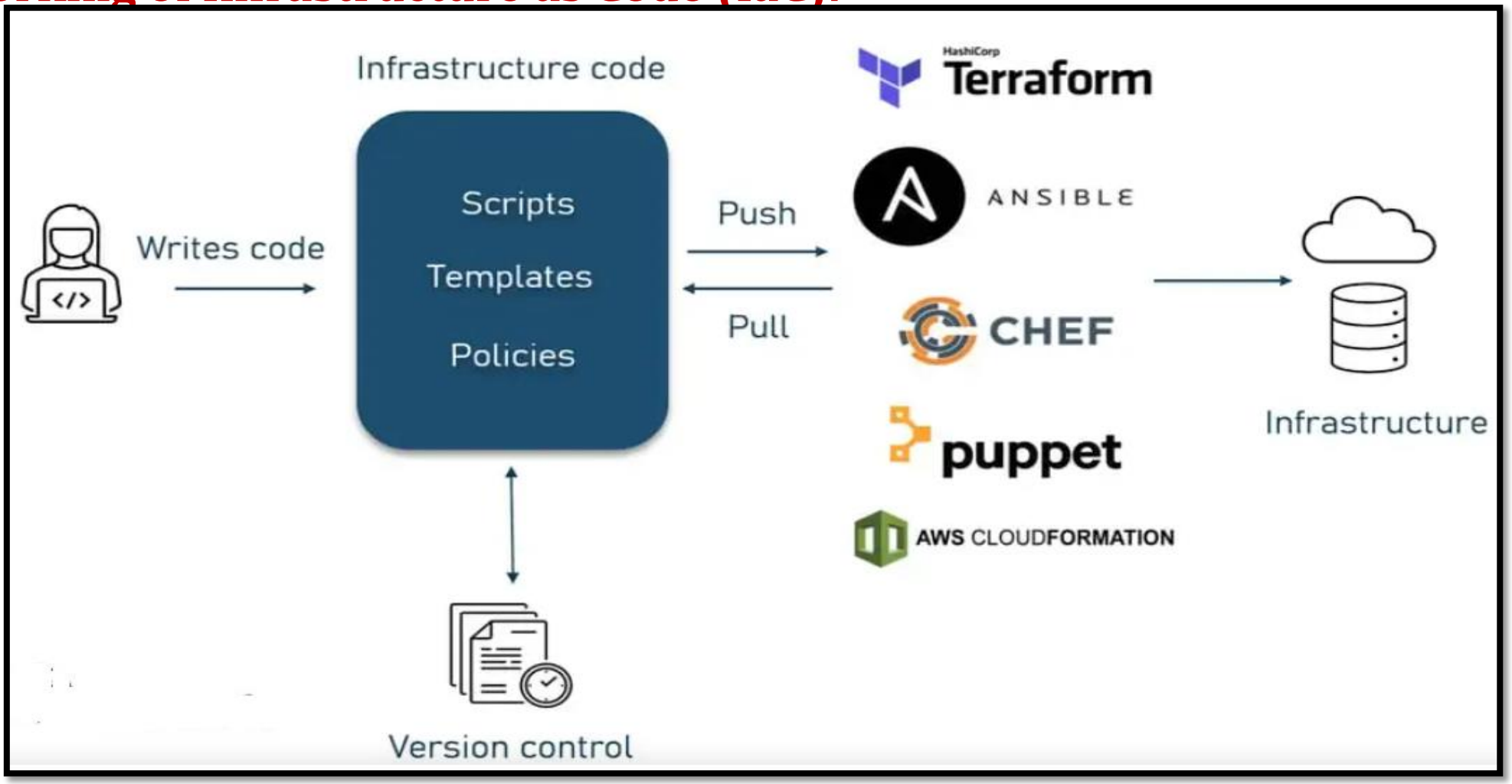
Data Center Automation and Scaling

Infrastructure as Code (IaC):-

- IaC uses software tools to automate administrative tasks by describing infrastructure in code.
- **Infrastructure as code (IaC) is the process of using software code for resource provisioning in data center and cloud environments, instead of hardware settings or configuration tools, with the benefits of automation and remote provisioning.**
- It's implemented like this:
- The team writes infrastructure configurations in a required programming language.
- The code files are sent to a code repository.
- An IaC tool runs the code and performs the required activities.

Data Center Automation and Scaling

Working of Infrastructure as Code (IaC):-



Data Center Automation and Scaling

Benefits of Infrastructure as Code (IaC):-

- **Reduced costs:-**
- Cloud computing by itself is more cost-effective since you don't have to spend money on hardware and people to maintain it.
- By automating with IoC, you also save on infrastructure costs since the effort to administer it decreases and your team gets to focus on more important tasks that bring your business value.
- **Consistency:-**
- manual deployment introduces many inconsistencies and variations.
- IaC avoids so-called configuration or environment drift by ensuring that deployments are repeatable and setting up the same configuration every time.

Data Center Automation and Scaling

Benefits of Infrastructure as Code (IaC):-

- **Easily duplicate an environment:-**
- The same environment can be deployed on a different system in another location using the same IaC, as long as the infrastructure resources are available.
- **Version control:-**
- Under IaC, infrastructure configurations are put into a code file that's easy to edit and distribute.
- As any other code, it can be checked into source control, versioned, and reviewed along with your application source code using existing practices.

Data Center Automation and Scaling

Choosing an IaC tool :- Declarative vs imperative approach

- There are two approaches to automating infrastructure: declarative and imperative.
- In a **declarative** approach, you only need to define the desired final condition of the infrastructure without going into detail on how it should be achieved, allowing the platform to handle it by itself.
- It's the dominant method because it requires less knowledge on the user's part and it's idempotent – it always produces the same result. This is done via a declarative language such as SQL, YAML, or JSON.
- In an **imperative (procedural)** approach, you must define specific step-by-step commands to achieve the required configuration.
- Here, you can be very detailed and create complex configurations, having more control over the task. But you do need a high level of skill to write those commands with languages like Java, Ruby, or Python.
- if you're building a large infrastructure, you might prefer the automation the declarative approach provides. But if you're planning to start with a few scripts, the imperative approach will work fine.

Data Center Automation and Scaling

- **Choosing an IaC tool :- Provisioning vs configuration management**
- Typically, IaC tools are divided into the ones that do configuration management and the provisioning tools. They are both steps of the deployment process, but there are differences.
- **Provisioning** is the process of setting up an IT infrastructure: virtual machines, databases, access to resources, and making them available to the users.
- The configuration job is then done by a different tool.
- **Configuration management** comes after provisioning and it means installing software, configuring servers' desired state and maintaining them.
- Some IaC tools can do both, but one is typically more equipped for a specific task than the other.

Data Center Automation and Scaling

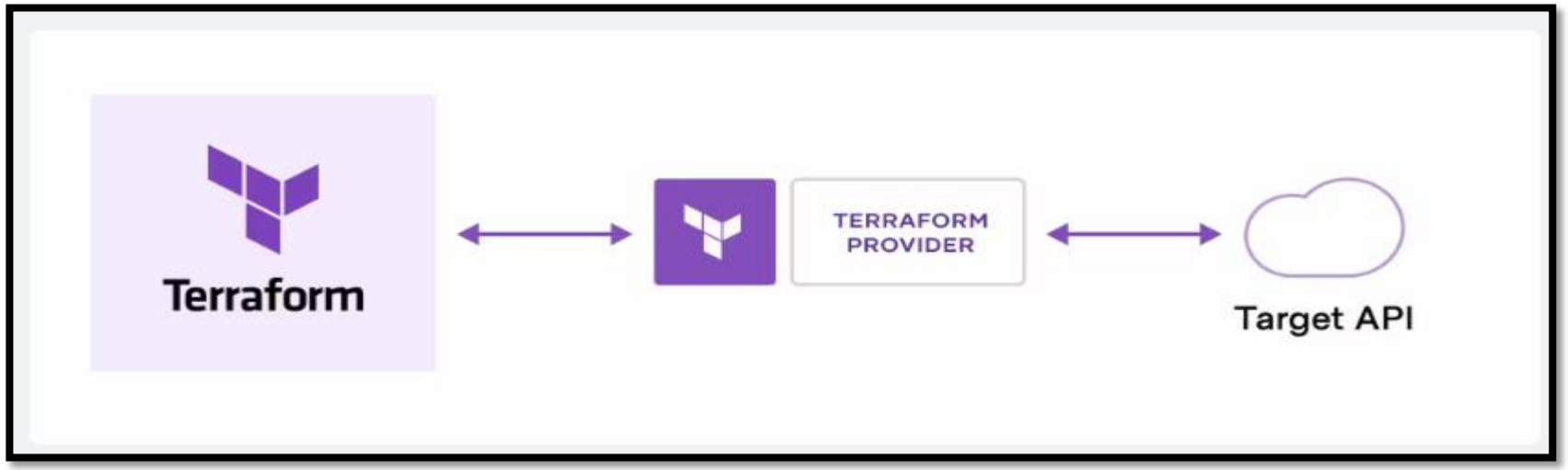
- **Choosing an IaC tool :- Mutable vs immutable infrastructure**
- Tools can also differ by what type of infrastructure they're building: **mutable or immutable**. Shifting from one to another is costly and heavy on operations, so carefully consider your needs.
- A **mutable** infrastructure can be changed to fit business needs. You can easily introduce updates to the existing version, apply patches, and scale. While this may be handy, each upgrade compromises your operation security. Besides, this may negate one of IoC purposes - consistency.
- An **immutable** infrastructure can't be changed once deployed. If you want to make alterations to it, you have to replace it with a new version. Since versions are independent, you can easily roll back, track mistakes, and enjoy consistency and predictability.
- Immutable infrastructures are considered a preferred method since it actually delivers on all the benefits of IaC.

Data Center Automation and Scaling

- **Automation tools for IaC :-**
- **Terraform**
- **AWS CloudFormation**
- **Ansible**
- **Puppet**
- **Chef**

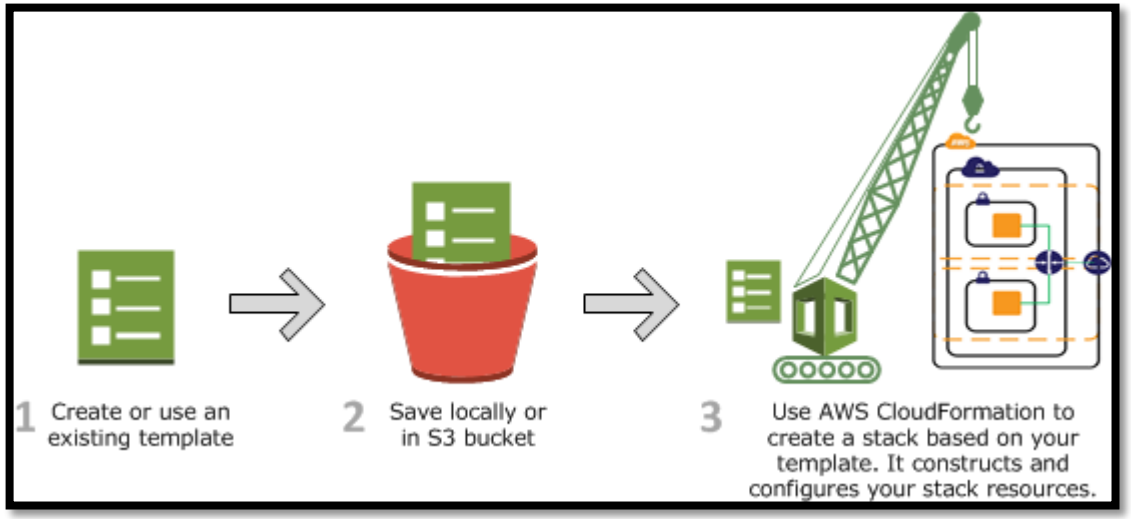
Data Center Automation and Scaling

- **Automation tools for IaC :- Terraform**
- Terraform is an open-source IaC tool developed by HashiCorp.
- It uses a declarative configuration language to define infrastructure and supports a wide range of cloud providers and on-premises infrastructure.
- Terraform creates and manages resources on cloud platforms and other services through their application programming interfaces (APIs). Providers enable Terraform to work with virtually any platform or service with an accessible API.



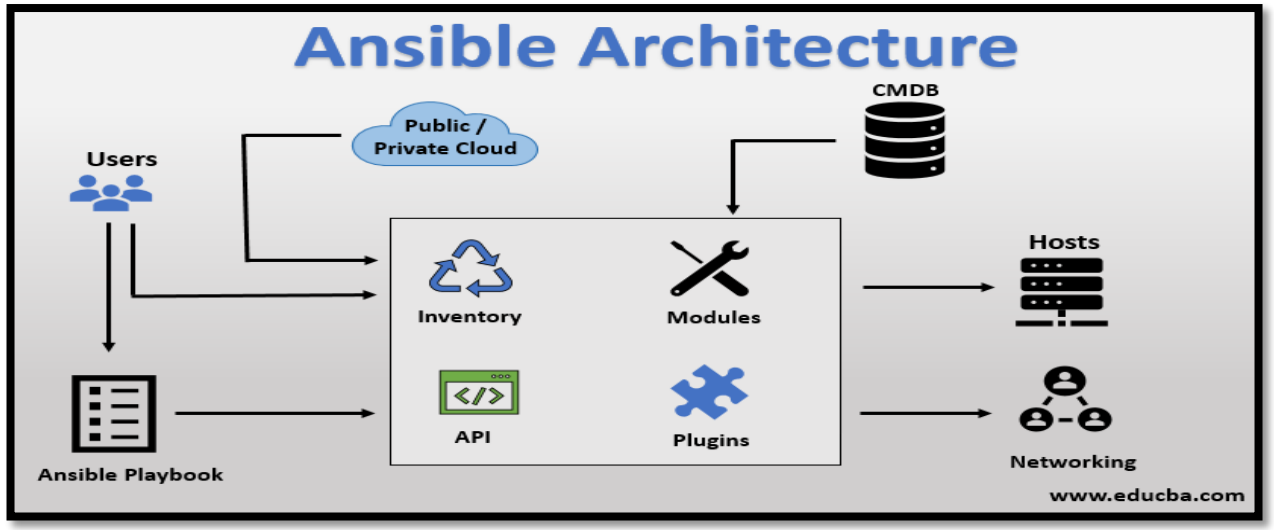
Data Center Automation and Scaling

- **Automation tools for IaC :- AWS CloudFormation**
- AWS CloudFormation is a service provided by Amazon Web Services for defining and provisioning AWS infrastructure as code.
- It uses JSON or YAML templates to describe the resources and dependencies of an AWS environment.
- Cloud Formation stacks can be versioned and changes can be applied with rolling updates or rollback capabilities.



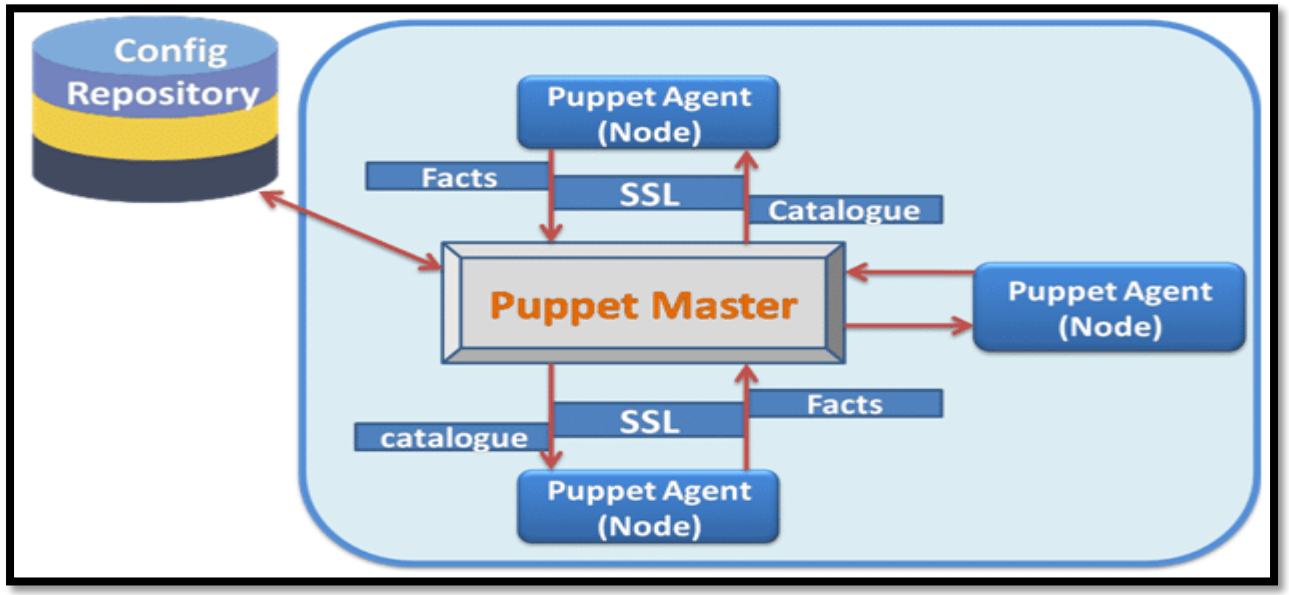
Data Center Automation and Scaling

- **Automation tools for IaC :-Ansible**
- Ansible is an open-source automation platform that supports IaC.
- It uses simple, human-readable YAML files to describe infrastructure and automation tasks, and it is known for its agentless architecture and broad support for various infrastructure components.
- It is not limited to cloud provisioning; it can automate various IT tasks, including configuration, management and application deployment.



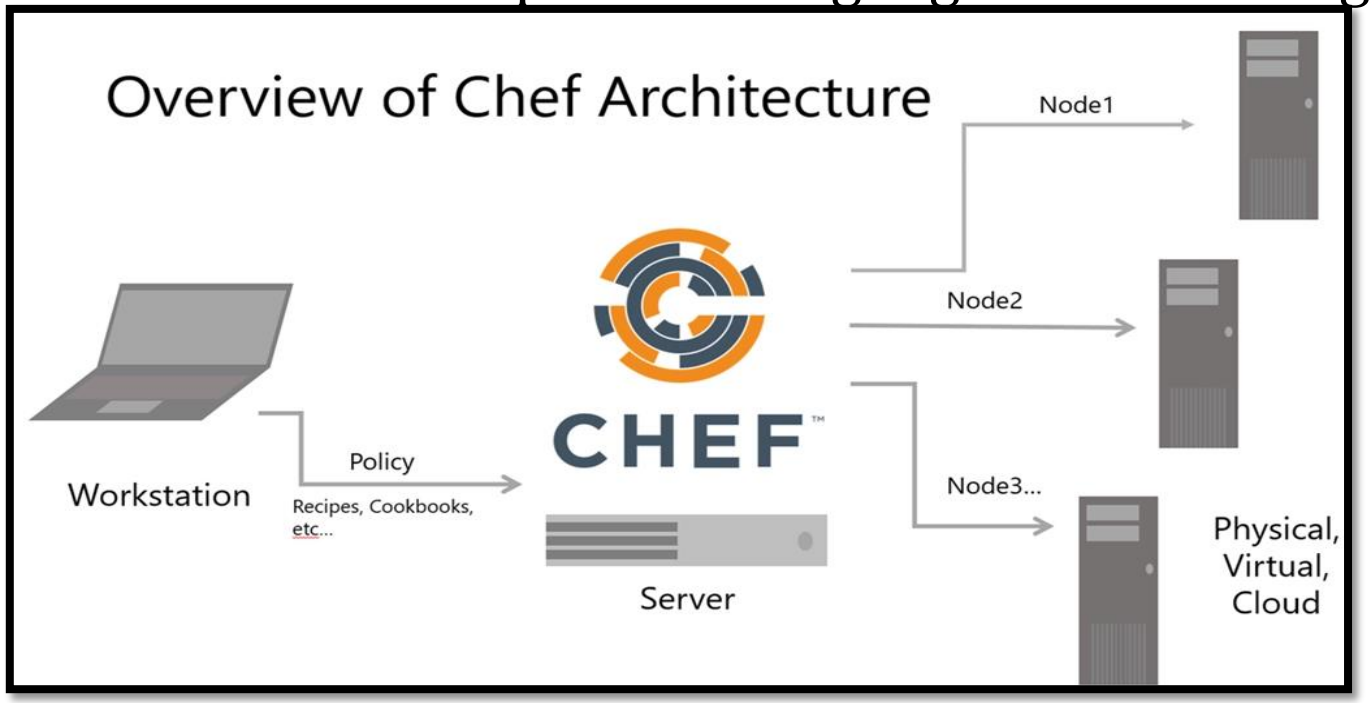
Data Center Automation and Scaling

- **Automation tools for IaC :- Puppet:-**
- Puppet is an automation tool that supports IaC by allowing the definition of infrastructure configurations in code.
- It supports both on premises and cloud environments.
- It uses declarative language to define system configuration, ensuring consistency across servers.



Data Center Automation and Scaling

- **Automation tools for IaC :- Chef:-**
- Chef is another automation tool that supports IaC practices.
- It uses code to define and automate the infrastructure, and it provides a rich set of resources for configuring and managing systems.
- It uses a Rubey-based Domain Specific Language for defining infrastructure configurations.



Data Center Automation and Scaling

- **Automation tools for IaC :-**

Name	Languages	Function	Approach	Infrastructure type
Terraform	HCL + Typescript, Python, Java, C#, Go with CDK	Provisioning	Declarative	Immutable
AWS CLOUDFORMATION	JSON, YAML + TypeScript, Python, Java, .NET, and Go with CDK	Provisioning	Declarative	Both
ANSIBLE	Python, Ruby, YAML	Configuration management	Imperative	Mutable
puppet	PuppetDSL, YAML	Configuration management	Declarative	Mutable
CHEF	Ruby	Configuration management	Declarative	Mutable

Data Center Automation and Scaling

- **Scalability in cloud data centers:-**
- Scalability in cloud data centers refers to the ability to dynamically adjust resources to accommodate changing workloads and demands.
- It enables organizations to efficiently manage their infrastructure by adding or removing resources in response to varying levels of usage.
- It is a key characteristic of cloud computing that provides flexibility and cost efficiency
- **Load Balancing:** Effective load balancing is essential for distributing workloads across multiple instances to maximize the benefits of horizontal scalability.
- **Auto-Scaling:** Leveraging auto-scaling capabilities provided by cloud platforms enables the automated addition or removal of resources based on predefined conditions, such as CPU utilization or network traffic.

Data Center Automation and Scaling

- **Scalability in cloud data centers:-**
- **Scalability in cloud data centers refers to the ability to dynamically adjust resources to accommodate changing workloads and demands.**
- It enables organizations to efficiently manage their infrastructure by adding or removing resources in response to varying levels of usage.
- It is a key characteristic of cloud computing that provides flexibility and cost efficiency
- **Load Balancing:** Effective load balancing is essential for distributing workloads across multiple instances to maximize the benefits of horizontal scalability.
- **Auto-Scaling:** Leveraging auto-scaling capabilities provided by cloud platforms enables the automated addition or removal of resources based on predefined conditions, such as CPU utilization or network traffic.

Data Center Automation and Scaling

- **Scalability in cloud data centers:-**
- **Cloud scalability is a flexible, reliable data infrastructure capable of scaling up or down in its amount of data, number of applications, and types of locations to support changing business demands and objectives.**
- Data storage capacity, processing power, and networking can all be increased by using existing cloud computing infrastructure.
- Scaling can be done quickly and easily, usually without any disruption or downtime.
- Third-party cloud providers already have the entire infrastructure in place; In the past, when scaling up with on-premises physical infrastructure, the process could take weeks or months and require exorbitant expenses.
- This is one of the most popular and beneficial features of cloud computing, as businesses can grow up or down to meet the demands depending on the season, projects, development, etc.

Data Center Automation and Scaling

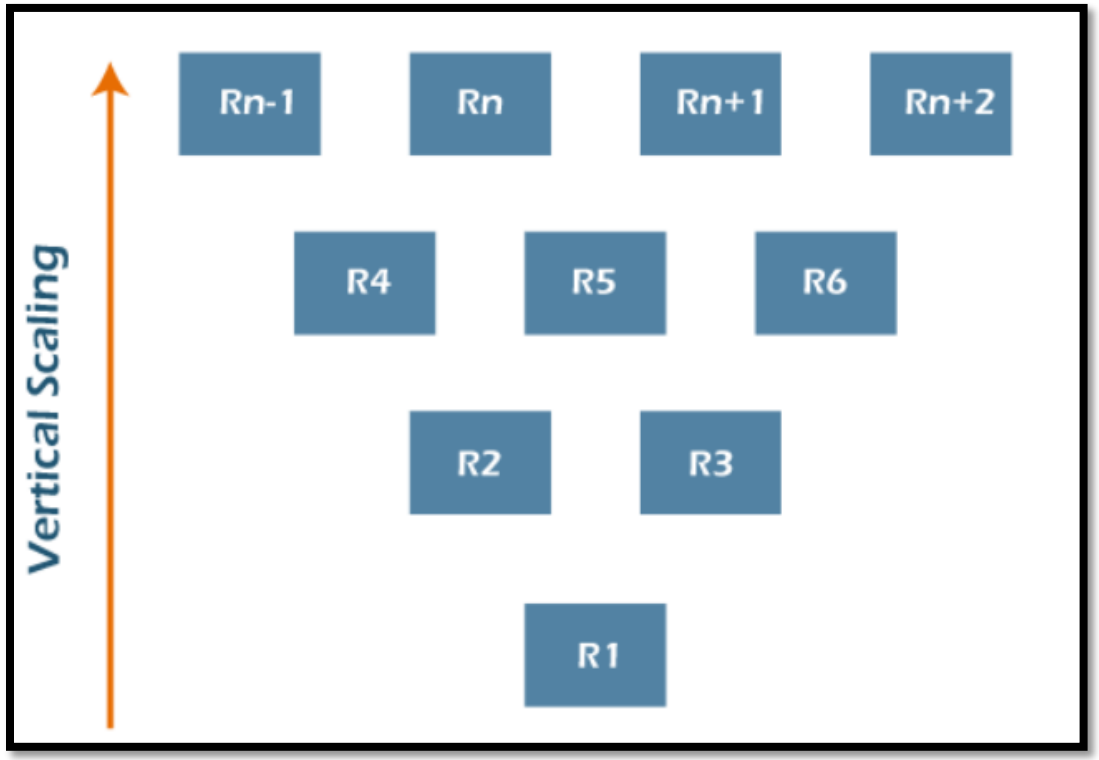
- **Scalability in cloud data centers:-**
- Types of scaling:-
- **Vertical Scalability (Scaled-up)**
- **horizontal scalability (Scaled-Out)**
- **diagonal scalability**

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Vertical Scalability (Scaled-up):-**
- Vertical scalability, also known as scale-up, involves increasing the capacity of a single resource, such as adding more memory, CPU, or storage to a virtual machine or server.
- This approach can be limited by the maximum capacity of the hardware.
- It automatically enables the system to allocate more or fewer resources to meet changing requirements.
- One of the primary benefits of vertical scalability is that it allows you to optimize your existing resources, which can help you save costs and reduce waste.

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Vertical Scalability (Scaled-up):-**



Data Center Automation and Scaling

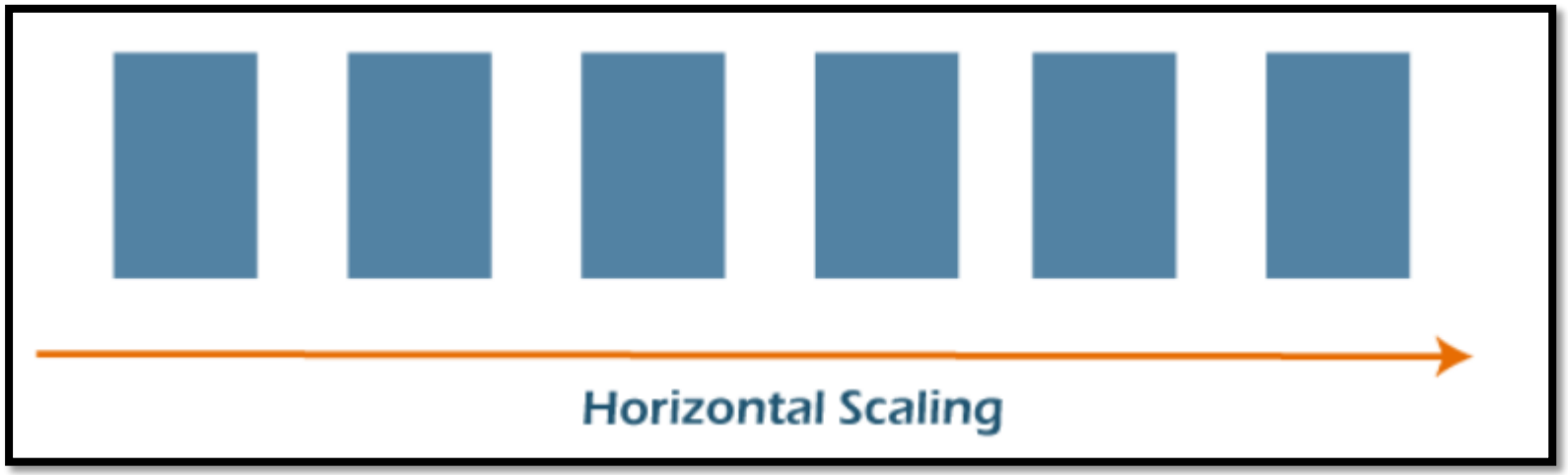
- **Types of Scalability :-**
- **Vertical Scalability (Scaled-up):-**
- **Features of Vertical Scalability:-**
- **Enhanced Performance:** Increasing the capacity of a single resource can enhance its performance and ability to handle more intensive workloads.
- **Consolidation of Resources:** Vertical scalability can be beneficial for consolidating workloads onto fewer, more powerful systems, reducing management overhead.
- **Hardware Limitations:** Vertical scalability may be limited by the maximum capacity of the underlying hardware, which can lead to constraints on scalability.

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Horizontal Scalability (Scaled-Out):-**
- Horizontal scaling refers to adding more servers to your network, rather than simply adding resources like with vertical scaling.
- This method tends to take more time and is more complex, but it allows you to connect servers together, handle traffic efficiently and execute concurrent workloads.
- Horizontal scalability involves adding more instances or nodes to distribute the workload across multiple machines.
- Instead of making individual resources larger, horizontal scalability adds more resources in parallel.
- This approach provides more flexibility and can potentially handle unlimited growth by adding more resources as needed.
- Horizontal scalability requires the application to be designed in a way that allows it to be distributed across multiple nodes.

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Horizontal Scalability (Scaled-Out):-**



Data Center Automation and Scaling

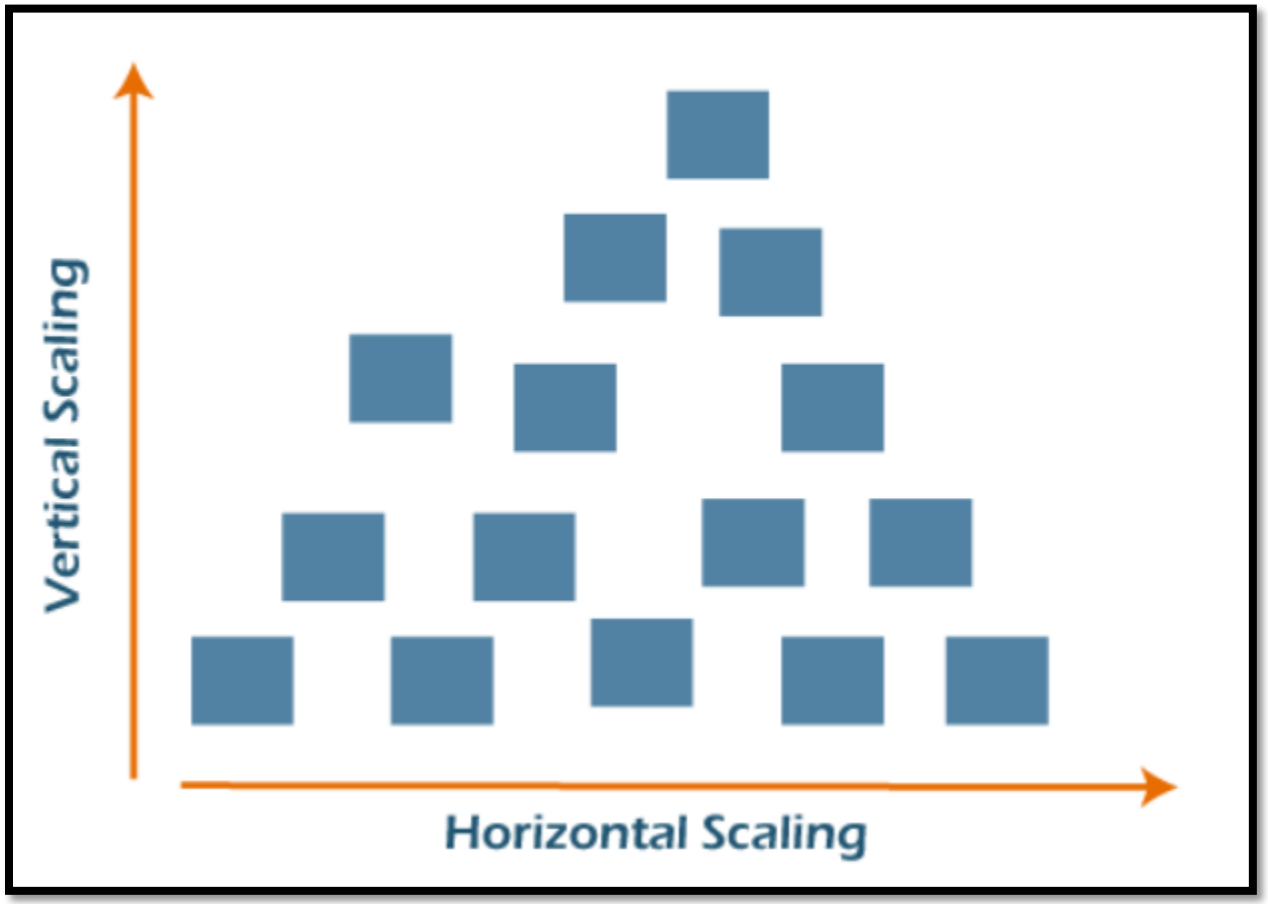
- **Types of Scalability :-**
- **Horizontal Scalability (Scaled-out):-**
- **Features of Horizontal Scalability:-**
- **Increased Performance and Workload Distribution:-** Horizontal scaling allows for improved performance by distributing the workload across multiple machines, thereby reducing the burden on individual resources.
- **Redundancy and Fault Tolerance:-** By adding multiple instances of resources, horizontal scaling provides redundancy and fault tolerance. If one node fails, the remaining instances can continue to handle the workload, enhancing system reliability.
- **Ease of Expansion:-** It offers the ability to easily expand the system by adding more instances or nodes as the workload increases, allowing for seamless scalability.
- **Cost-Effectiveness:-** Horizontal scaling can be cost-effective as it typically involves using commodity hardware and adding more instances as needed, rather than investing in larger, more expensive individual resources.

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Diagonal Scalability:-**
- It is a mixture of both Horizontal and Vertical scalability where the resources are added both vertically and horizontally.
- Diagonal scaling, which allows you to experience the most efficient infrastructure scaling.
- When you combine vertical and horizontal, you simply grow within your existing server until you hit the capacity.
- Then, you can clone that server as necessary and continue the process, allowing you to deal with a lot of requests and traffic concurrently.
- This type of scalability allows you to add more instances or resources as needed while also optimizing your existing resources to achieve maximum efficiency.
- Hybrid scalability is often used for complex applications that require a combination of processing power, storage, and bandwidth.

Data Center Automation and Scaling

- **Types of Scalability :-**
- **Diagonal Scalability:-**



Data Center Automation and Scaling

- **Benefits of cloud Scalability :-**
- **Convenience:** Often, with just a few clicks, IT administrators can easily add more VMs that are available-and customized to an organization's exact needs-without delay.
- Teams can focus on other tasks instead of setting up physical hardware for hours and days. This saves the valuable time of the IT staff.
- **Flexibility and speed:** As business needs change and grow, including unexpected demand spikes, cloud scalability allows IT to respond quickly. Companies are no longer tied to obsolete equipment-they can update systems and easily increase power and storage.
- **Cost Savings:** Thanks to cloud scalability, businesses can avoid the upfront cost of purchasing expensive equipment that can become obsolete in a few years.
- Through cloud providers, they only pay for what they use and reduce waste.
- **Disaster recovery:** With scalable cloud computing, you can reduce disaster recovery costs by eliminating the need to build and maintain secondary data centers.

Data Center Automation and Scaling

- **Use of cloud Scalability :-**
- Successful businesses use scalable business models to grow rapidly and meet changing demands. It's no different with their IT.
- Cloud scalability benefits help businesses stay agile and competitive.
- Scalability is one of the driving reasons for migrating to the cloud.
- Whether traffic or workload demands increase suddenly or increase gradually over time, a scalable cloud solution enables organizations to respond appropriately and cost-effectively to increased storage and performance.
- Scalable cloud architecture is made possible through virtualization.
- Unlike physical machines whose resources and performance are relatively set, virtual machines virtual machines (VMs) are highly flexible and can be easily scaled up or down.
- They can be moved to a different server or hosted on multiple servers at once; workloads and applications can be shifted to larger VMs as needed.

Data Center Automation and Scaling

- **Elasticity in Cloud Data Center :-**
- **Elasticity** refers to the ability of a cloud to automatically expand or compress the infrastructural resources on a **sudden up and down** in the requirement so that the workload can be managed efficiently.
- This elasticity helps to minimize infrastructural costs.
- This is not applicable for all kinds of environments, it is helpful to address only those scenarios where the resource requirements **fluctuate up and down** suddenly for a specific time interval.
- It is not quite practical to use where persistent resource infrastructure is required to handle the heavy workload.
- Cloud elasticity enables you to access more resources when necessary and release them when they are no longer needed.

Data Center Automation and Scaling

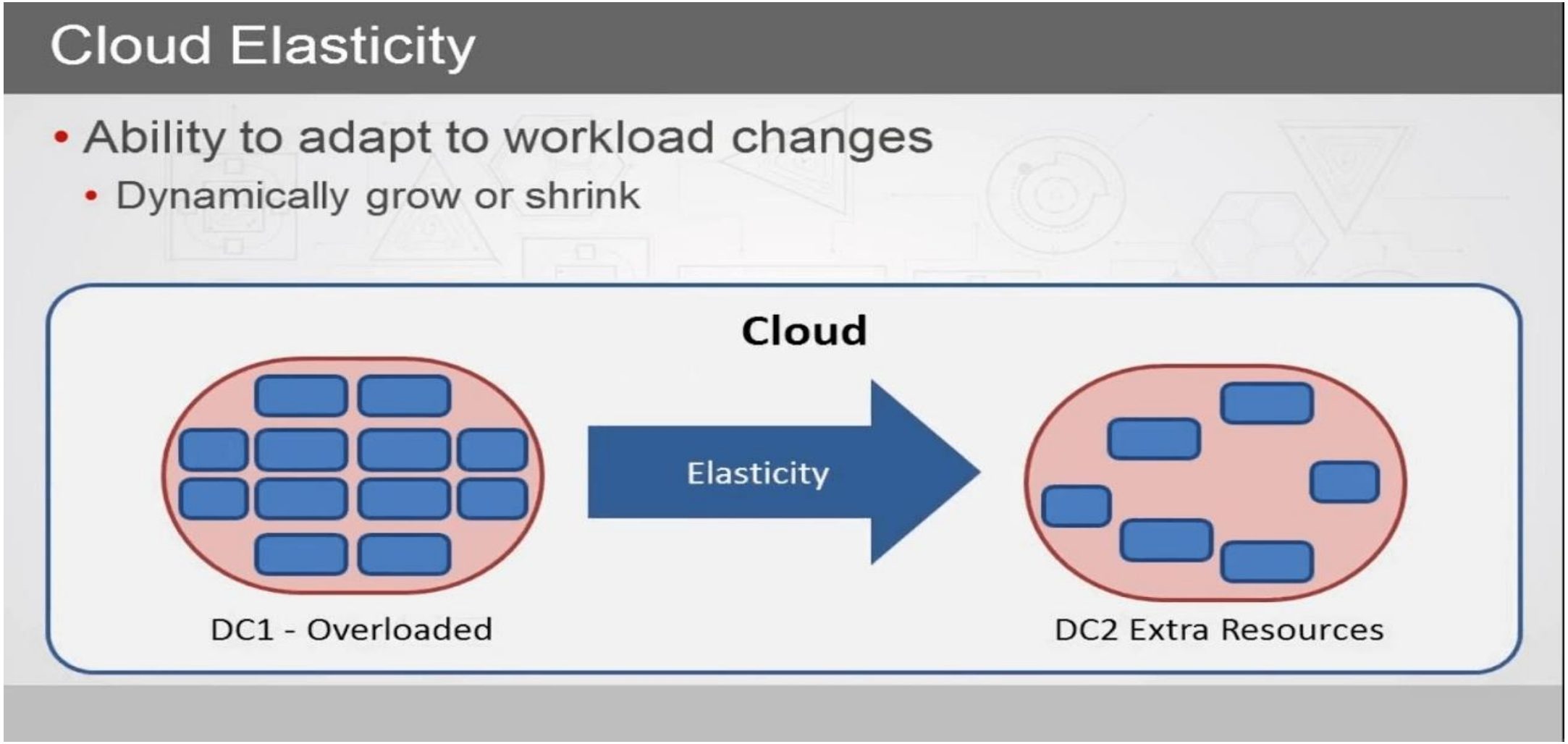
- **Elasticity in Cloud Data Center :-**
- The elasticity process often needs to happen quickly.
- A delay in increasing capacity could overload your system, potentially causing service outages.
- In contrast, if you delay shrinking, some of your servers will have little to do. But you'd still have to pay for the idle capacity, which is a waste of your cloud budget.
- Elastic environments match resource allocation to dynamic workloads, allowing you to take up more resources or release those you no longer need. If the process occurs quickly or in real time, it is called rapid elasticity.

Data Center Automation and Scaling

- **Elasticity in Cloud Data Center :-**
- **Example:**
- Consider an online shopping site whose transaction workload increases during festive season like Christmas.
- So for this specific period of time, the resources need a spike up. In order to handle this kind of situation, we can go for a Cloud-Elasticity service rather than Cloud Scalability.
- As soon as the season goes out, the deployed resources can then be requested for withdrawal.

Data Center Automation and Scaling

❑ **Elasticity in Cloud Data Center :-**



Data Center Automation and Scaling

- **Key characteristics of Elasticity in Cloud Data Center :-**

- **Auto Scaling -** Cloud resources can scale out and in automatically based on criteria like workload demand, utilization metrics, etc. The scaling is handled programmatically without manual intervention.

- **Rapid Elasticity -** Scaling up and down can happen very quickly, in minutes or even seconds in some cases. This allows capacity to closely match sporadic changes in demand.

- **Fine-grained Control -** Scaling can happen at a very granular level, such as adding/removing individual VMs, storage blobs, containers, etc. instead of monolithic blocks.

- **Flexible Capacity -** Elasticity allows capacity to expand or shrink based on actual real-time resource needs. There are no fixed limits on how much you can scale.

- **Usage-based Costing -** You only pay for the resources you currently use. Elasticity enables releasing unneeded resources and saves money during low demand periods.

Data Center Automation and Scaling

Cloud Elasticity	Cloud Scalability
Elasticity is used just to meet the sudden up and down in the workload for a small period of time.	Scalability is used to meet the static increase in the workload.
Elasticity is used to meet dynamic changes, where the resources need can increase or decrease.	Scalability is always used to address the increase in workload in an organization.
Elasticity is commonly used by small companies whose workload and demand increases only for a specific period of time.	Scalability is used by giant companies whose customer circle persistently grows in order to do the operations efficiently.
Elasticity is automated.	Scalability is usually done manually
Elasticity is reactive and responds to real-time changes in demand.	Scalability uses a predictive approach based on known demand changes.
Elasticity only incurs costs for resources currently being used.	Scalability incurs costs even when excess resources are idle.
It is a short term planning and adopted just to deal with an unexpected increase in demand or seasonal demands.	Scalability is a long term planning and adopted just to deal with an expected increase in demand.