ADC501

Cloud Computing

Module 4

Cloud Deployment Techniques

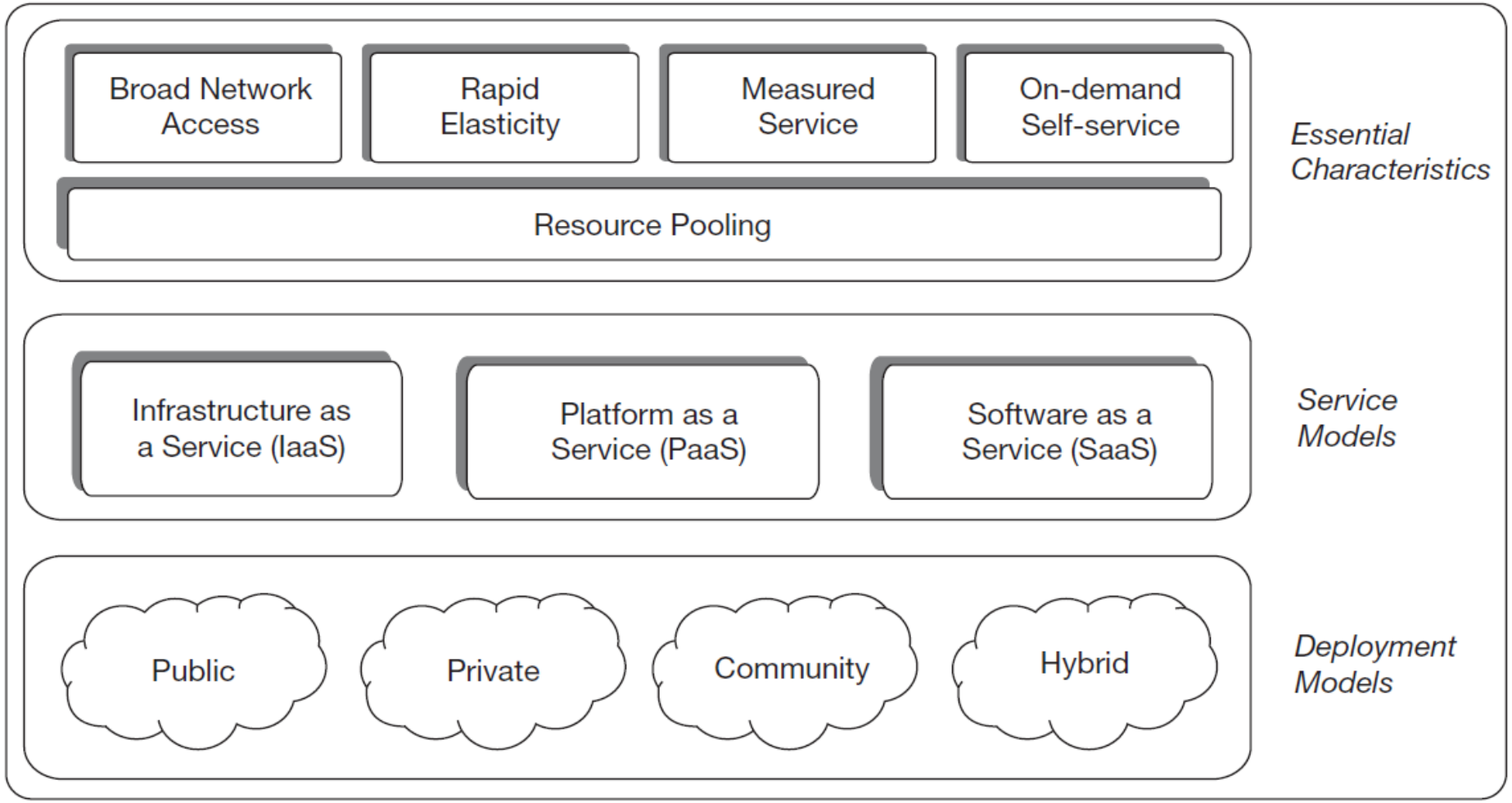


FIG 4.1: The NIST cloud computing model

NIST defines cloud computing by describing five essential characteristics, three cloud service models and four cloud deployment models.

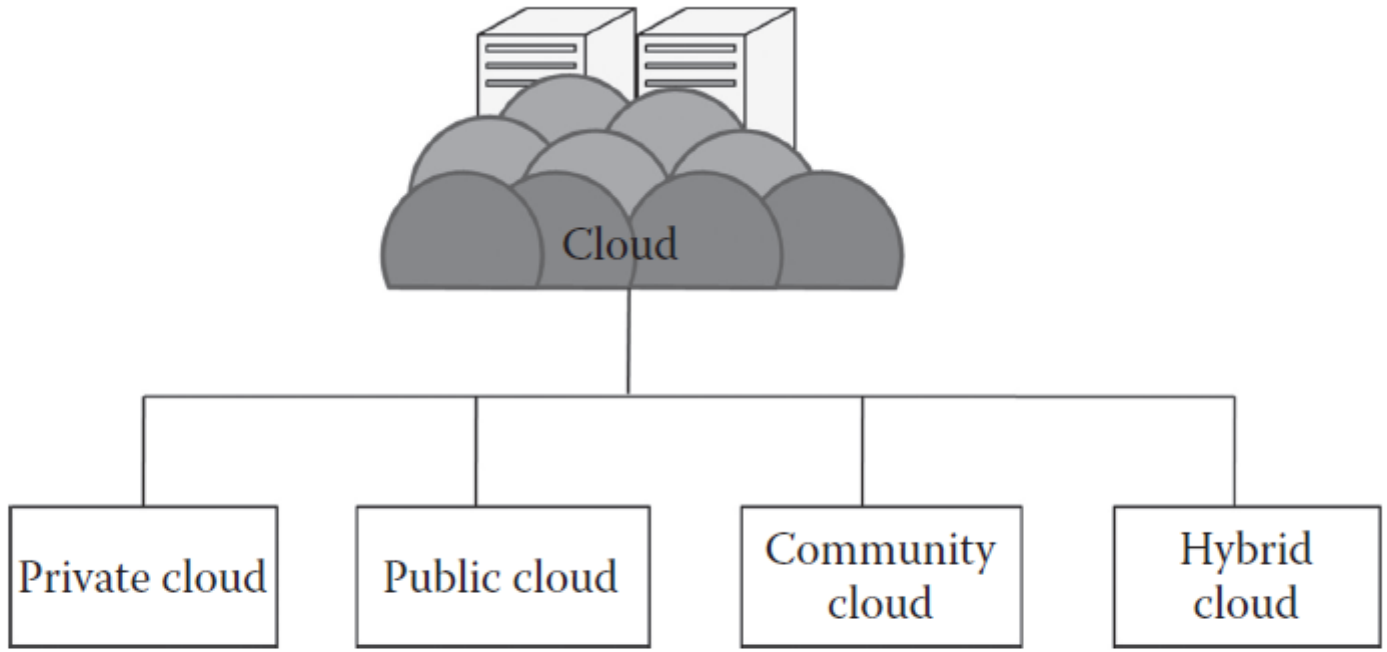


FIGURE 4.1
Cloud deployment models.

Four types of deployment models in the cloud

1. **Private** cloud
   - most basic deployment model
   - deployed by a single organization for its personal use
   - not shared by other organizations
   - it is not allowed for public use

2. **Community** cloud
   - extension of the private cloud
   - mutual benefits among the participating organizations

3. **Public** cloud
   - allows access from any place in the world and is open to the public
   - biggest in size among all
   - public cloud service provider charges the users on an hourly basis and serve the users according to the service-level agreements (SLAs)

4. **Hybrid** cloud
   - combination
   - Usually, it consists of the private and public clouds combined
   - one of the upcoming cloud models growing in the industry

Private Cloud

According to the National Institute of Standards and Technology (NIST) - private cloud can be defined as

the cloud infrastructure that is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units)

It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises

Private cloud can be deployed using Opensource tools such as Openstack , Eucalyptus .

private cloud is small in size

Characteristics of Private Cloud

- **Secure**
- **Central control**
- **Weak SLAs**

  high availability and good service may or may not be available
  depends on the organization that is controlling the cloud

Suitability

- The organizations or enterprises that require a separate cloud for their personal or official use.

- The organizations or enterprises that have a sufficient amount of funds as managing and maintaining a cloud is a costly affair.

- The organizations or enterprises that consider data security to be important.

- The organizations that want autonomy and complete control over the cloud.

- The organizations that have a less number of users.

- The organizations that have prebuilt infrastructure for deploying the cloud and are ready for timely maintenance of the cloud for efficient functioning.

- Special care needs to be taken and resources should be available for troubleshooting.

private cloud platform is not suitable for the following

- The organizations that have high user base
- The organizations that have financial constraints
- The organizations that do not have prebuilt infrastructure
- The organizations that do not have sufficient manpower to maintain and manage the cloud

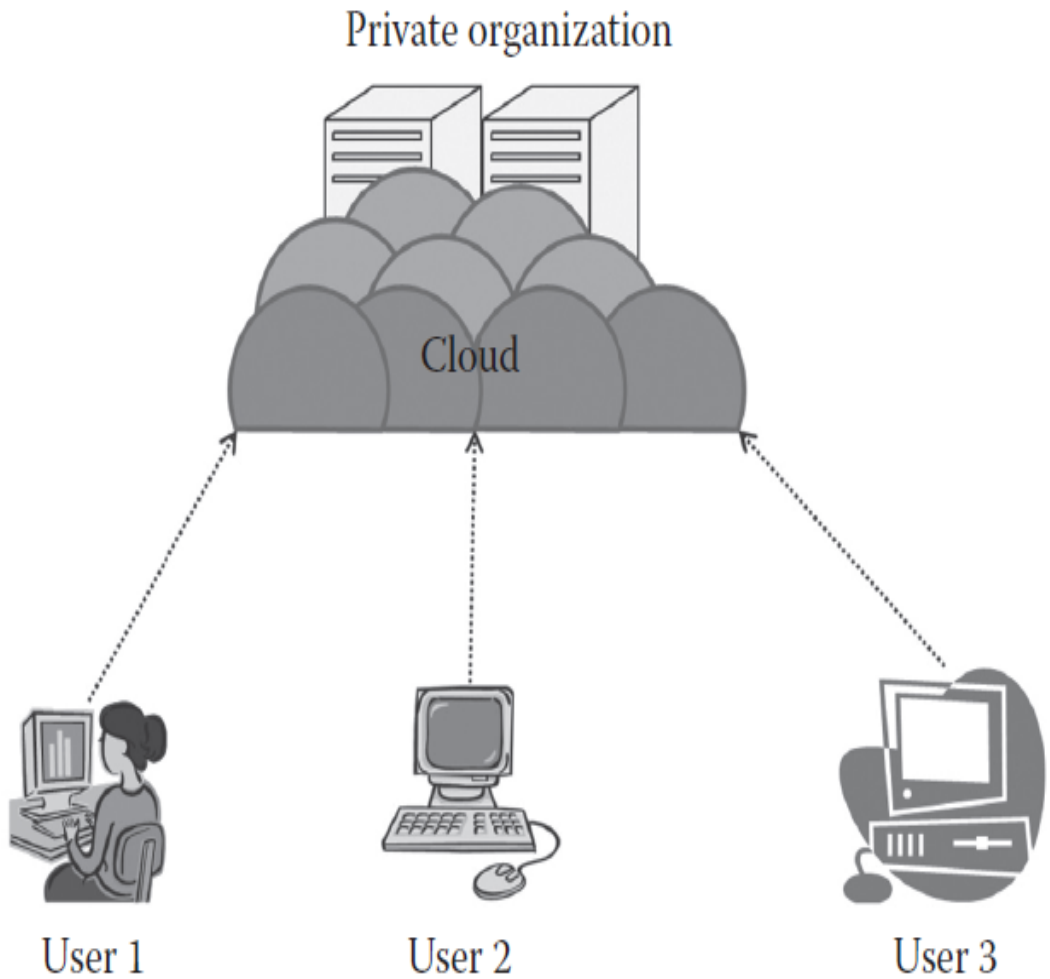


FIGURE 4.2
On-premise private cloud.

private cloud can be classified into several types based on their location and management

- On-premise private cloud

Issues

SLA - Usually, these users have broader access rights than the general public cloud users
service providers are able to efficiently provide the service because of the small user base and mostly efficient network

Network - high bandwidth and has a low latency Network management is easier in this case, and resolving a network issue is easier

Performance - good performance

Security and data privacy – private cloud is more resistant to attacks

Location - If a company has several physical locations, then the cloud is distributed over several places , virtual private network [VPN]

Cloud management - resource scheduling, resource provisioning, and resource management

Multitenancy - less effect

Maintenance - easy

Outsourced private cloud

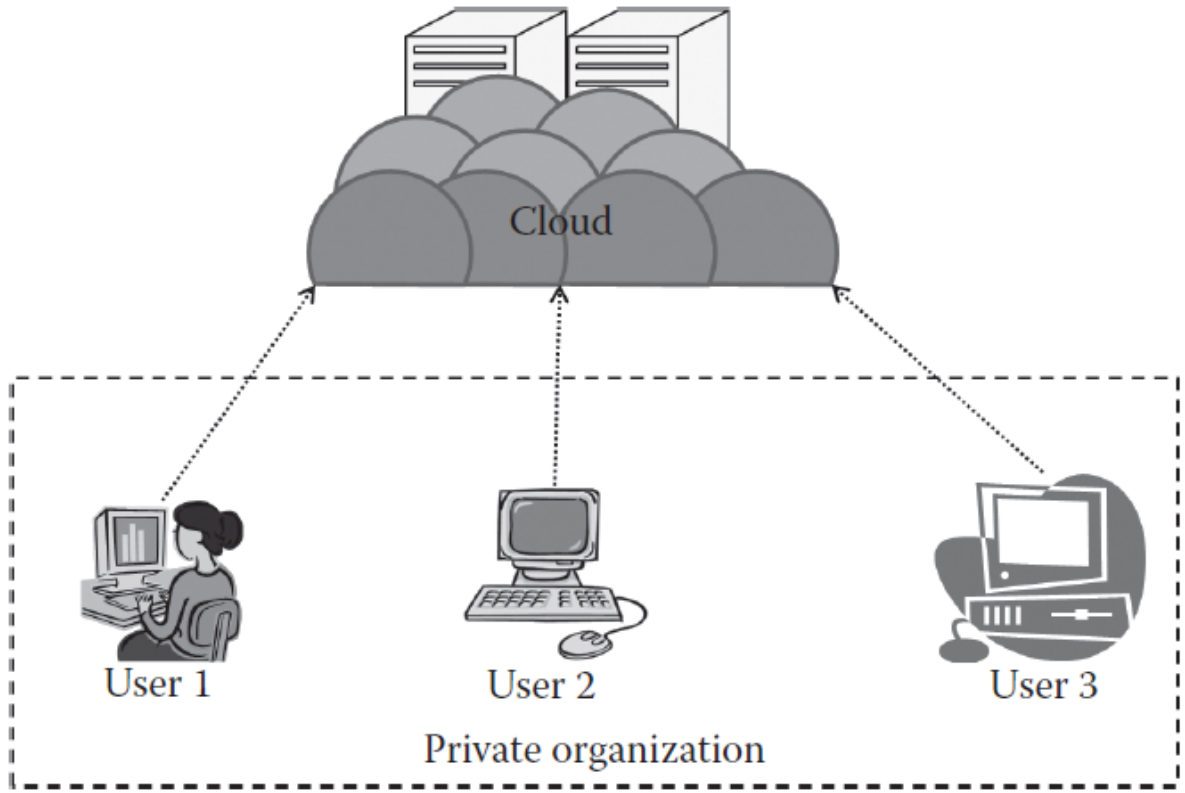


FIGURE 4.3
Outsourced private cloud.

The issues that are specific to outsourced private cloud

- SLA - The SLAs are usually followed strictly as it is a third-party organization
- Network
- Security and privacy
- Laws and conflicts - If this cloud is deployed outside the country
- Location
- Performance
- Maintenance

- **Advantages of Private Cloud**

➢ The cloud is small in size and is easy to maintain

➢ It provides a high level of security and privacy to the user

➢ It is controlled by the organization

- **Disadvantages of Private Cloud**

➢ For the private cloud, budget is a constraint.

➢ The private clouds have loose SLAs.

Community Cloud

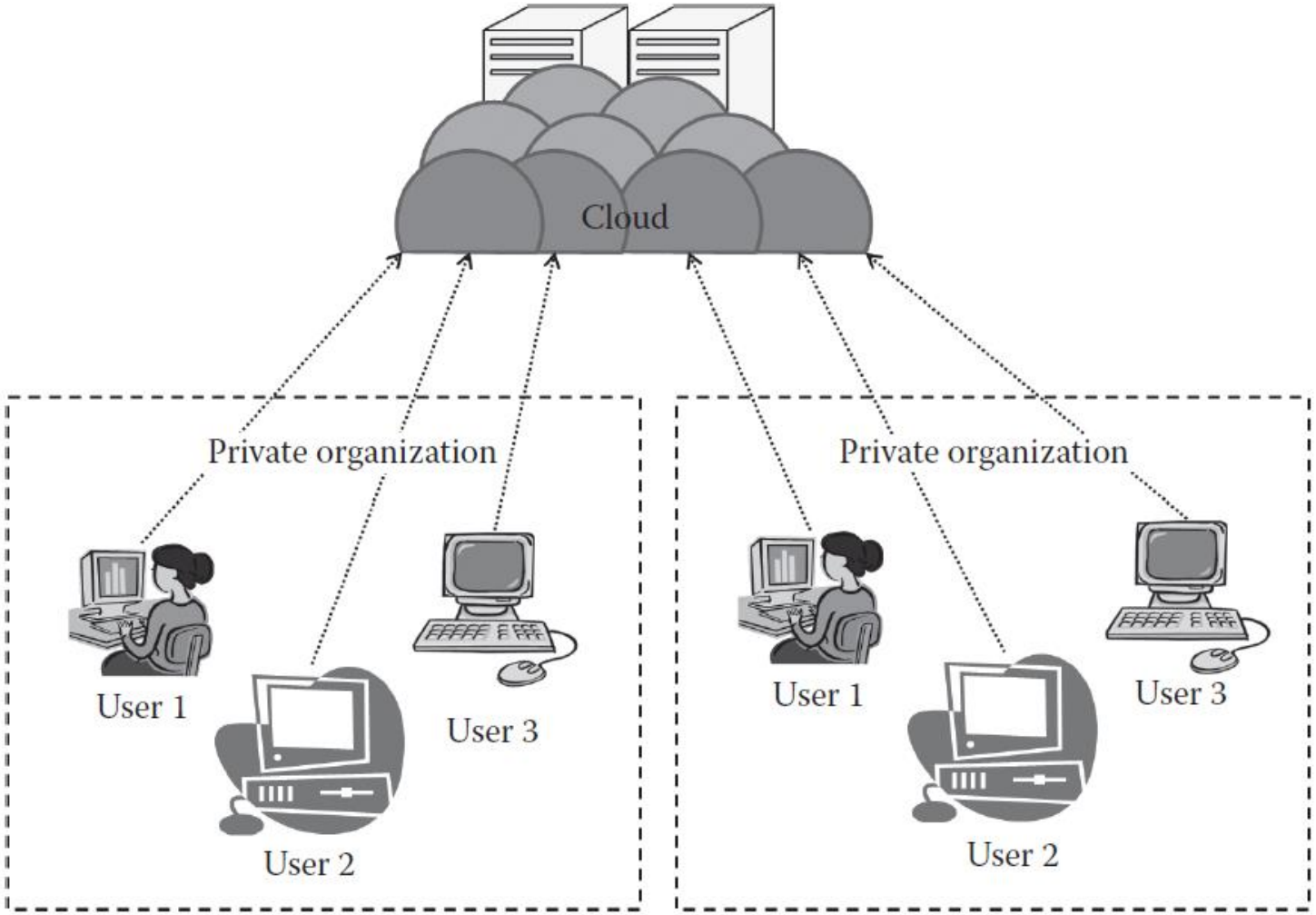


FIGURE 4.5
Community cloud.

Community Cloud

- According to NIST, the community cloud is the cloud infrastructure that is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations)

- further extension of the private cloud

- This model is very suitable for organizations that cannot afford a private cloud and cannot rely on the public cloud either

Characteristics of Community Cloud

- Collaborative and distributive maintenance

- Partially secure

- Cost effective

Suitability of Community Cloud

- Want to establish a private cloud but have financial constraint
- Do not want to complete maintenance responsibility of the cloud
- Want to establish the cloud in order to collaborate with other clouds
- Want to have a collaborative cloud with more security features than the public cloud

This cloud is **not suitable** for organizations that

- Prefer autonomy and control over the cloud
- Does not want to collaborate with other organizations

Sr. No.	Issues	On-premise community cloud	Outsourced community cloud
1.	SLA	more stringent than the private cloud but is less stringent than the public cloud	SLA here is stringent as it involves a third party
2.	Network	each organization will have a separate network, and they will connect to the cloud, It is the responsibility of each organization to take care of their own network, The network is not big and complex as in the public cloud	organizations are responsible for their own network and the service provider is responsible for the cloud network
3.	Performance	it is on the maintenance and management team that the performance depends	The service provider is responsible for efficient services, except for the network issue in the client side
4.	Multitenancy	unprivileged access into interorganizational data may lead to several problems	
5.	Location	organizations have to access the cloud from another location	
6.	Security and privacy	the situation is more like that of a public cloud with less users	
7.	Laws and conflicts	This applies if organizations are located in different countries	If the service provider is outside the country, then there is conflict related to data laws in that country
8.	Cloud management	Cloud management is done by the organizations collectively	The complexity of managing and maintenance increases with the number of organizations in the community
9.	Cloud maintenance	Cloud maintenance is done by the organizations collectively	

Advantages of Community Cloud

- It allows establishing a low-cost private cloud.
- It allows collaborative work on the cloud.
- It allows sharing of responsibilities among the organization.
- It has better security than the public cloud.

Disadvantages

- Autonomy of an organization is lost.
- Security features are not as good as the private cloud.
- It is not suitable if there is no collaboration.

Public Cloud

According to **NIST**, the public cloud is the cloud infrastructure that is provisioned for open use by the general public.

- exists on the premises of the cloud provider
- Public cloud consists of users from all over the world
- A user can simply purchase resources on an hourly basis and work with the resources
- no need of any prebuilt infrastructure

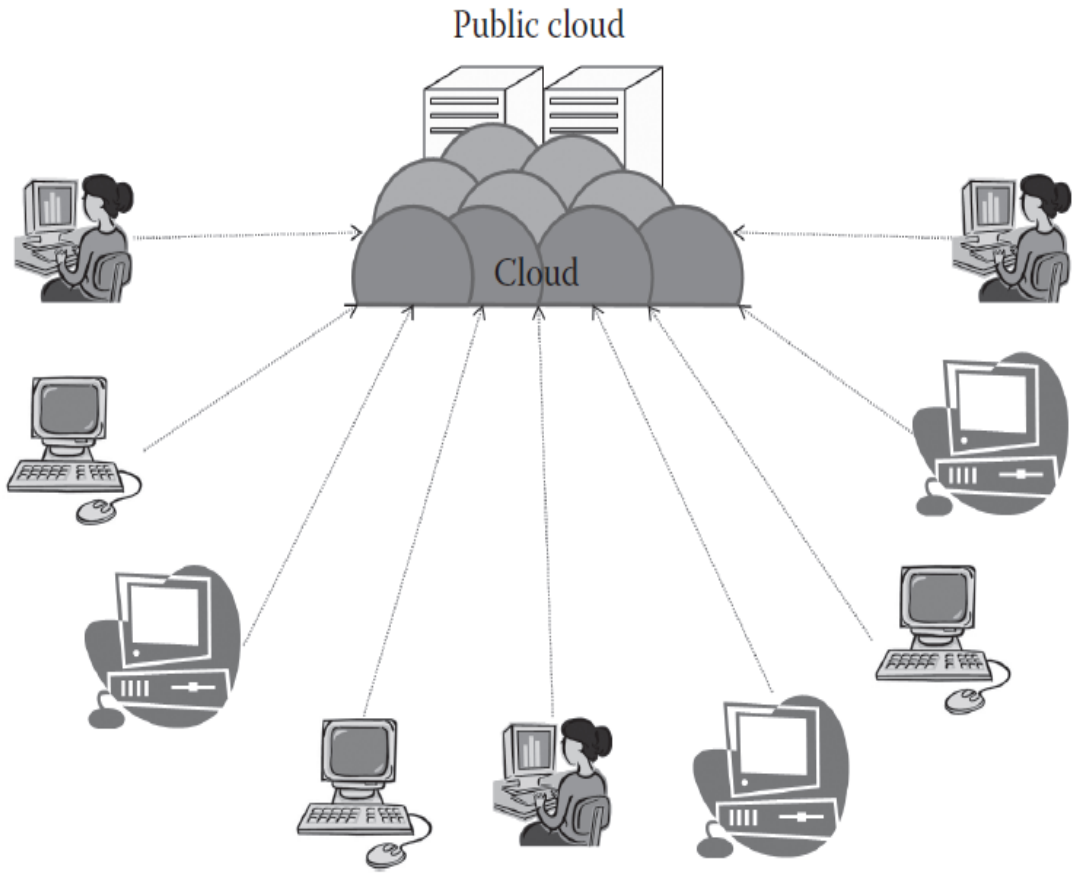


FIGURE 4.4
Public cloud.

Characteristics of Public Cloud

- Highly scalable - The resources in the public cloud are large in number and the service providers make sure that all the requests are granted
- Affordable - pay-as-you-go basis (does not involve any cost related to the deployment)
- Less secure
- Highly available
- Stringent SLAs - SLAs are very competitive

Suitability of Public Cloud

- The requirement for resources is large, that is, there is large user base.
- The requirement for resources is varying.
- There is no physical infrastructure available.
- An organization has financial constraints.

The public cloud is **not suitable**

- Security is very important.
- Organization expects autonomy.
- Third-party reliability is not preferred.

Issues of Public Cloud

- SLA - number of users is more and so are the numbers of service agreements , users here are diverse
- Network
- Performance - As the number of users increases, it is a challenging task for the service providers to give good performance
- Multitenancy- high risk of data being leaked or a possible unprivileged access
- Location - As the public cloud is fragmented and is located in different regions, the access to these clouds involves a lot of data transfers through the Internet
- Security and data privacy
- Laws and conflicts
- Cloud management
- Maintenance

Advantages of Public Cloud

- There is no need of establishing infrastructure for setting up a cloud.
- There is no need for maintaining the cloud.
- They are comparatively less costly than other cloud models.
- Strict SLAs are followed.
- There is no limit for the number of users.
- The public cloud is highly scalable.

Disadvantages of Public Cloud

- Security is an issue.
- Privacy and organizational autonomy are not possible**.**

Hybrid Cloud

- According to NIST, the hybrid cloud can be defined as the cloud infrastructure that is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability

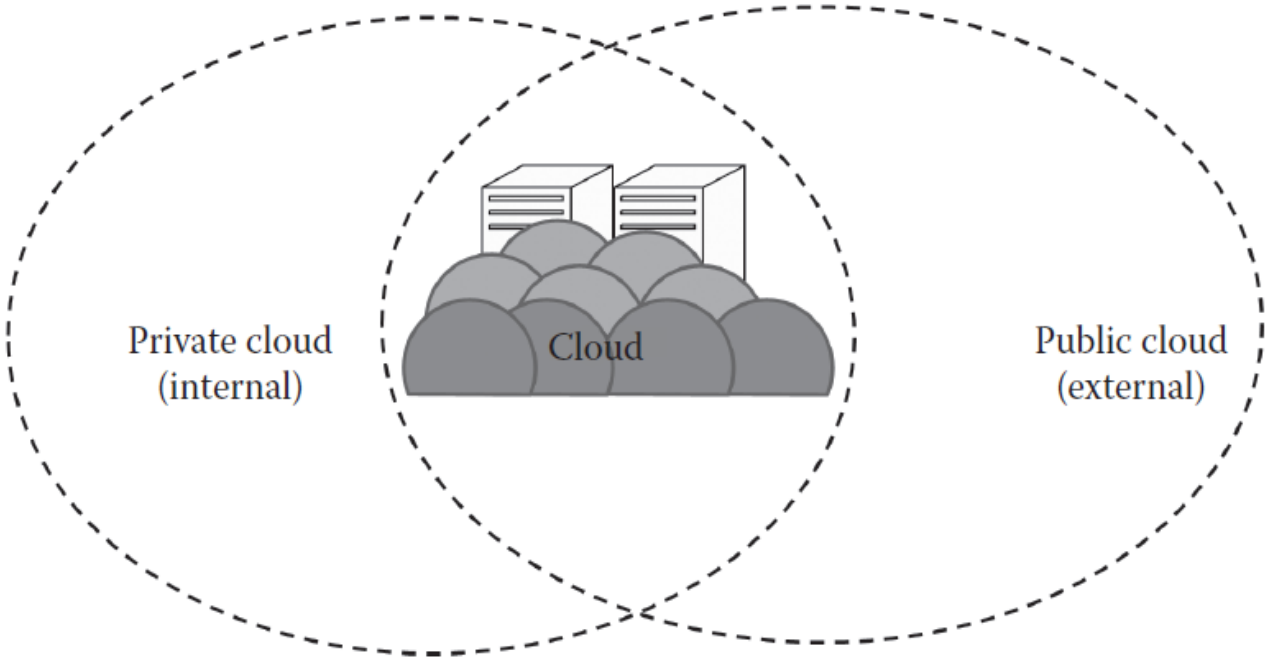


FIGURE 4.6
Hybrid cloud.

Characteristics of Hybrid Cloud

- Scalable - as the public cloud is scalable, the hybrid cloud with the help of its public counterpart is also scalable
- Partially secure
- Stringent SLAs
- Complex cloud management

Suitability of Hybrid Cloud

The hybrid cloud environment is suitable for

- Organizations that want the private cloud environment with the scalability of the public cloud
- Organizations that require more security than the public cloud

The hybrid cloud is **not suitable** for

- Organizations that consider security as a prime objective
- Organizations that will not be able to handle hybrid cloud management

Issues of Hybrid Cloud

- SLA - The private cloud does not have stringent agreements, whereas the public cloud has certain strict rules to be covered, There is a right combination of SLAs between the clouds
- Network - The organization takes the responsibility from the network
- Performance
- Multitenancy
- Location
- Security and privacy
- Laws and conflicts
- Cloud management
- Cloud maintenance - It involves a high cost of maintenance

Advantages of Hybrid Cloud

- It gives the power of both the private and public clouds.
- It is highly scalable.
- It provides better security than the public cloud.

Disadvantages

- The security features are not as good as the public cloud.
- Managing a hybrid cloud is complex.
- It has stringent SLAs.

Cloud Architecture

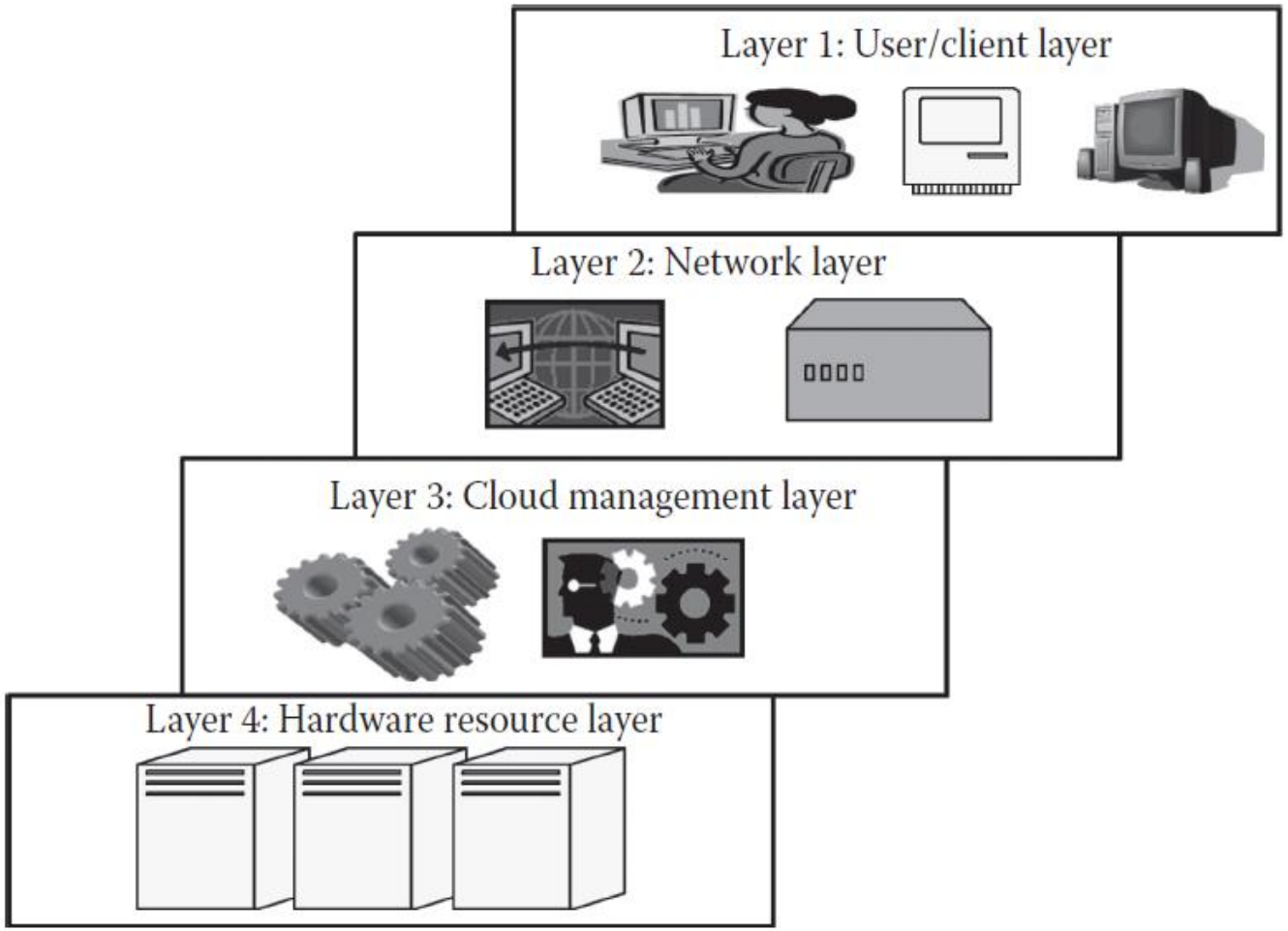


FIGURE 3.1
Cloud architecture.

Cloud Architecture

- **Layer 1 (User/Client Layer)**

The client can be any device such as a thin client, thick client, or mobile or any handheld device that would support basic functionalities to access a web application.

- **Layer 2 (Network Layer)**

The whole cloud infrastructure is dependent on this connection where the services are offered to the customers
private cloud - local area network (LAN)
This layer does not come under the purview of service-level agreements (SLAs),

- **Layer 3 (Cloud Management Layer)**

softwares that are used in managing the cloud
cloud operating system (OS), resource management (scheduling, provisioning, etc.), optimization (server consolidation, storage workload consolidation), and internal cloud governance
This layer comes under the purview of SLAs

Cloud Architecture

- **Layer 4 (Hardware Resource Layer)**

data center, which is a huge collection of hardware resources interconnected to each other that is present in a specific location or a high configuration system

This layer comes under the purview of SLAs

the data center consists of a high-speed network connection and a highly efficient algorithm to transfer the data from the data center to the manager.

There can be a number of data centers for a cloud, and similarly, a number of clouds can share a data center

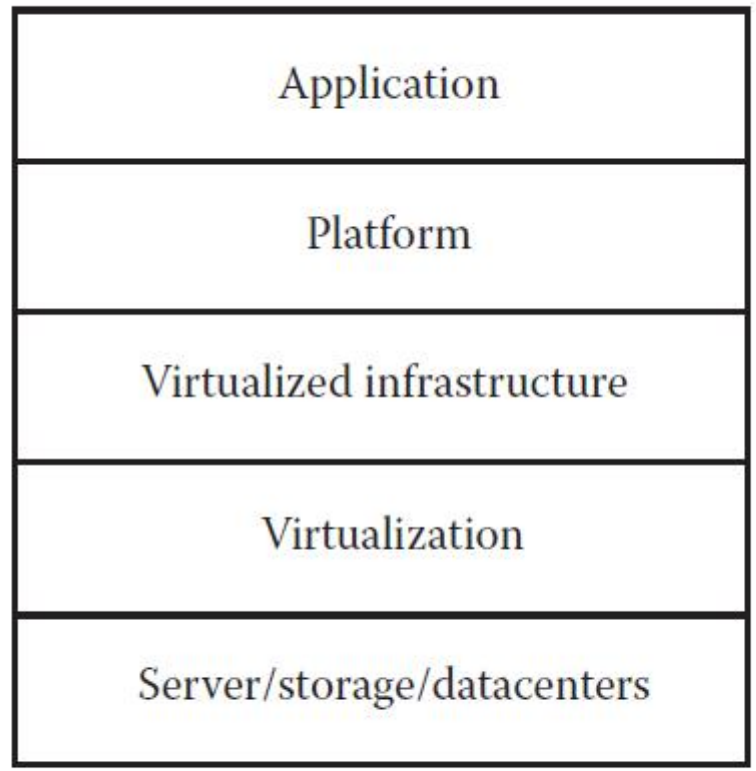
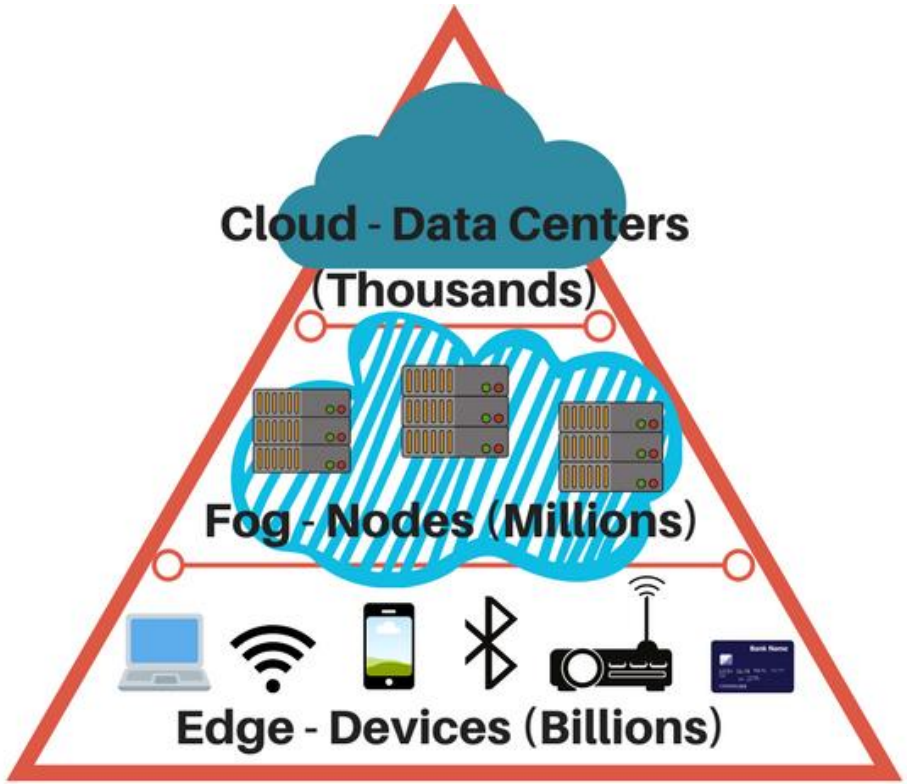


FIGURE 3.2
Cloud structure.

Mobile Cloud Computing (MCC)

- mobile computing, cloud computing and wireless communication
- IoT – Smart Devices- EDGE/FOG computing
- explosive growth in uses of mobile devices
- laptop, notebook, tablet and mobile phone, IoT Devices
- battery-powered and enabled for wireless communication
- *Mobile communication* technology – 2G, 3G, 4G, 5G….
- *radio wave transmission* through towers installed around the world



- *mobile computing* is a technology where computing-capable devices can transmit data without being connected to any device or network physically, even when in transit.

Limitations of Mobile Cloud Computing (MCC)

- MCC- location and connectivity issues are critical factors to be considered every time a task is to be executed
- Performance concerns for resource-constrained mobile devices
- highly-configured mobile devices which increases the cost
- lightweight and small enough to make themselves portable
- This limits many attributes of mobile computing devices like processing capability, storage capacity and battery lifetime
- *vendor-lock-in* problem - a particular platform (like Symbian, Android, Windows etc.)

Mobile Cloud Computing (MCC)

- ***task offloading***

- resolves the application performance and data storage issues but the energy issue also as processing of tasks consume battery power heavily

- conserving mobile energy (battery power) and extending storage and enhancing data security

- mobile devices mainly play the roles of human-computer interfacing with less processing burden

- Initially the process of offloading used to be happen from mobile devices to some stationary computer connected through network or Internet

- The computing systems having rich resources used in the offloading process were referred as *surrogate systems*

- Lesser processing obligation improves energy conservation in mobile devices which is a critical factor in mobile computing technology

Mobile Cloud Computing (MCC)

- Integration of Cloud Computing into Mobile Computing

- cloud computing enables it to provide high-performance computational delivery by using easily available and cheaper commodity hardware

- reduces the computational cost and makes cloud computing cost-effective

- Cloud provider ensures consistent performance of application irrespective of the number of concurrent users of the application at any moment

- Cloud computing also offers attractive data storage facility

Benefits of Mobile Cloud Computing

- Extended life of battery
- Improved processing performance
- Enhanced storage capacity
- Improved reliability
- Enhanced data security - data is replicated over multiple locations
- Relief from vendor-lock-in

Networking in Cloud Enabled Data Centers (CEDCs)

- Data centers are used in two major applications –
- the ones that aim to provide online services to users, for example, Google, Facebook, and Yahoo
- to provide resources to users, for example, Amazon Elastic Compute Cloud (EC2) and Microsoft Azure

- Based on the fault-tolerance capacity and service uptime, today's data centers are classified into four tiers

FIGURE 4.8

A huge data center that is 11 times the size of a football field, housing 400,000 to 1 million servers.

(*Courtesy of Dennis Gannon [26]*)

- The cloud is built on massive data centers.
- as large as a shopping mall (11 times the size of a football field) under one roof
- 400,000 to 1 million servers

TABLE 10.1

Classification of Data Centers

Tiers	Features	Uptime (%)
I	Nonredundant capacity components (single uplinks and servers)	99.671
II	Tier I + redundant capacity components	99.741
III	Tier I + Tier II + dual-powered equipments and multiple links	99.982
IV	Tier I + Tier II + Tier III + all components are fault tolerant including uplinks, storage, HVAC systems, servers + everything is dual powered	99.995

- Tier IV data center is considered to be most robust and less prone to failures.
- designed to host mission critical servers
- computer systems with fully redundant subsystems (cooling, power, network links, storage, etc.)
- compartmentalized security zones controlled by biometric access control methods
- The simplest is Tier I data center, which is usually used by small shops

- HVAC – heating , ventilation & air-conditioning

Architecture of Classical Data Centers

- *multitier* model

- Typically, the following three tiers are used -

1. Web server
2. Application
3. Database

- Resiliency is improved

- Security is improved

Physical Organization

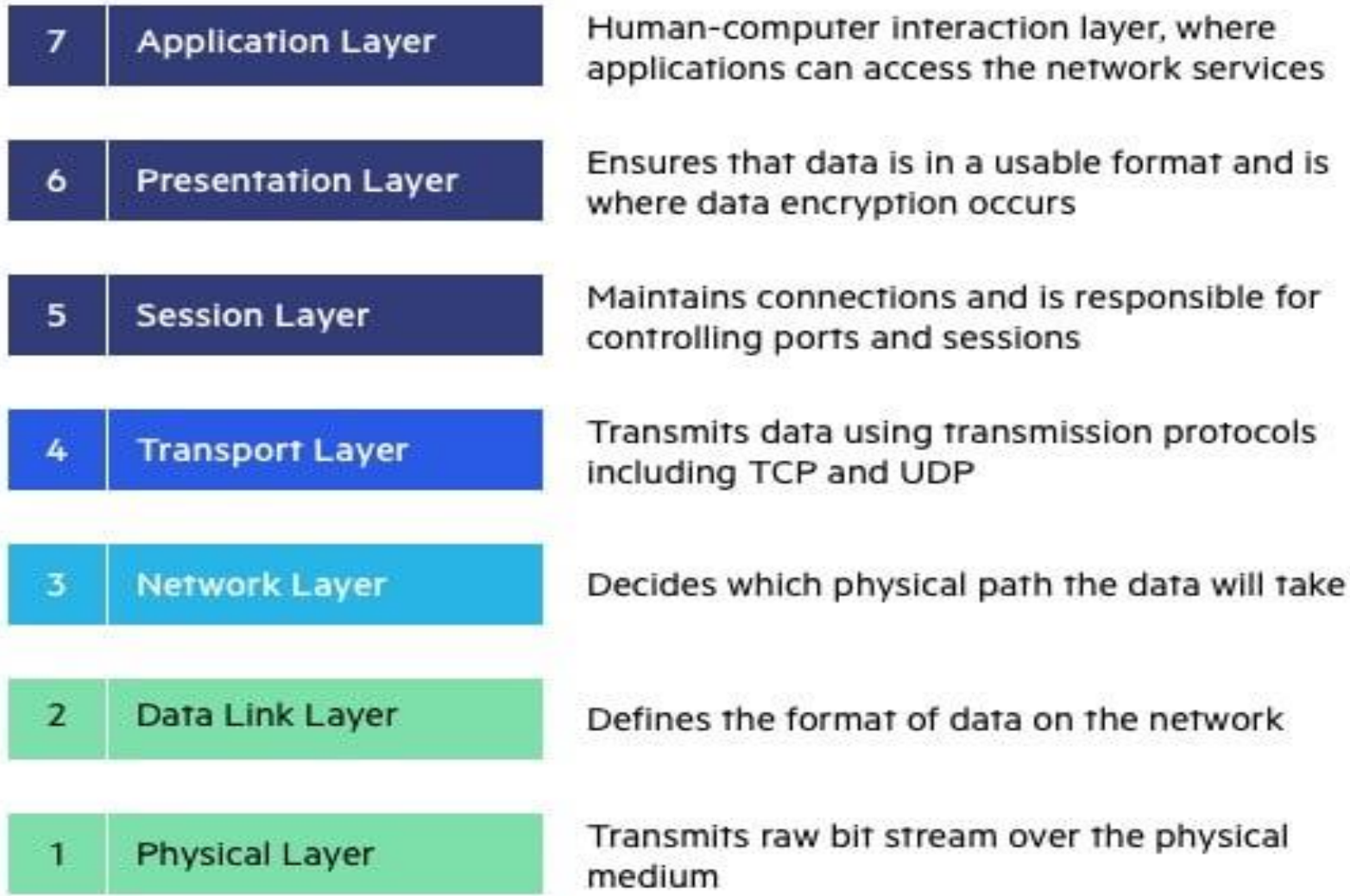
- the server racks in a data center
- large number of servers that are mounted in rack cabinets
- placed in single rows forming corridors between them, so as to allow access
- a few equipments such as storage devices are often as large as the racks – placed alongside racks
- In the event of a failure or when upgrades are required, the entire containers are replaced rather than replacing an individual server

FIGURE 10.1
Physical organization of a data center.

Storage and Networking Infrastructure

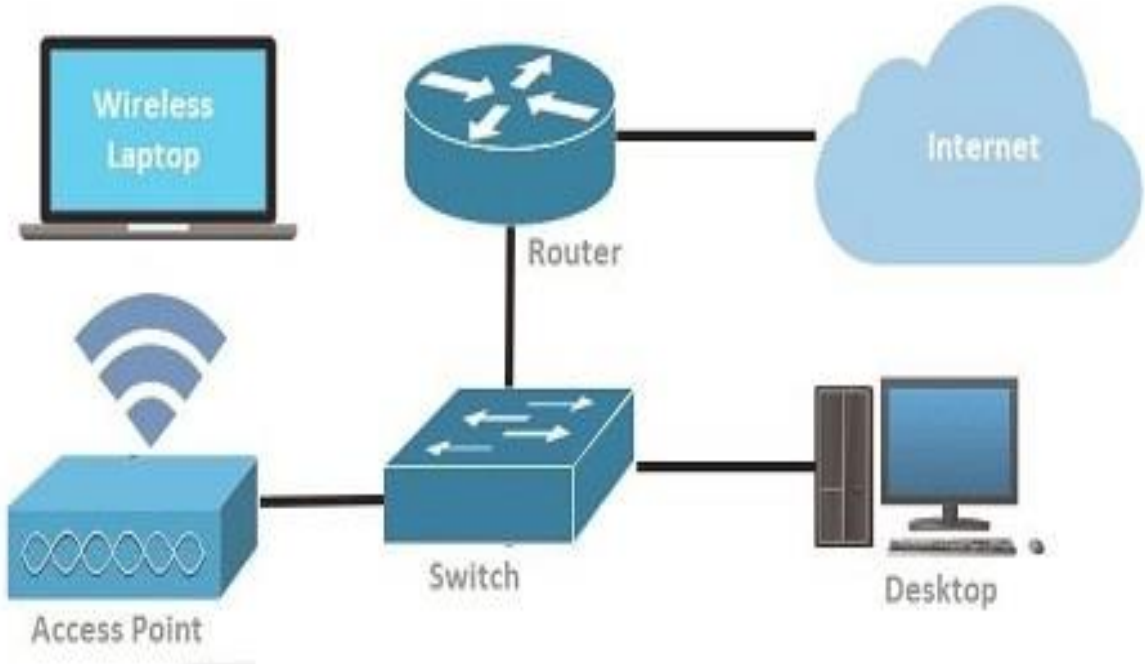
- data centers require four different types of network accesses
- could use four different types of physical networks
- 1. Client–server network: To provide external connectivity to the data center - Traditional wired Ethernet or Wireless LAN technologies
- 2. Server–server network: To provide high-speed communication among the servers of the data center , Ethernet, InfiniBand (IBA), figure 10.2
- 3. Server–storage network: To provide high-speed connectivity between the servers and storage devices, Usually Fiber Channel is used
- 4. Other networks - network required to manage the data center , Ethernet is used but the cabling may be different from the mainstream networks

FIGURE 10.2
Networking infrastructure in data centers.

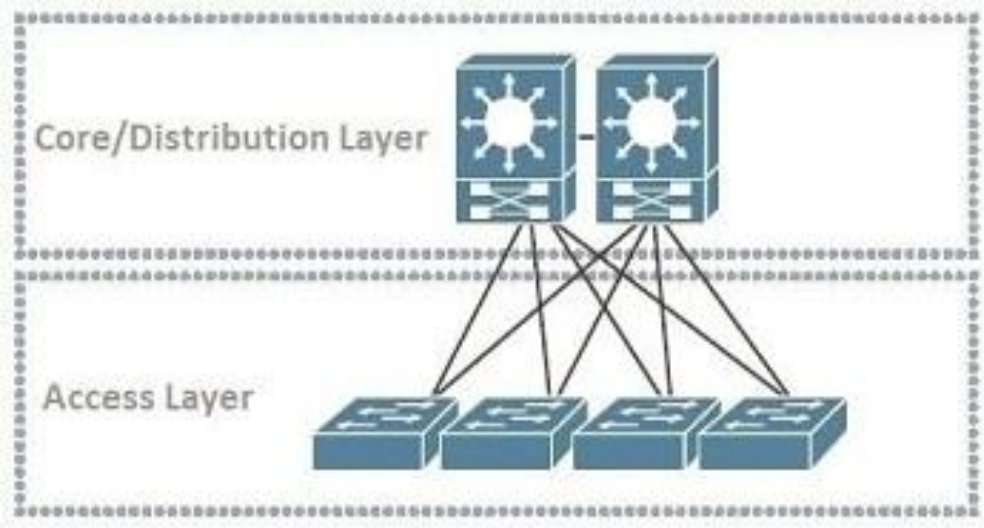


The Open Systems Interconnection (OSI) model describes seven layers that computer systems use to communicate over a network

Small Office/Home Office (SOHO) Network Topology

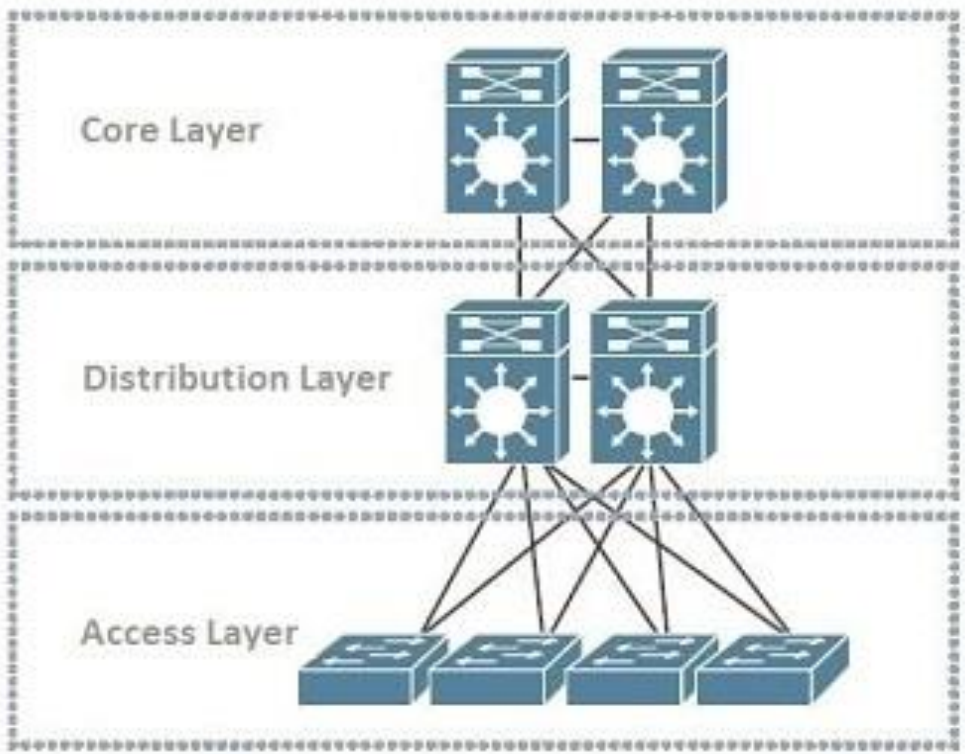


- A **two-tier network topology** is very popular in DCNs today
- network is smaller
- scalability and complexity are not much concern
- In a small office network



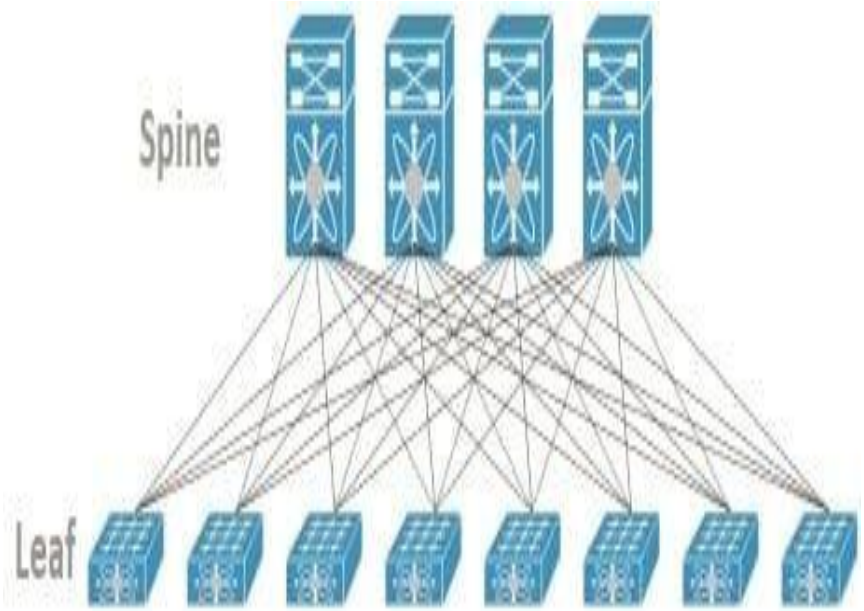
Three-tier Network Topology

- better scalability, flexibility, and network segmentation
- In an enterprise network



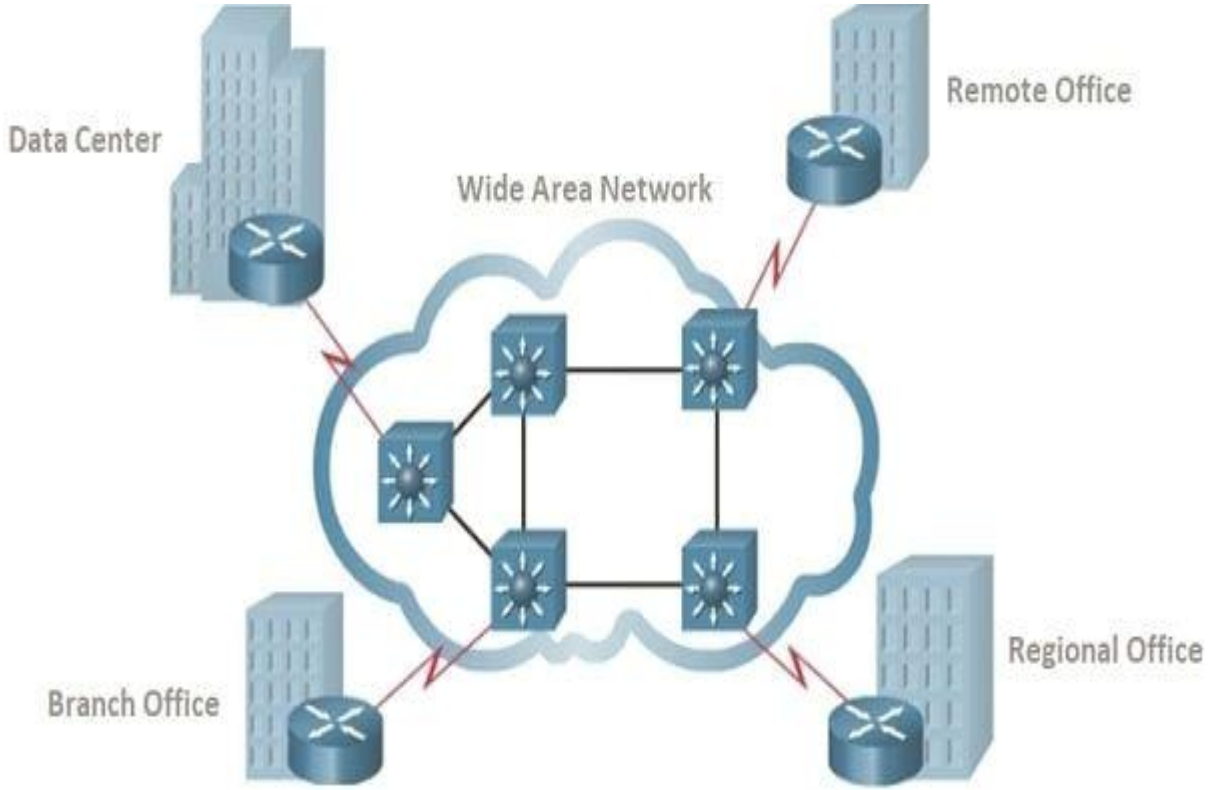
Spine-Leaf Network Topology

- highly scalable and high-performance
- frequently employed in large data centers or cloud environments.
- facilitates low-latency and non-blocking communication among devices, ensuring efficient and rapid data transmission
- connecting servers or storage devices
- ensures that any device in the network can reach any other device with minimal latency



WAN (Wide Area Network) Network Topology

- multi-site organization
- to interconnect geographically dispersed locations or branch offices
- thousands of branches
- leased lines, MPLS (Multi-Protocol Label Switching), VPN (Virtual Private Network), or SD-WAN (Software-Defined Wide Area Network)



Nature of Traffic in Data Centers

TABLE 10.3

Data Center Traffic: Applications and Performance Requirements

Traffic Type	Examples	Requirements
Mice traffic (<100 kB)	Google search, Facebook	Short response times
Cat traffic (100 kB to 5 MB)	Picasa, YouTube, Facebook photos	Low latency
Elephant traffic (>5 MB)	Software updates, video on demand	High throughput

Source: Reproduced from Kant, K., *Comput. Netw.*, 53(17), 2939, December 2009.

- the round-trip time (RTT) in DCNs can be as less as 250 μs in the absence of queuing

Networking Issues in Data Centers

- **Availability**

  to provide maximum uptime for the services that are offered to the users

  downtime may lead to violation of service-level agreements (SLAs) between the cloud user and the cloud provider

  largely affecting the cloud provider's revenues

  to replicate the data and take regular backups

- **Poor Network Performance**

  three basic performance requirements of a DCN are high burst tolerance, low latency, and high throughput

- **Security**

  Keeping a cloud user's data secure during transit, or while it is at rest

Transport Layer Issues in DCNs

- success of the Internet, in fact, can be partly attributed to the congestion control mechanisms implemented in TCP

- TCP has evolved to keep up with the changing network conditions and has proven to be scalable and robust

- challenges faced by the state-of-the-art TCP in DCNs
  - TCP Incast - synchronized mice collide

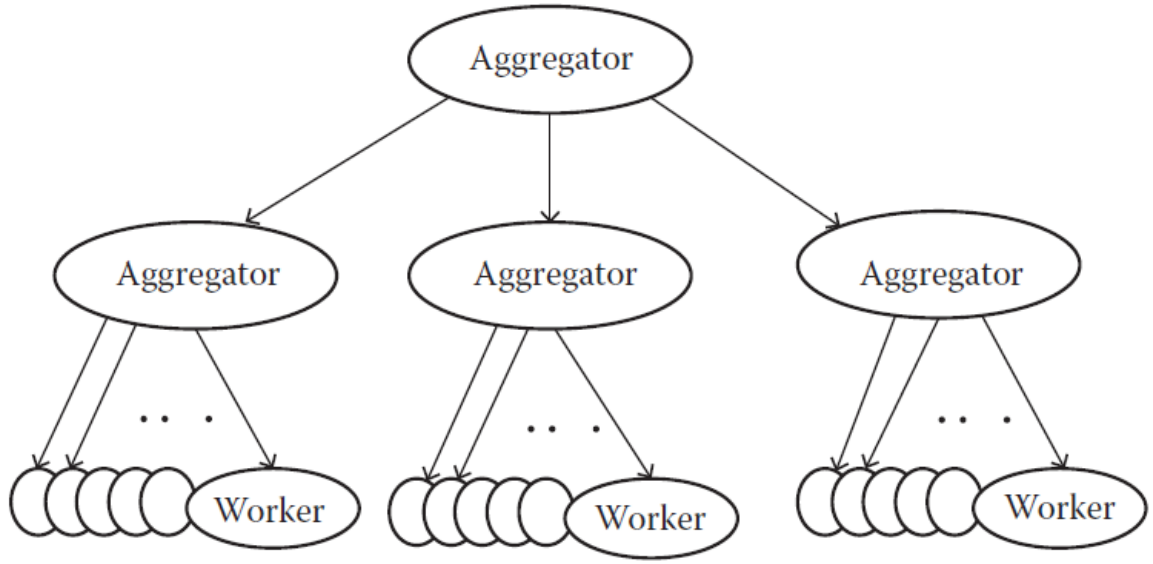


FIGURE 10.5
TCP incast.

TCP Incast

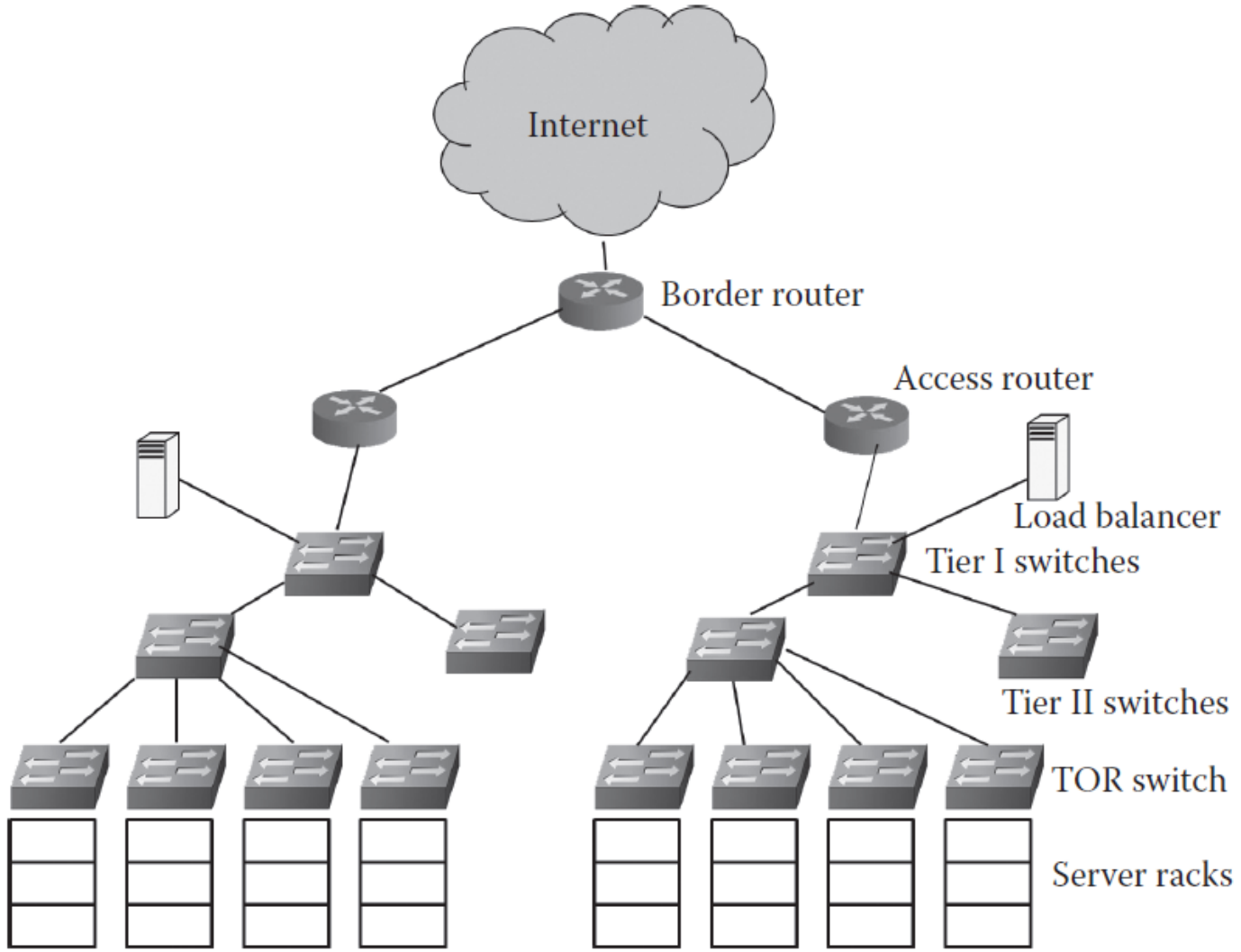
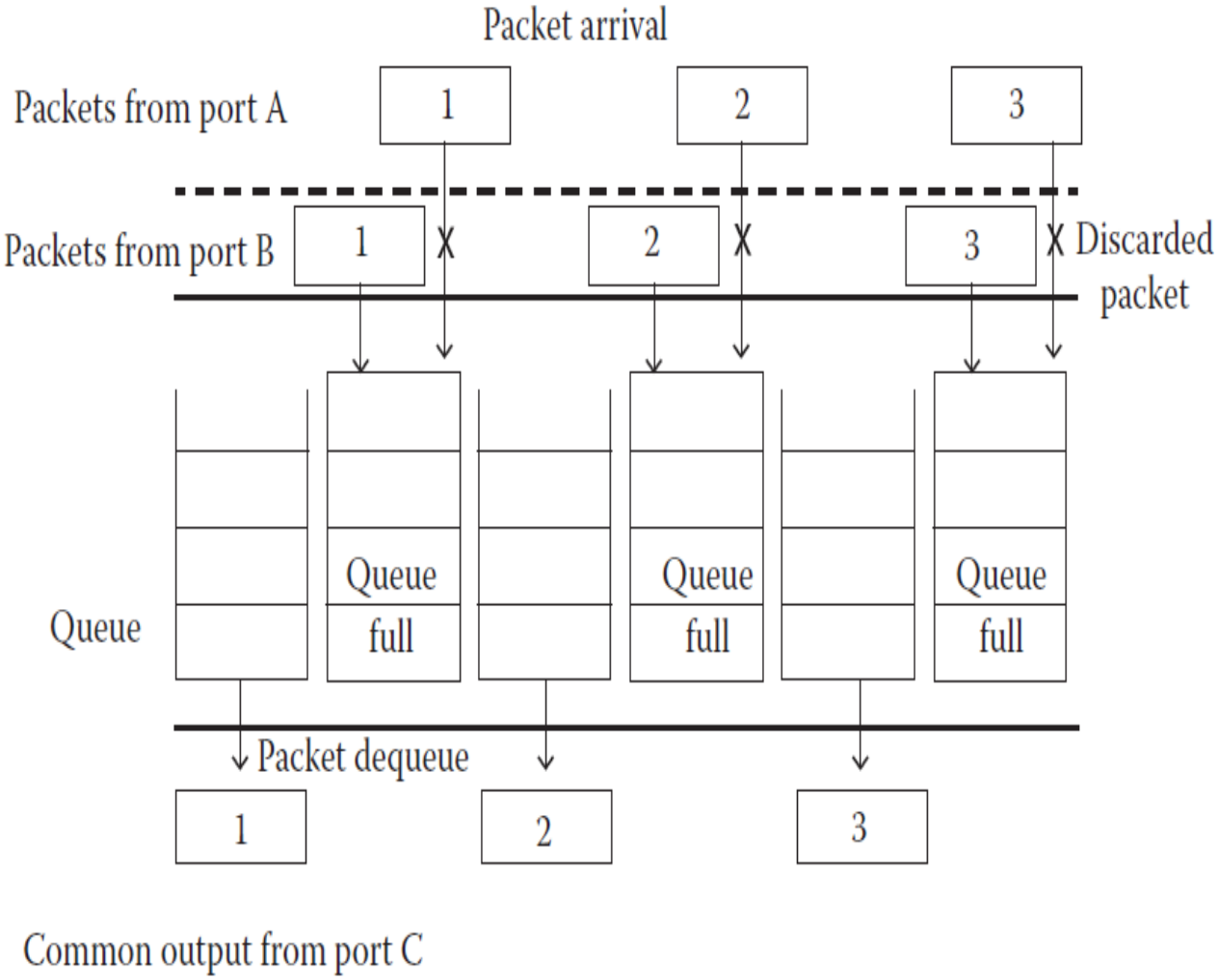


FIGURE 10.4
Partition/aggregate application structure. (From Kurose, J.F. and Ross, K.W., *Computer Networking: A Top Down Approach*, 6th edn., Addison-Wesley, 2012.)

challenges faced by the state-of-the-art TCP in DCNs

- TCP Outcast -mainly occurs in data center switches that employ drop-tail queues
- the small set of flows lose out on their throughput share significantly
- drop-tail queues is termed as *port blackout*
- throughput of a TCP flow is inversely proportional to the RTT of that flow
- This behavior of TCP leads to RTT bias



challenges faced by the state-of-the-art TCP in DCNs

- Queue Buildup - performance of mice traffic is significantly affected due to the presence of the elephant traffic

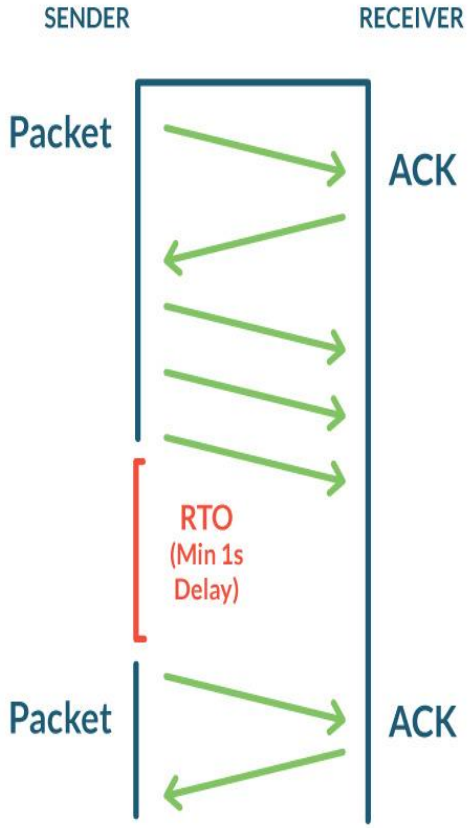
- Buffer Pressure -the elephant traffic lasts for a longer time and keeps most of the buffer space occupied

- Pseudocongestion Effect - Amazon EC2 data center reveals that virtualization dramatically deteriorates the performance of TCP and UDP in terms of both throughput and end-to-end delay
  hypervisor scheduling latency increases the waiting time for each VM to obtain an access to the processor
  Once RTO occurs, VM sender assumes that the network is heavily congested and significantly brings down the sending rate

- RTO - Recovery Time Objective (RTO) is the maximum acceptable amount of time for restoring a network or application and regaining access to data after an unplanned disruption

TCP Enhancements for DCNs

- **TCP with Fine-Grained RTO (FG-RTO)**



- default value of minimum RTO in TCP is generally in the order of milliseconds (around 200 ms).
- This value of RTO is suitable for Internet-like scenarios where the average RTT is in the order of hundreds of milliseconds
- it is significantly larger than the average RTT in data centers, which is in the order of a few microseconds
- reducing the minimum RTO from 200 ms to 200 μs significantly alleviates the problems of TCP in DCNs and improves the overall throughput by several orders of magnitude.
- real-time deployment of fine-grained timers is a challenging issue because the present operating systems lack the high-resolution timers required for such low RTO values
- *reactive* approach ,this approach significantly improves the network performance by reducing post–packet loss delay
- does not alleviate the TCP incast problem for loss-sensitive applications

TCP Enhancements for DCNs

TCP with FG-RTO + Delayed ACKs Disabled

- Delayed ACKs are mainly used for reducing the overhead of ACKs on the reverse path
- the receiver sends only one ACK for every two data packets received
- If only one packet is received, the receiver waits for delayed ACK timeout period before sending an ACK
- This timeout period is usually 40 ms.
- reducing the delayed ACK timeout period to 200 μs while using FG-RTO achieves far better throughput

TCP Enhancements for DCNs

DCTCP

- Additive increase/multiplicative decrease (AIMD) is the cornerstone of TCP congestion control algorithms
- The congestion window (cwnd) is increased by 'one' segment 'per' RTT
- congestion window (*cwnd*) AI

$$cwnd = cwnd + \frac{1}{cwnd}$$

- MD

$$cwnd = \frac{cwnd}{2}$$

- Data center TCP (DCTCP) employs an efficient multiplicative decrease mechanism that reduces the *cwnd* based on the *amount of congestion* in the network rather than reducing it by half

$$cwnd = cwnd \times \left(1 - \frac{\alpha}{2}\right)$$

- where $\alpha(0 \leq \alpha \leq 1)$ is an estimate of the fraction of packets that are marked

- when congestion is low (α *is near* 0), *cwnd* is reduced slightly and
- when congestion is high (α *is near* 1), *cwnd* is reduced by half,
- DCTCP is a novel TCP variant that alleviates TCP incast, queue buildup, and buffer pressure problems in DCNs.