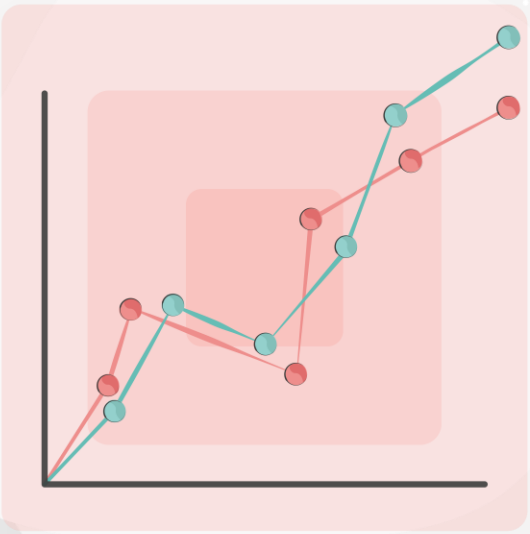
Level 5 Data Engineer
Module 4 Topic 1

Introduction to Networks for Data Engineers

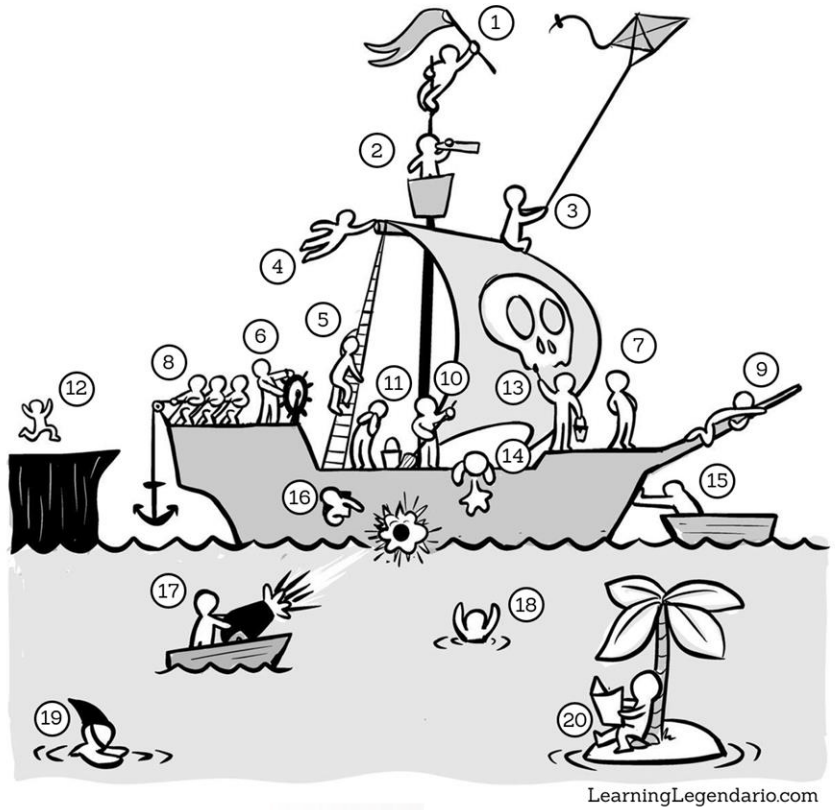
Welcome to today's webinar.



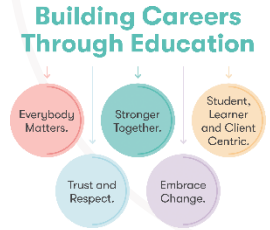
Ice breaker: Discussion

A bit of fun to start…

Which character are you when it comes to networks and cyber security?

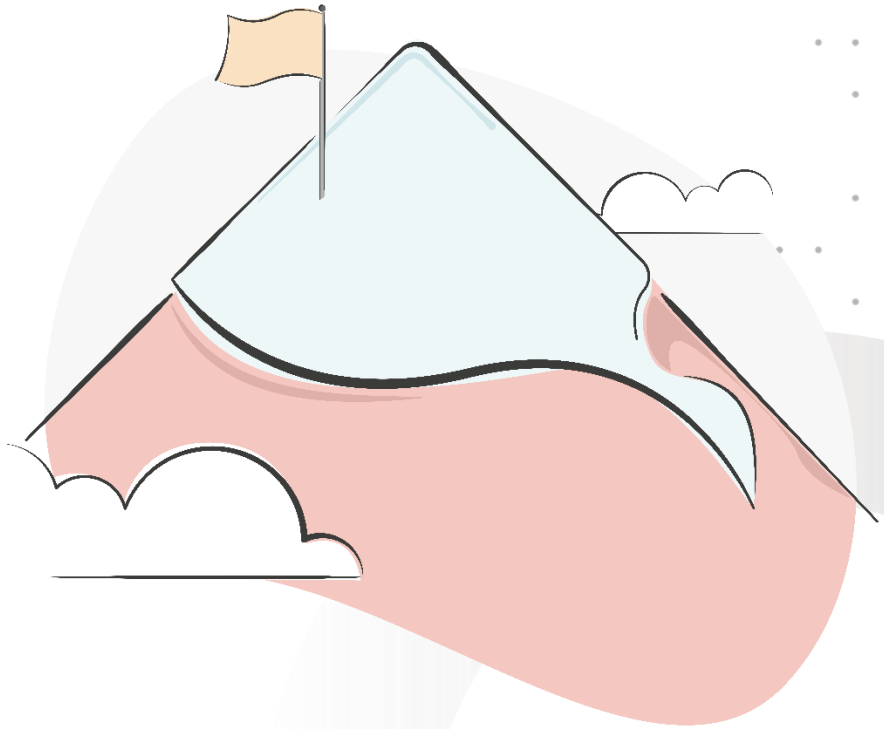


Session aim and objectives



This webinar supports the following learning outcomes:

- Explain the principles of computer networks, including IP addressing, TCP/IP and the OSI model.
- Demonstrate awareness of modern networking practices.
- Learn about network infrastructure costs and sustainability.
- Explain the role of different types of network devices.

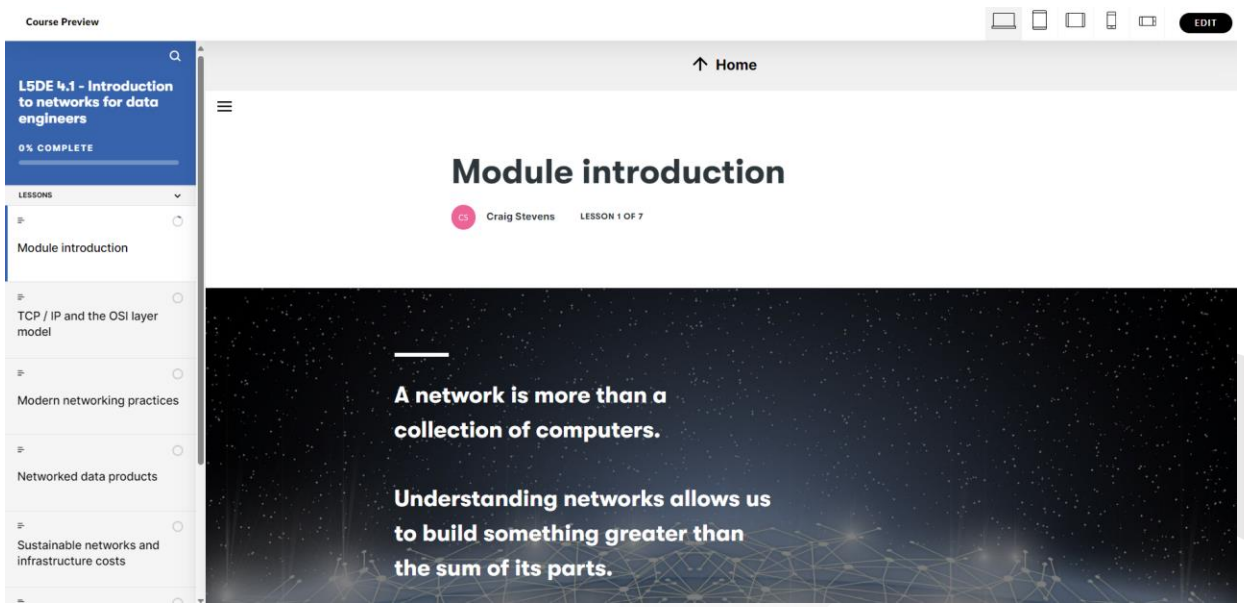


BPP

Recap of e-learning

Are you happy with your learning?

- What was the most interesting thing you learned?
- What is TCP/IP?
- How is the OSI model useful?
- Which network infrastructure elements do you remember?



A screenshot of topic 1 e-learning

Webinar Agenda

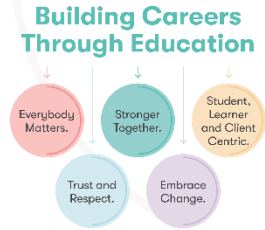
What we will cover in the webinar:

1. Basic principles of Networks
2. IP Addresses
3. Binary and hexadecimal numbers – a Data Engineer has to be able to read those!
4. Modern networking practices

Collaborate activities:

- **Practical lab** (tutor-led)

Basic principles of networks



The ‘what’ = Network Services…

- Routing & Switching
- Security Appliances & Firewalls
- VoIP & Unified Communications
- Wireless
- IPSec & SSL VPN
- Quality of Service (QoS)

BPP

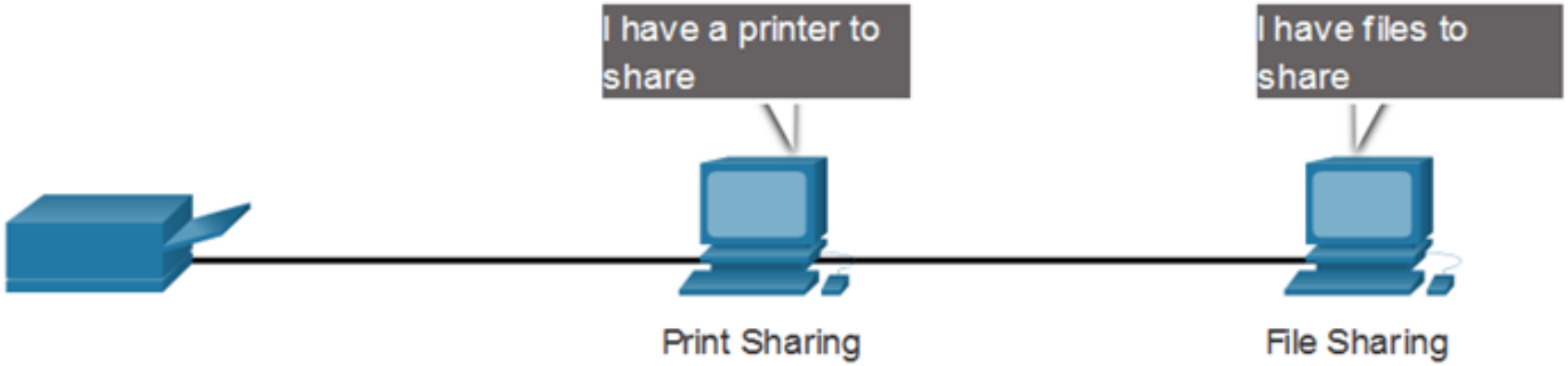
Basic principles of networks

The ‘who’ = Common network services…

- **Hub**
  - Multiport repeater
- **Switch**
  - Collision domains
  - MAC address learning
- **Router**
  - Broadcast domains
  - ‘Gateway’
- **Firewall**
  - Stateful packet inspection
- **VPN concentrator**
  - VPN termination point

Networking models

Peer to peer…



Advantages	Disadvantages
Easy to set up	No centralised administration
Scalable	Not as secure
Lower cost	Limited reliability
Used for simple tasks: transferring files and sharing printers	Slower performance

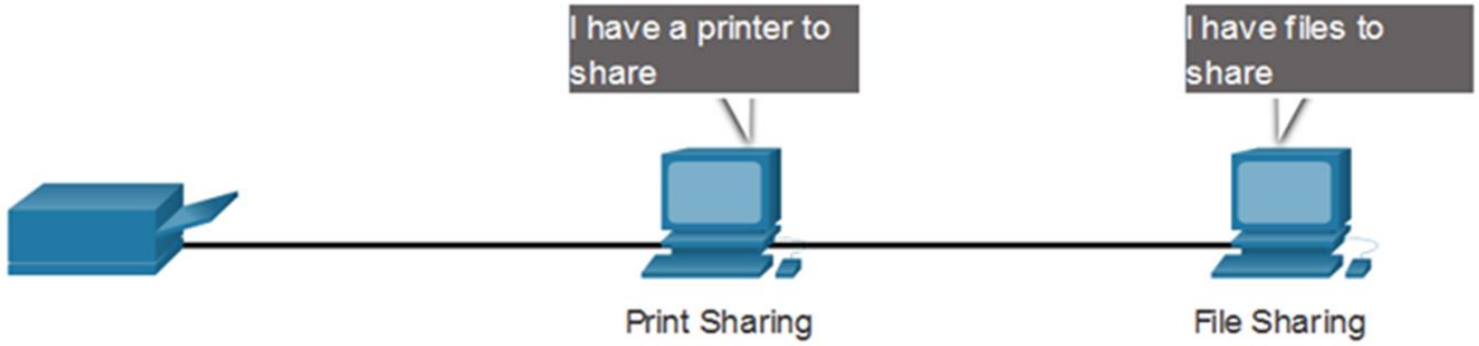
Networking models

Client/Server…

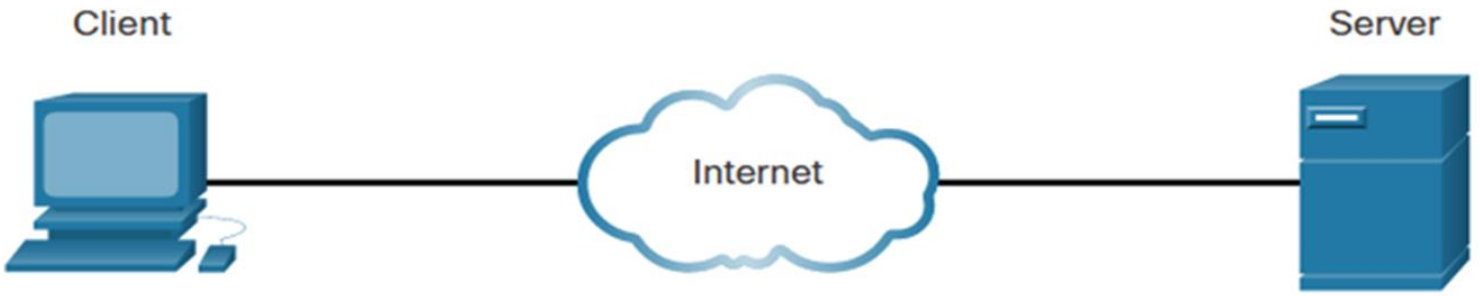
Server Type	Description
Email	Email server runs email server software. Clients use client software to access email.
Web	Web server runs web server software. Clients use browser software to access web pages.
File	File server stores corporate and user files. The client devices access these files.

Discussion

- **What computing solutions are more suited to peer-to-peer networking systems rather than client/server systems?**
- **Why?**



Peer-to-peer



Client server

Submit your responses to the chat!

Network components

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Intermediary network devices…

Wireless Router

LAN Switch

Router

Intermediary Devices

Multilayer Switch

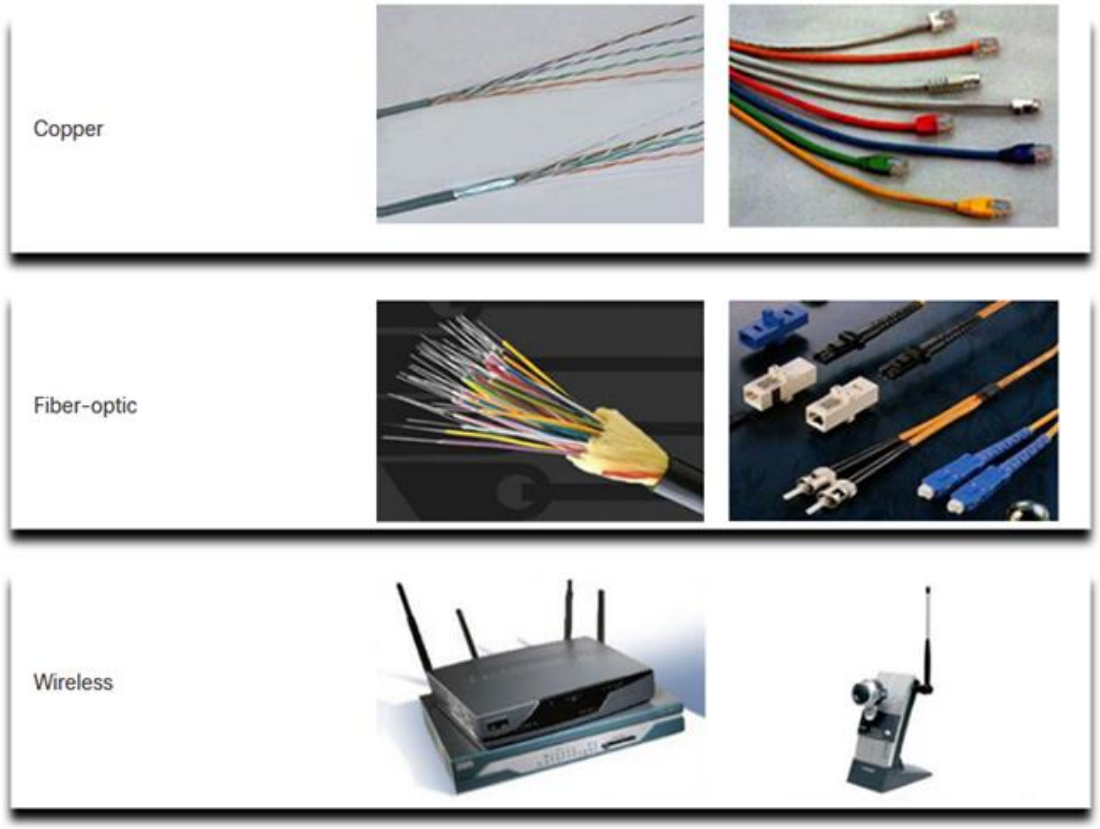
Firewall Appliance

BPP

Network components

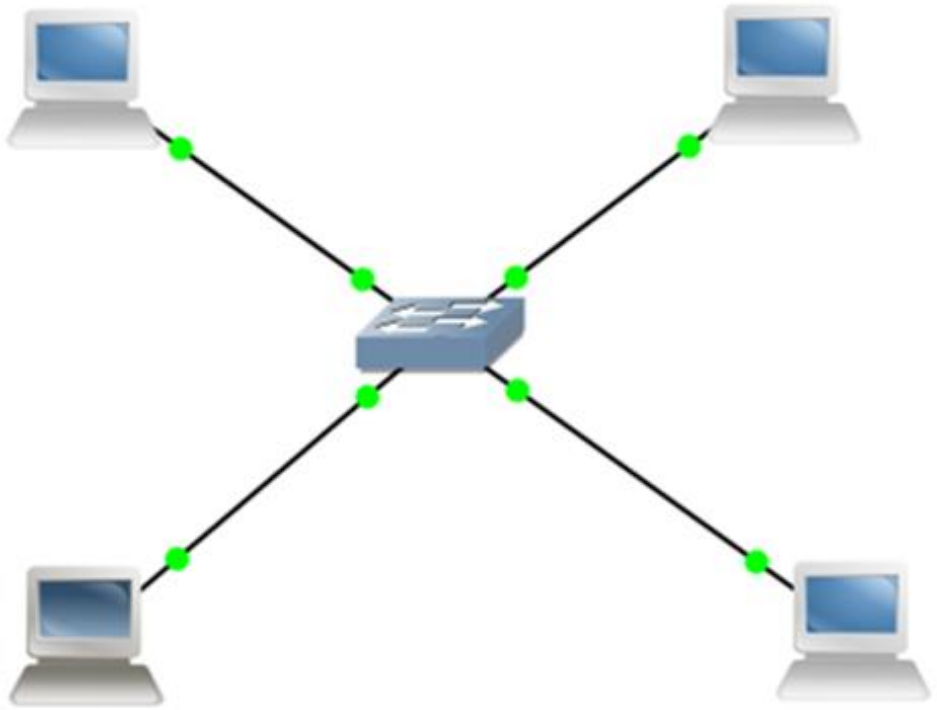
Network media…

Media Types	Description
Metal wires within cables	Uses electrical impulses
Glass or plastic fibers within cables (fiber-optic cable)	Uses pulses of light.
Wireless transmission	Uses modulation of specific frequencies of electromagnetic waves.

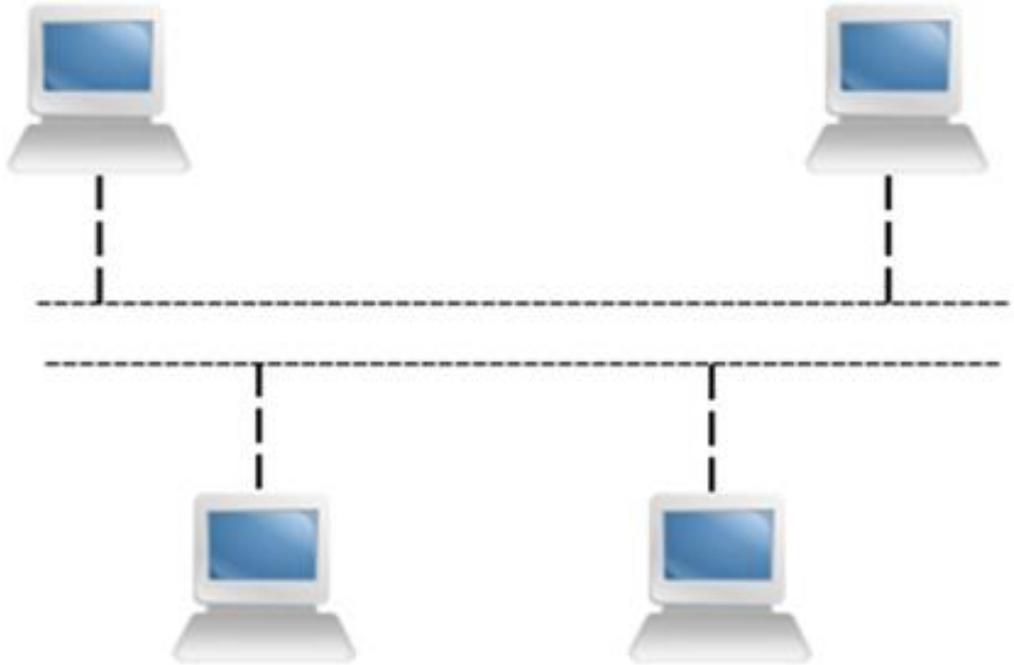


Network LAN Topologies

Most LANs are Ethernet based which have a physical star topology but logical bus topology…

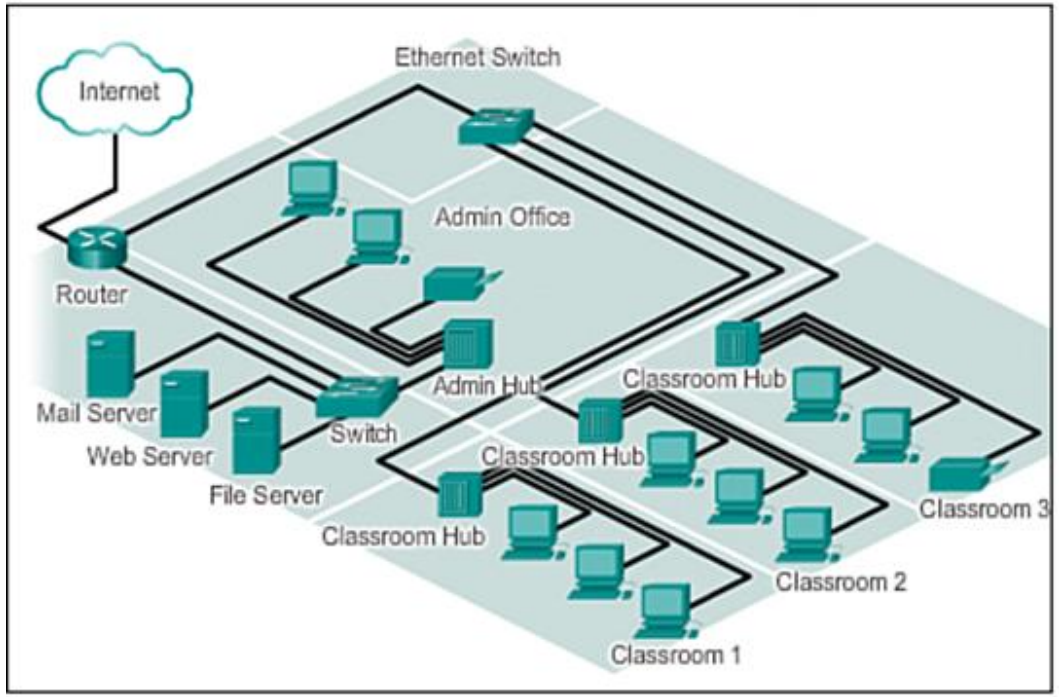


Ethernet Network Physical topology

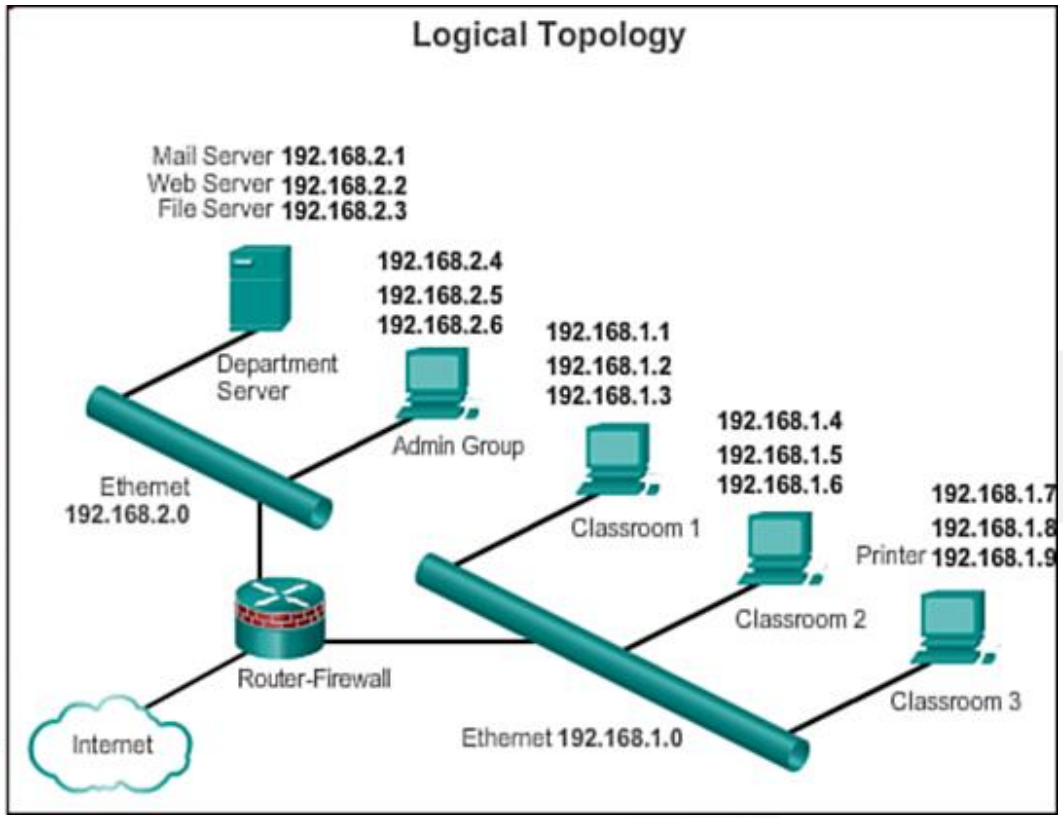


Ethernet Network Logical topology

Physical vs logical topologies



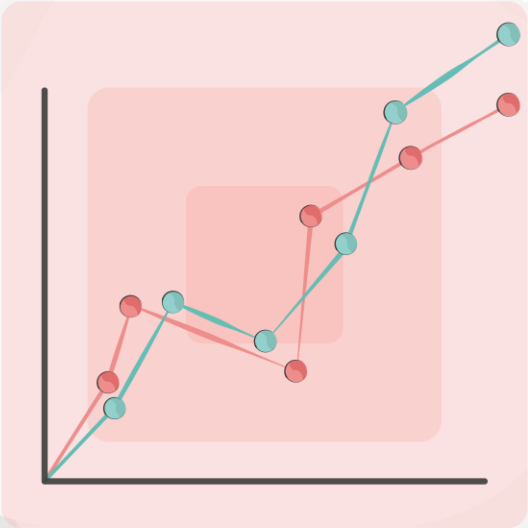
Physical topology diagram



Logical topology diagram

BPP

IP Addresses



IP Addresses

What you need to know…

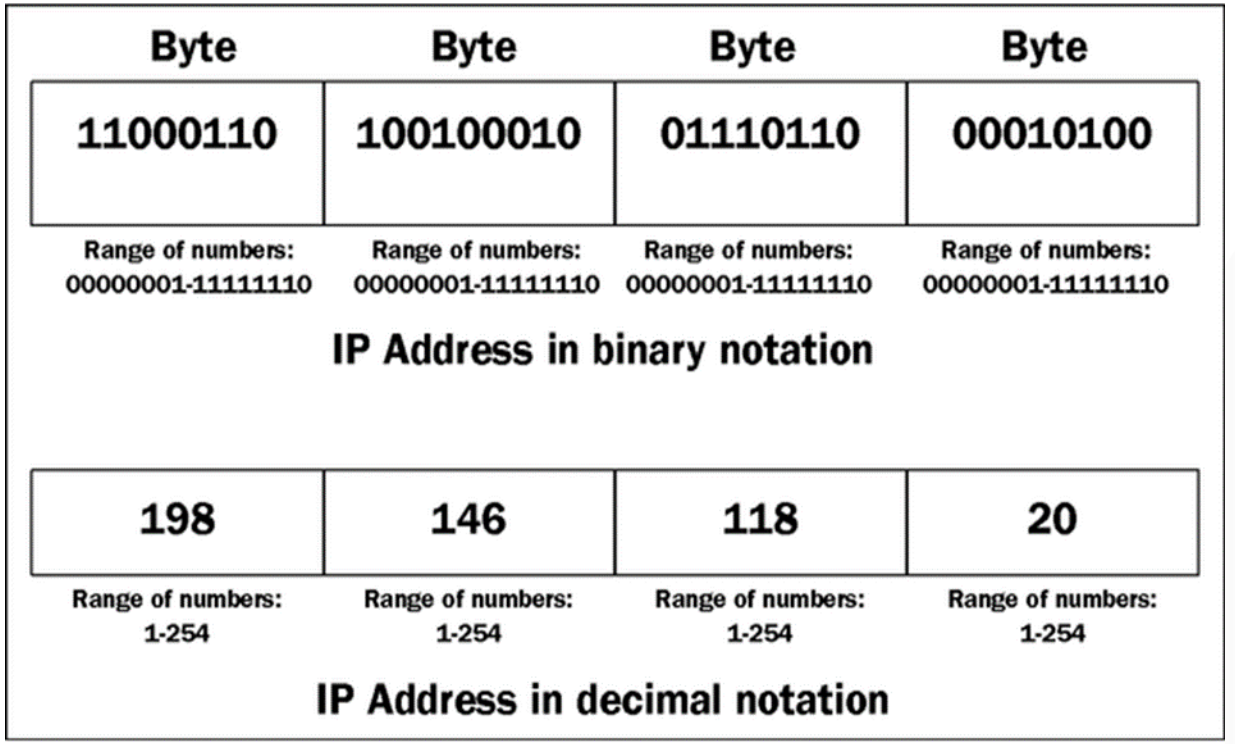
- An identifier for a computer or device on a TCP/IP network.
- 99% of the world still use IP version 4 (IPv4)
- IP version 6 is newer but adopted slowly due to compatibility issues

BPP

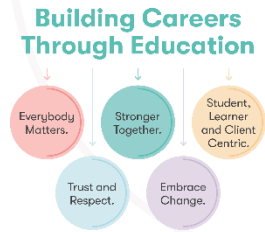
IP Addressing (IPv4)

What you need to know…

- An IPv4 address is four bytes (octets). Total 32 bits.
- Each byte is a number from 1 to 254 (0 and 255 are special)
- Stored in Big Endian order
- Written in dotted notation, e.g. 192.168.21.76



Maths Revision: Converting decimal to binary



Example convert the decimal number 198 to binary number?

Decimal: 198
Binary: 11000110

The binary number is the reminder written from the bottom to the top

Reminder	Divisor	Decimal Number
0	2	198
1	2	99
1	2	49
0	2	24
0	2	12
0	2	6
1	2	3
1	2	1
		0

BPP

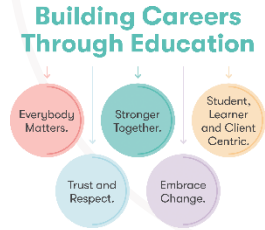
Maths Revision: Converting decimal to binary

Converting the binary number 11000110 to decimal number…

1	1	0	0	0	1	1	0
1×2^7	$+ 1 \times 2^6$	$+ 0 \times 2^5$	$+ 0 \times 2^4$	$+ 0 \times 2^3$	$+ 1 \times 2^2$	$+ 1 \times 2^1$	$+ 0 \times 2^0$
128 +	64 +	0 +	0 +	0 +	4 +	2 +	0

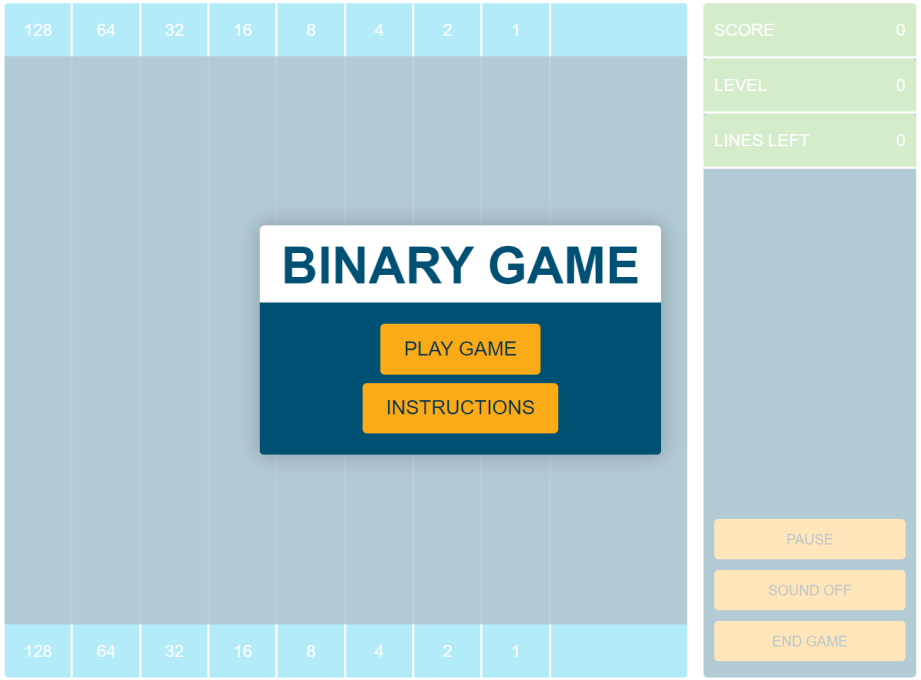
Decimal Number = 198

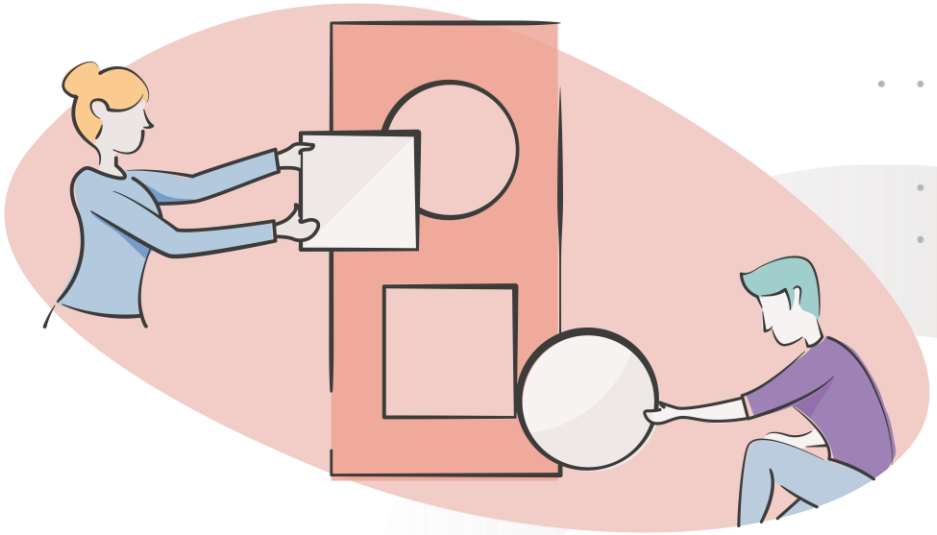
Activity



- Play the binary game
- Data Engineers have to be proficient in reading the binary system of numbering

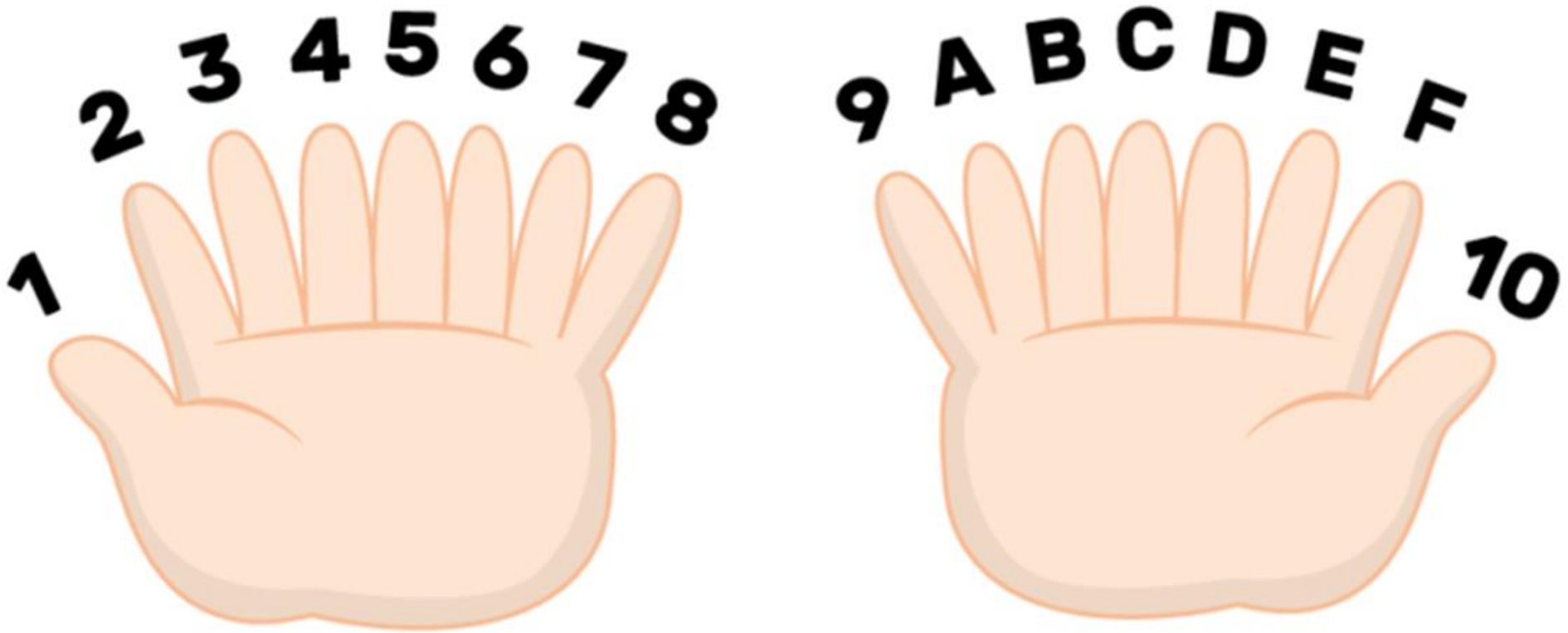
https://learningcontent.cisco.com/games/binary/index.html





BPP

Hexadecimal



Maths

Converting to Hex…

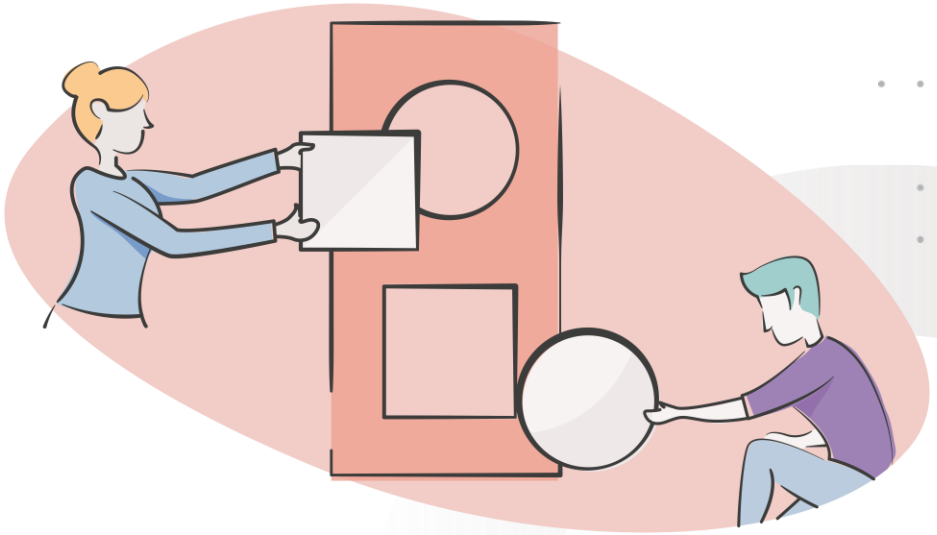
- Let’s convert 200 to hex.
- How many times 16 fits into 200?
- 200 / 16 = 12, remainder 8.
- So, we have 12 lots of 16, with 8 units left
- Denary 12 is hexadecimal C, and denary 8 is hexadecimal 8
- So, denary 200 = (12 × 16) + (8 × 1) = C8 in hexadecimal.

BPP

Activity

To practice, try converting the following hexadecimal numbers into denary:

- 45
- 2D
- E9



IPv4 vs. IPv6

- **IPv4**
  - 4,294,967,296 total addresses available
  - Addresses not assigned by geographic region
- **IPv6**
  - 128 bits used for address
  - IPv6 address written as eight 4-digit (16-bit) hexadecimal numbers separated by colons.
  - E.g. 1080:0:0:0:0:800:0:417A
  - 340,282,366,920,938,463,374,607,431,768,211,456 addresses
  - That about 3.7x10^21 addresses per square inch of the earth's surface
  - Addresses will be assigned by geographic region

BPP

Loopback address

- 127.0.0.0
  - Network number that cannot be assigned to any network
- 127.0.0.1
  - The loopback address
  - Used for diagnostic testing of the local TCP/IP installation

BPP

Automatic Private IP Addressing (APIPA)

- 169.254.x.x
- IP addresses are self-assigned when the computer asks for an IP address, and no one responds.
- i.e. the computer cannot reach a DHCP server over the network.

BPP

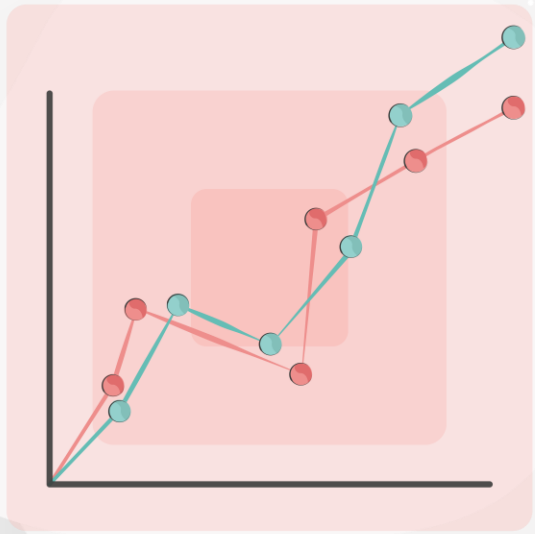
Network & Broadcast addresses

- Special IP address
- You cannot assign the highest number on a network to a host. This address is interpreted as a broadcast message for the subnet.
- For example, 255.255.255.255 broadcasts to all networks and nodes, while 172.16.255.255 broadcasts to all subnets and hosts on the network 172.16.0.0.
- You cannot assign a network number to a computer or any other host on the network.

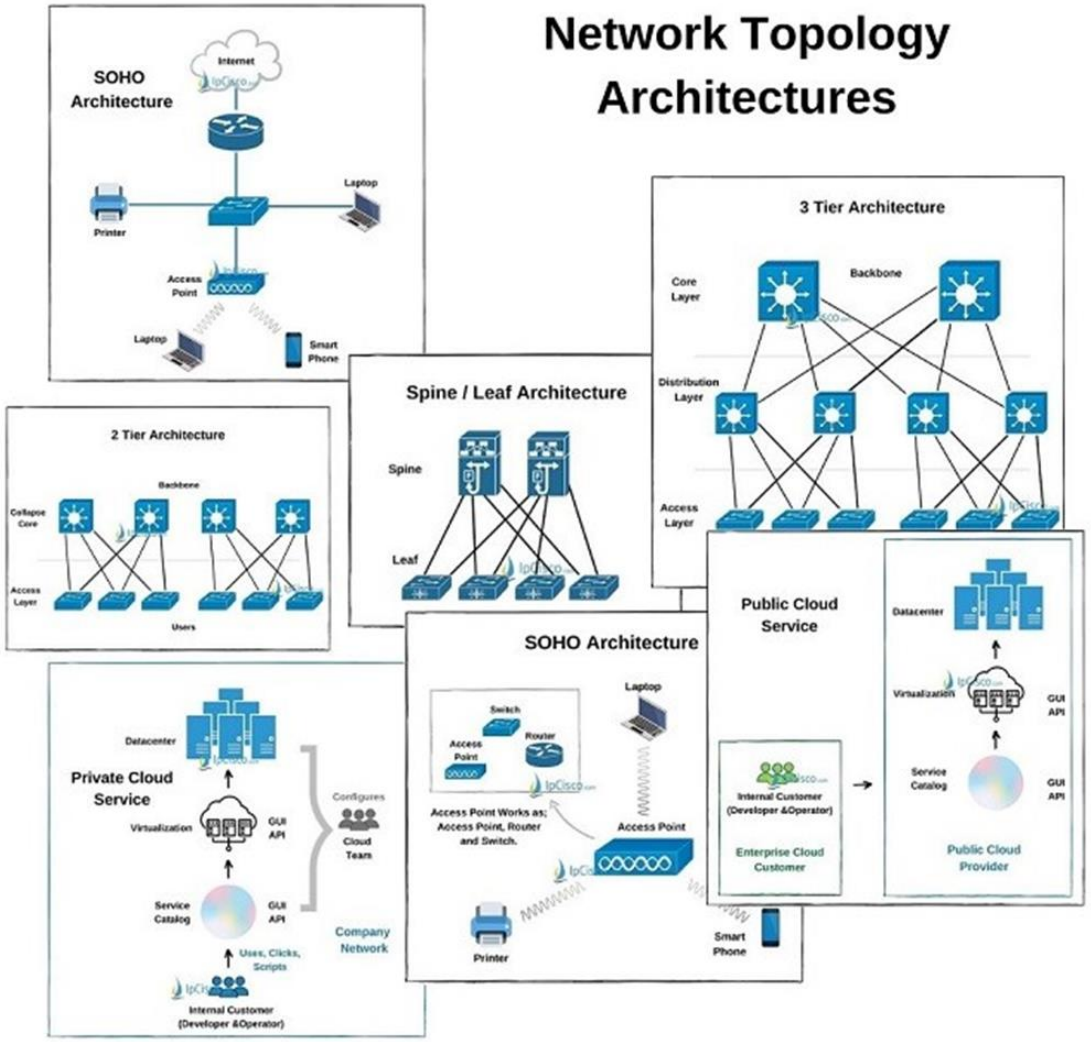
BPP

BPP

From networked infrastructures to networked data products

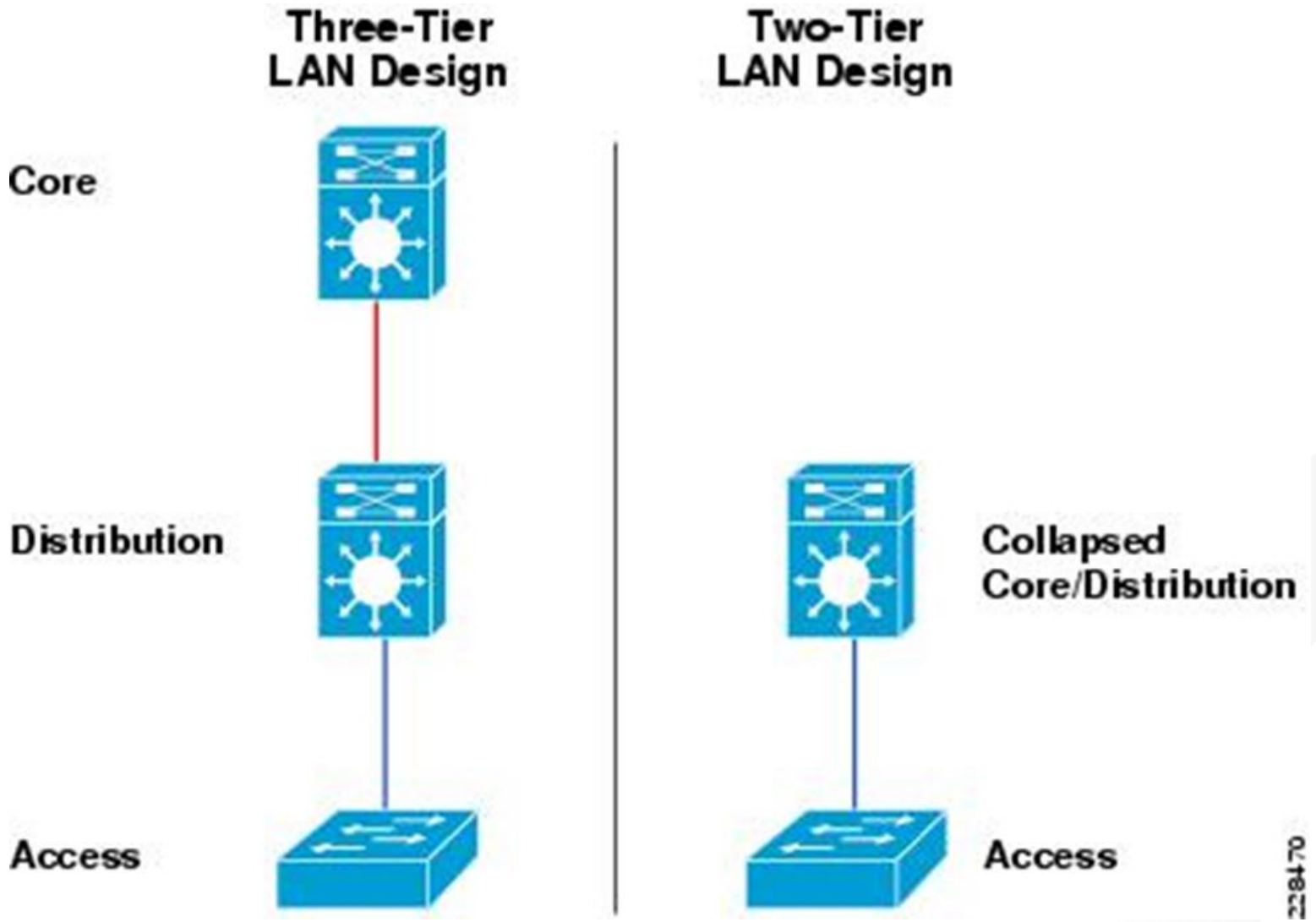


Network topology architectures



BPP

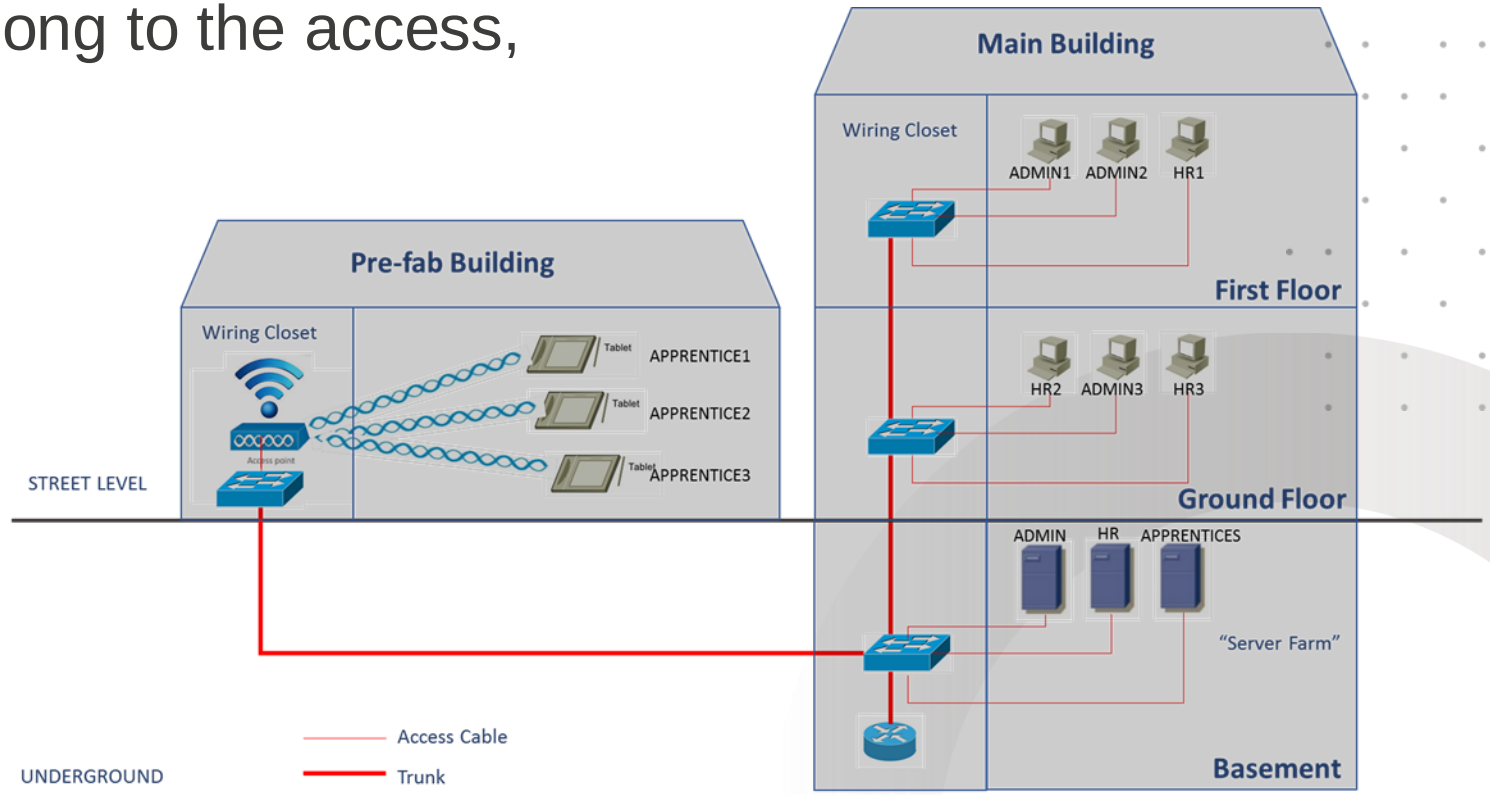
LAN Infrastructure Architecture



Activity walk-through

Campus LAN Architecture

- In this set-up, which devices belong to the access, distribution and core layers?
- Is this a 3-tier or 2-tier model?
- What is missing?

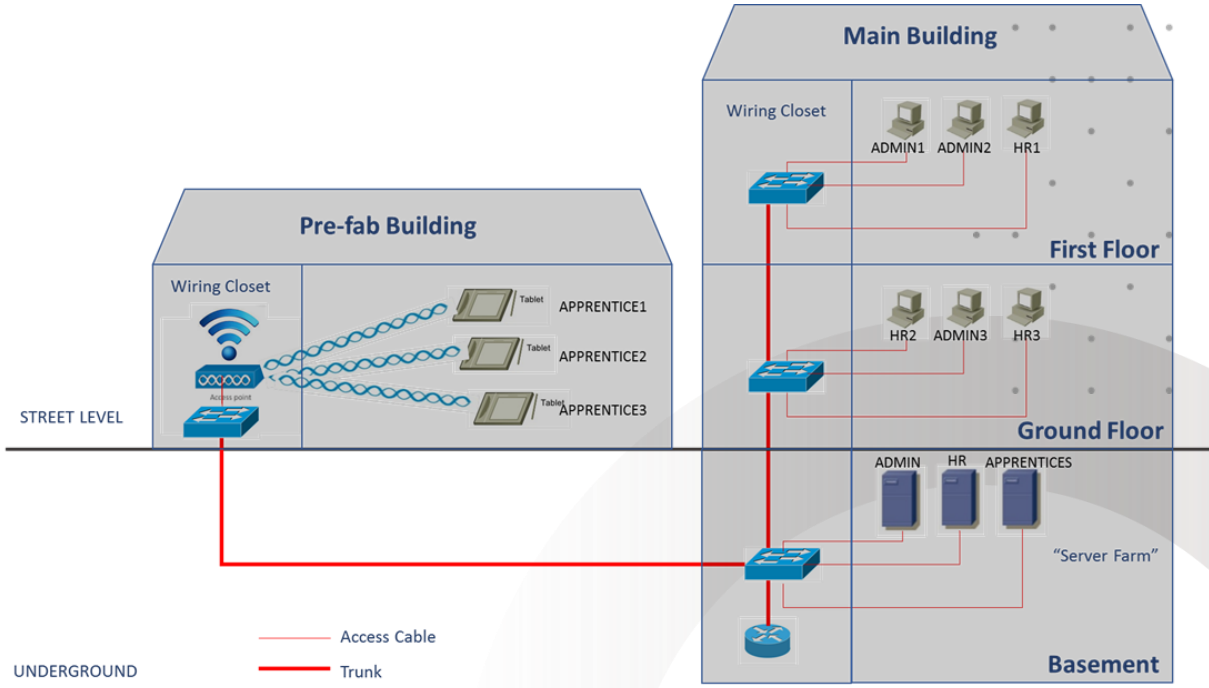


Analysis

- **Access Layer:** The devices labeled "APPRENTICE1", "APPRENTICE2", "APPRENTICE3", "HR1", "ADMIN", "HR2", "HR3", "APPRENTICES", "ADMINS", "JOB1", and "JOB2" are likely part of the access layer. These devices are at the edge of the network where end devices connect.

- **Distribution Layer:** The presence of "Wiring Closet" in both the Main Building and the Pre-fab Building suggests that these are distribution layer devices, likely aggregating connections from access layer devices and providing connectivity to the core layer if present.

- **Core Layer:** There are no clearly marked core layer devices or network segments that explicitly indicate core layer functions, such as high-speed backbone connectivity or centralized routing services.



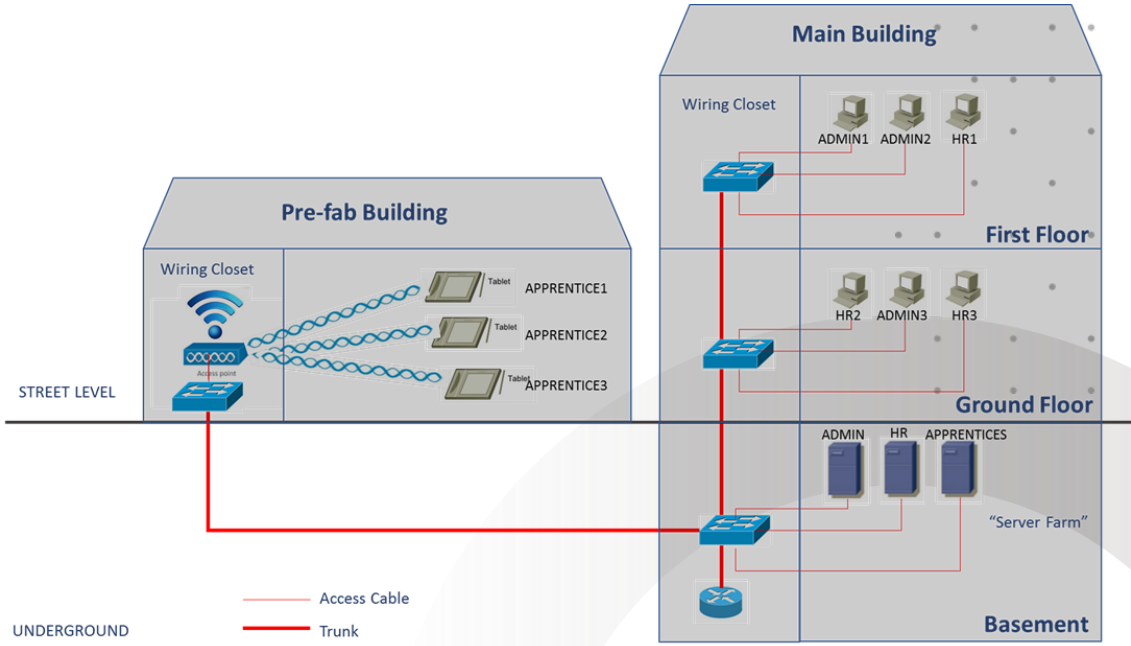
Missing Items

- **Core Layer Devices:** If this were a 3-tier architecture, we might expect to see dedicated high-speed backbone switches or routers forming a core layer distinct from the distribution layers.

- **Redundancy:** There is no explicit indication of redundancy in critical components like routers, switches, or connectivity paths, which is crucial for maintaining network availability and reliability.

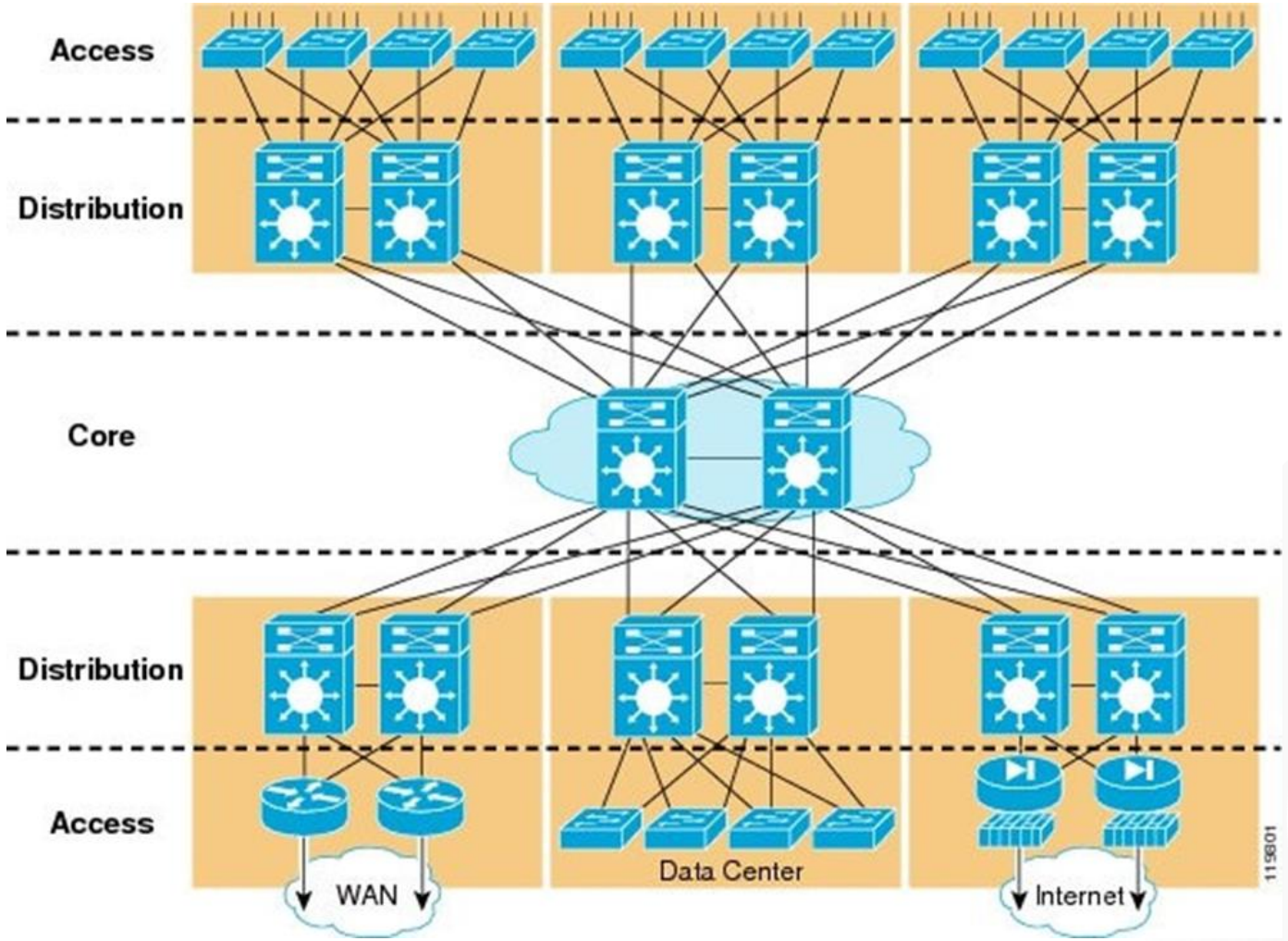
- **Security Devices:** Firewalls, intrusion detection/prevention systems (IDS/IPS), and other security measures are not marked but are critical in protecting against external and internal threats.

- **Labeling for Uplinks and Connectivity Types:** More detailed labeling on the nature of connections (fiber, Ethernet) and speeds or roles (uplinks, interconnects) could help in understanding the network architecture more clearly



Redundancy and high availability

Redundant links for High Availability



The server farm

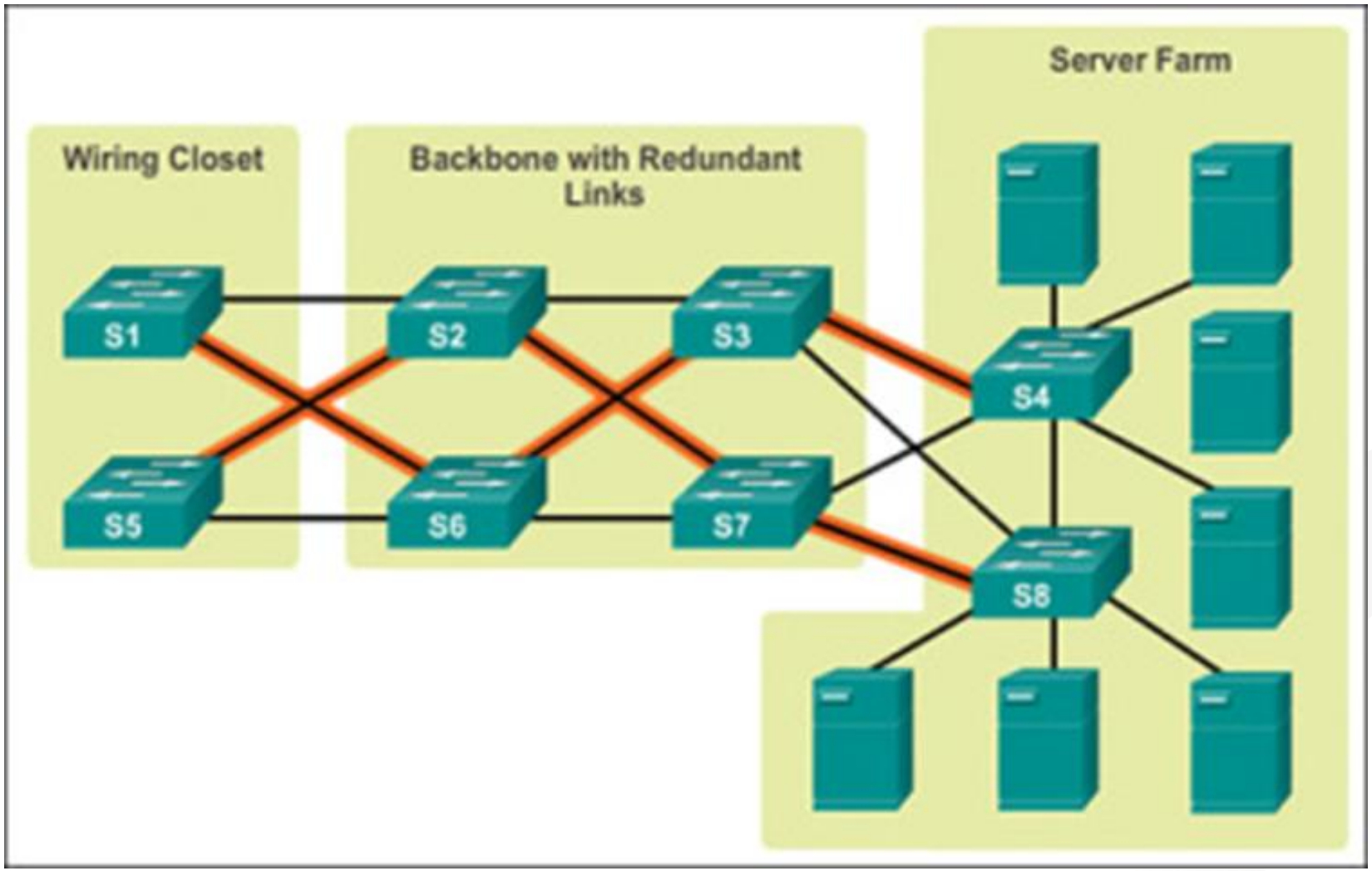
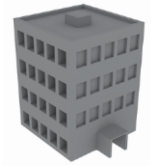
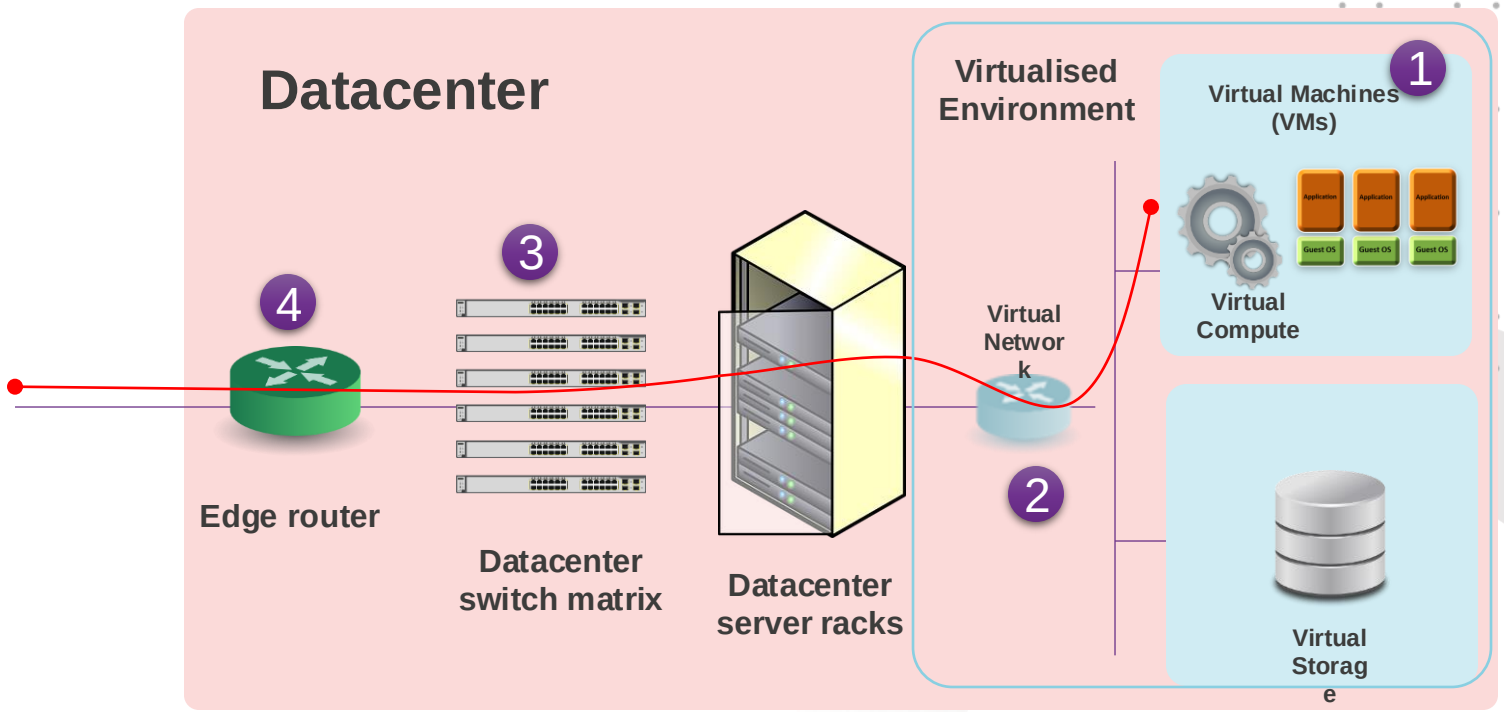


Image source Ciscopress.com:
https://www.ciscopress.com/articles/article.asp?p=2189637&seqNum=4

Data centre infrastructure

Data Centre Components





Data Centre Equipment

Data Centre Infrastructure

- Switches, racks, servers
- Leaf and spine datacenter topology

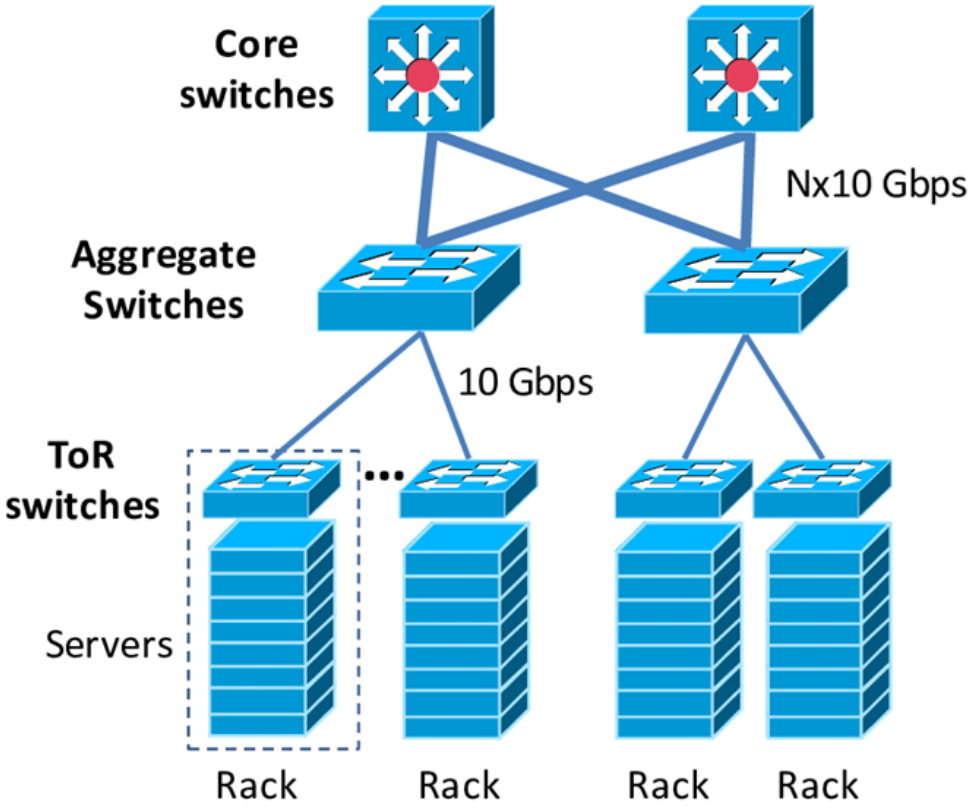


Image Source: opennebula.org

Data Centre Equipment

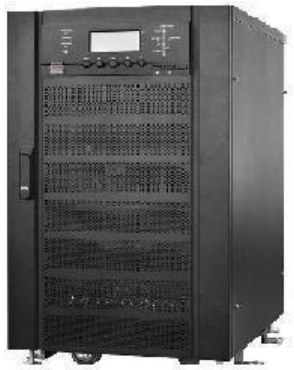
- Datacentres contain aisles, with each aisle containing rows of cabinets.
- Each cabinet contains multiple rack servers.

Cabinet containing individual rack servers

Data Centre Equipment

- To ensure a reliable uninterrupted supply of power, typically a data centre would use backup generators in case of a power outage…

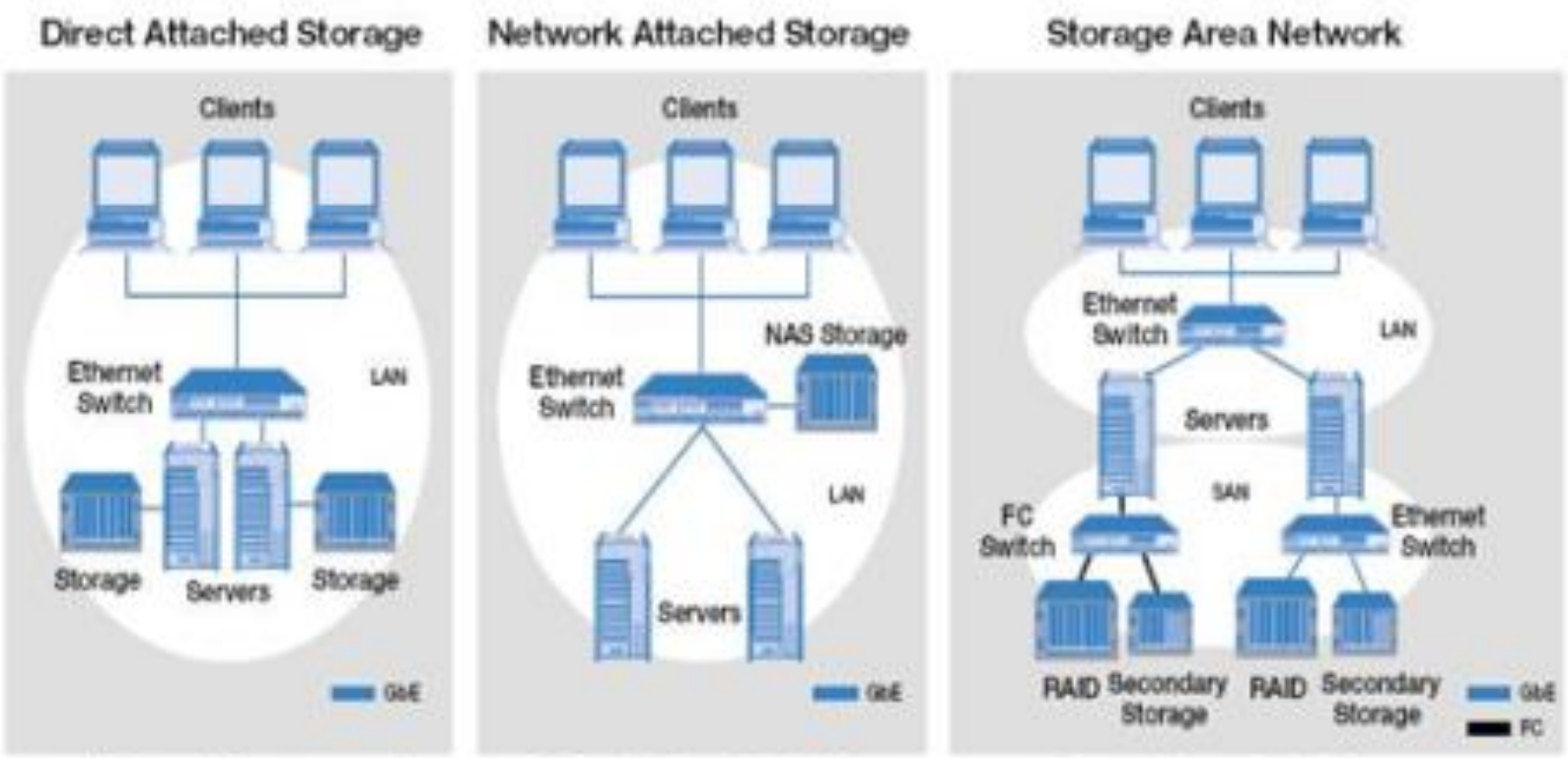
- However, these can take a few minutes to come online, so in the meantime..



… a large set of batteries keep critical systems running for long enough to either shutdown properly, or transition to the power supplied by the generator.

This is a UPS – Uninterruptable Power Supply

The evolution of network storage



- High cost of management
- Inflexible
- Expensive to scale

- Transmission optimized for file transactions
- Storage traffic travels across the LAN

- Transmission optimized for file transactions
- Separate LAN and SAN
- Increases data availability
- Flexible and scalable

BPP

Discussion

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

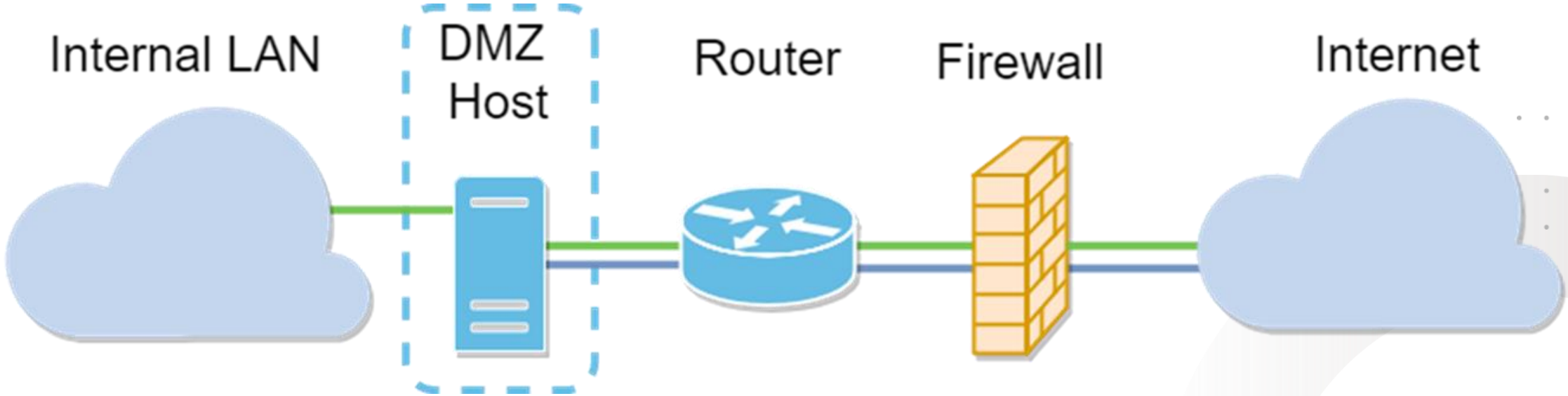
List and describe some of the challenges an in-house IT team would encounter in setting up an on-premise datacenter.

Submit your responses to the chat!

BPP

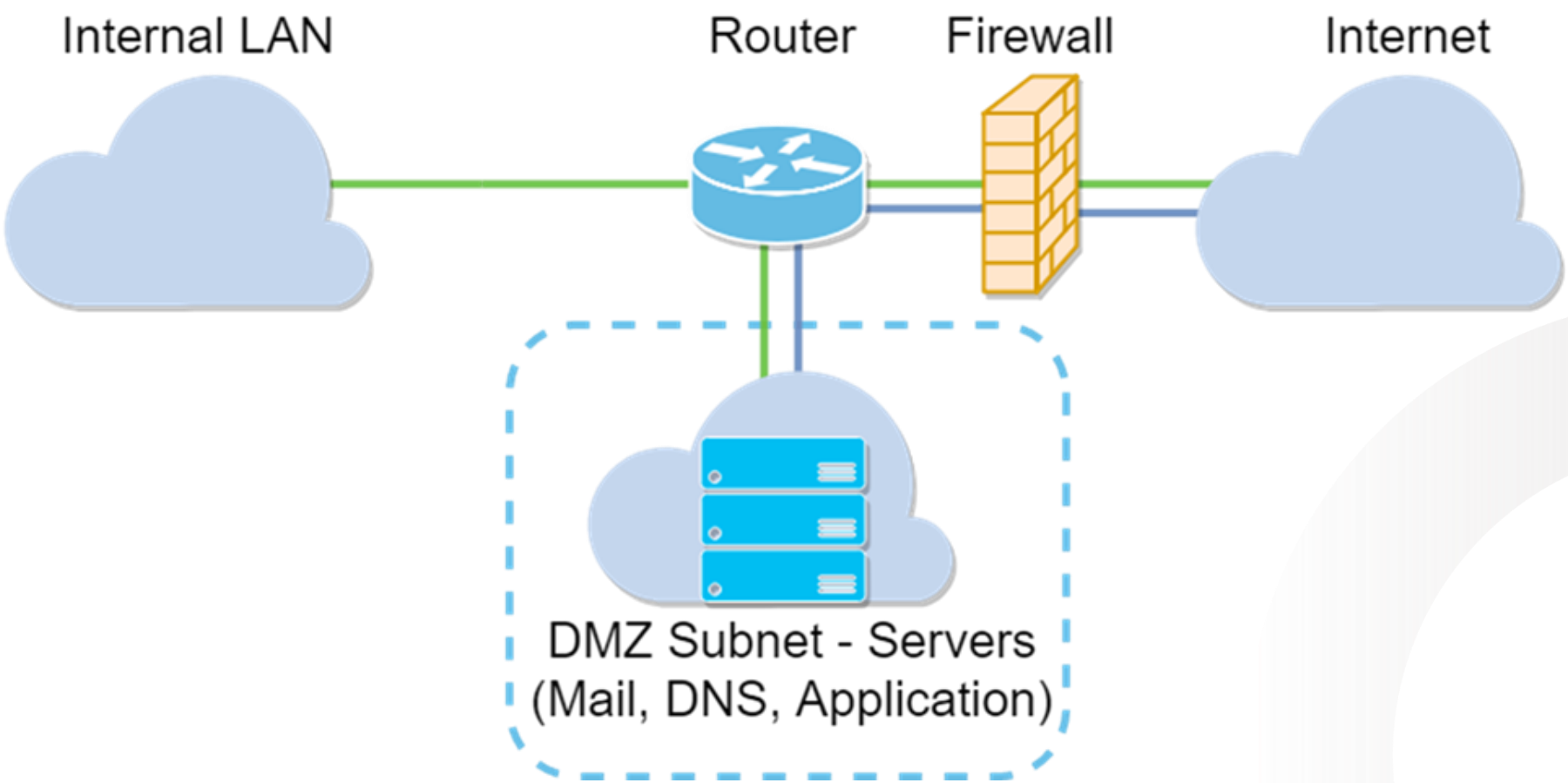
DMZ

Options for using a Firewall to create a DMZ – Option 1



DMZ

Options for using a Firewall to create a DMZ – Option 2

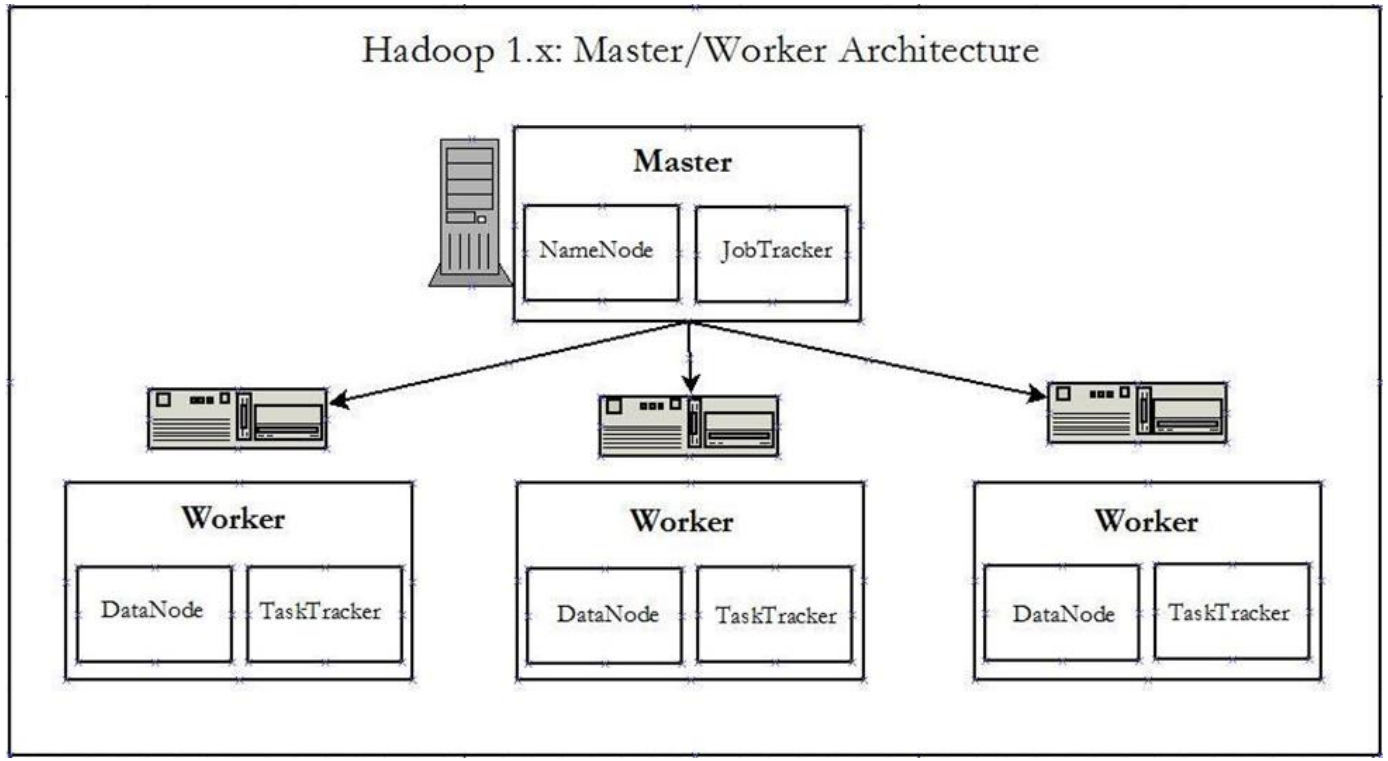


Discussion – Logical data products

Building Careers Through Education

Everybody Matters. Stronger Together. Student, Learner and Client Centric. Trust and Respect. Embrace Change.

Logical cluster architecture



Submit your responses to the chat!

Question: What problems can you anticipate with this sort of architecture?

BPP

Distributed data products

- **Clients send heartbeats to the server** every three seconds
- After a period with no heartbeats, a client is marked as lost

➢**Data recovery will kick in**

- A client can rejoin the cluster at any time

➢Alternatively, **whitelists** can be created by administrators to control which worker hosts are allowed

➢Similarly, blacklists can be created

BPP

High availability (HA)

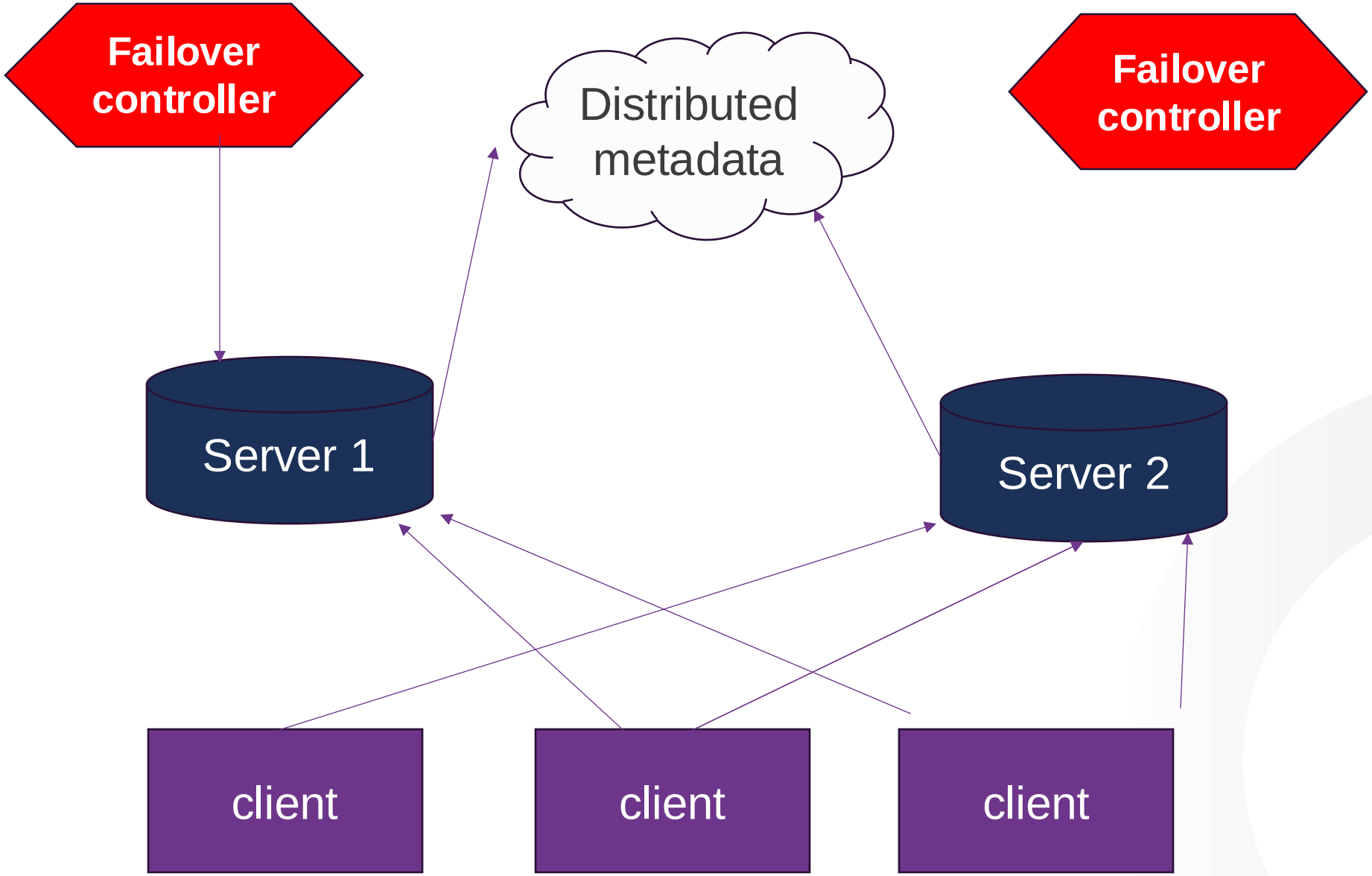
- You may have realised that if the server becomes **unresponsive**, you cannot use the data product

- **High availability** is a mode for mitigating this **single point of failure:**

➢Two servers – standby and active

- New components keep tracking who is standby and manage **failover control**

Advanced distributed networked architectures



What normally sits on top of clusters

- Data processing frameworks, like Spark

➢Use **data locality** to optimise data crunching

- No-SQL data access

➢**Such as Key-value** stores, optimised for thousands of inserts per second

- **SQL** engines

➢They abstract the distributed data providing a SQL layer on top

BPP

Data locality

- Bringing computation to the data
- Modern data products will schedule computation on the nodes that already contain the data to be crunched
- This minimises network chatter
- Rack awareness plays a role – scheduling computation in the same rack

BPP

Network considerations for clusters

- **Scalability –** master/worker architecture and the “shared nothing” paradigm in the workers

➢Limited by the master’s resources (recommended to read up on Federation in your spare time)

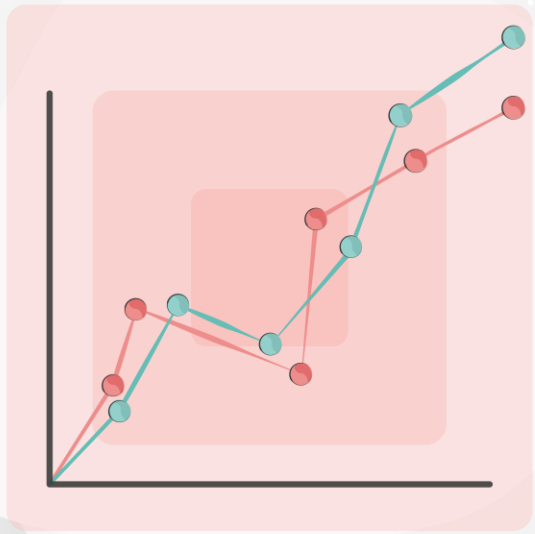
- **Reliability** – checkpointing, heartbeating, rack awareness, bit-rot prevention

➢Recommended to read up on back-up and disaster recovery solutions in your spare time

- **Security –** authentication, authorisation, non-repudiation, encryption, availability and isolation

BPP

Modern networking practices



Network segmentation

- Segmentation is a useful concept

➢Dividing a network into multiple smaller networks
➢Traffic on one network is separated from another network's traffic

- Accomplish the following:

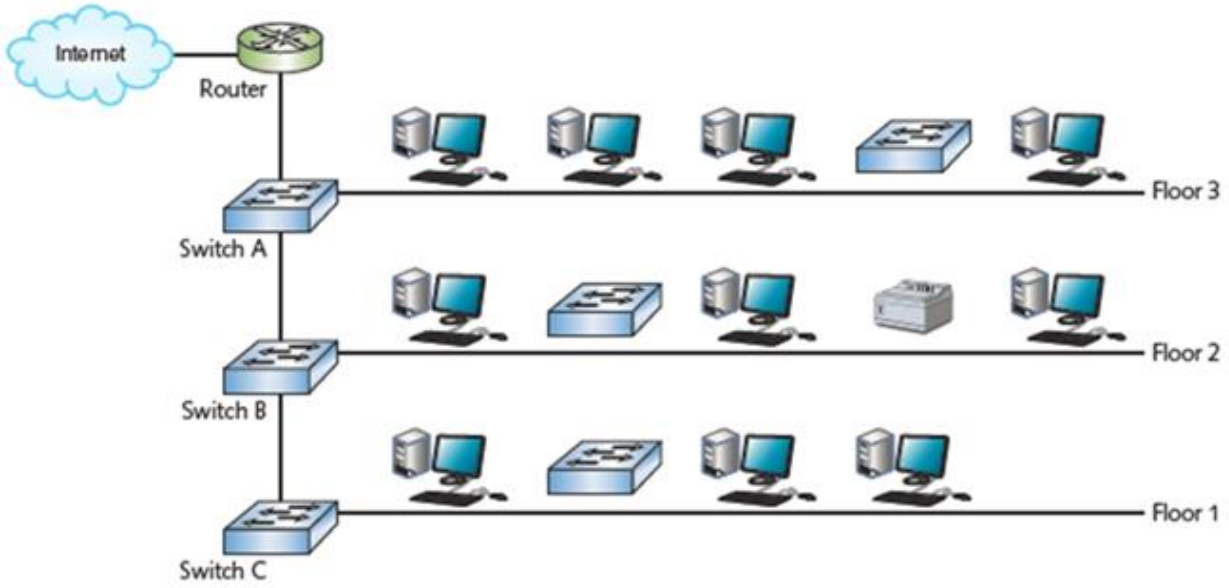
➢Enhance security
➢Improve performance
➢Simplify troubleshooting

BPP

Network segmentation

Example

- A business has grown from 20-30 computers to having a few hundred computers on three floors
  - ➢ There is only a single LAN or broadcast domain
  - ➢ One router serves as the default gateway for the entire network
- To better manage network traffic, segment the network so that each floor contains one LAN, or broadcast domain (using subnets)



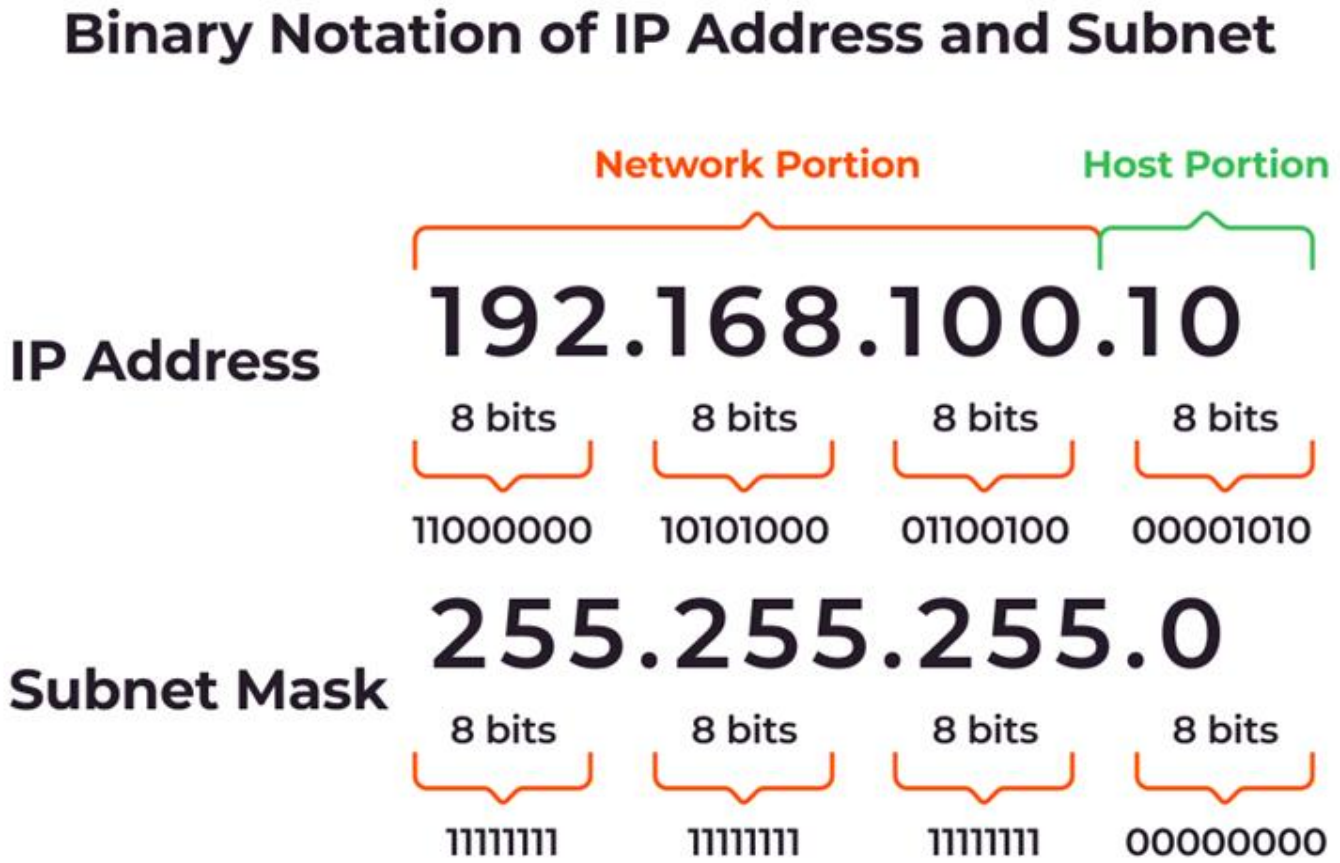
A single LAN with several switches and a router

Visualisation

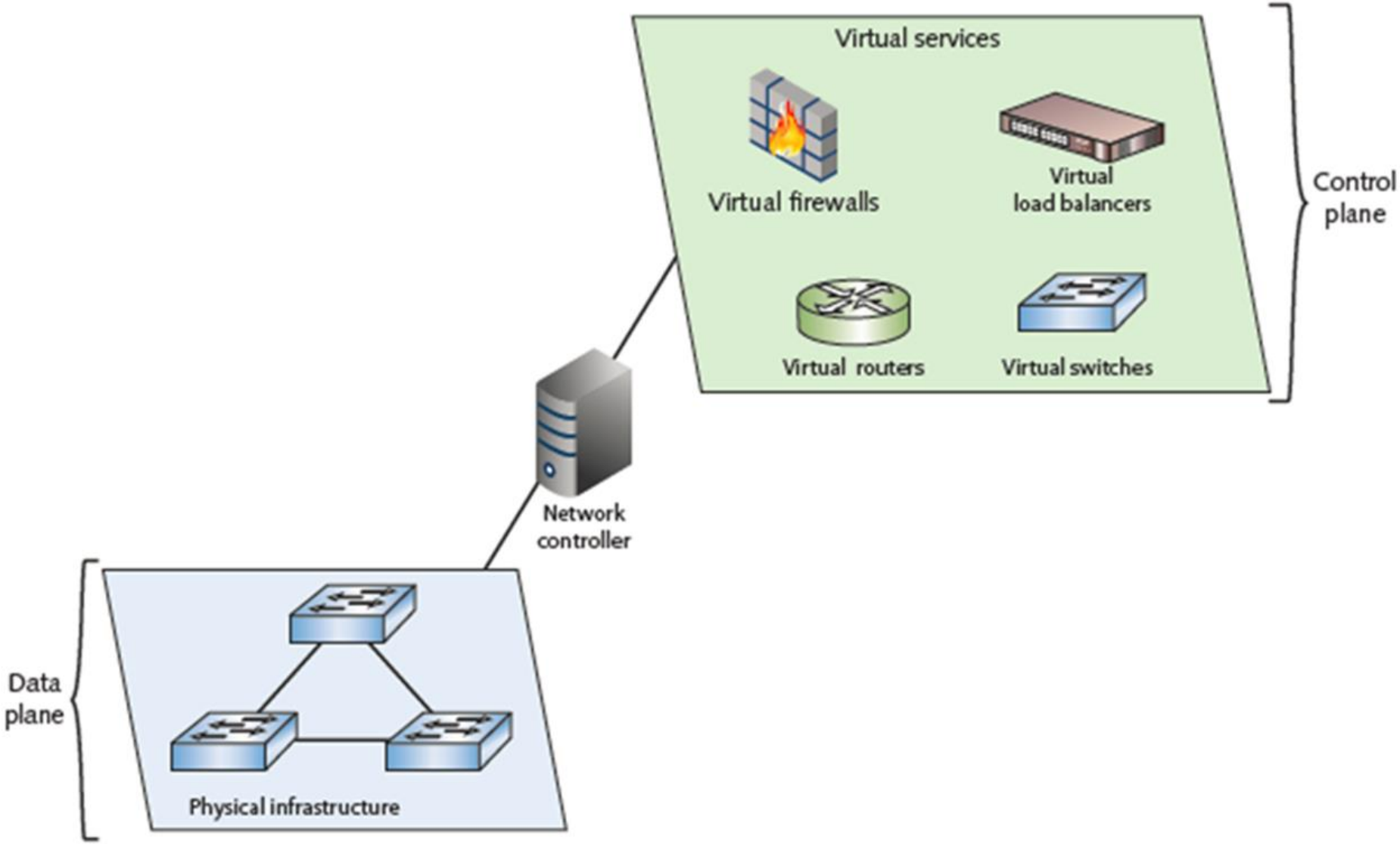
BPP

Networking hosts

In networking, hosts are on the same subnet if they have the same subnet mask.



SDN (Software Defined Networking)



Virtual network components

- Virtual network
  - ➢Can be created to consist solely of virtual machines on a physical server
- Most networks combine physical and virtual elements

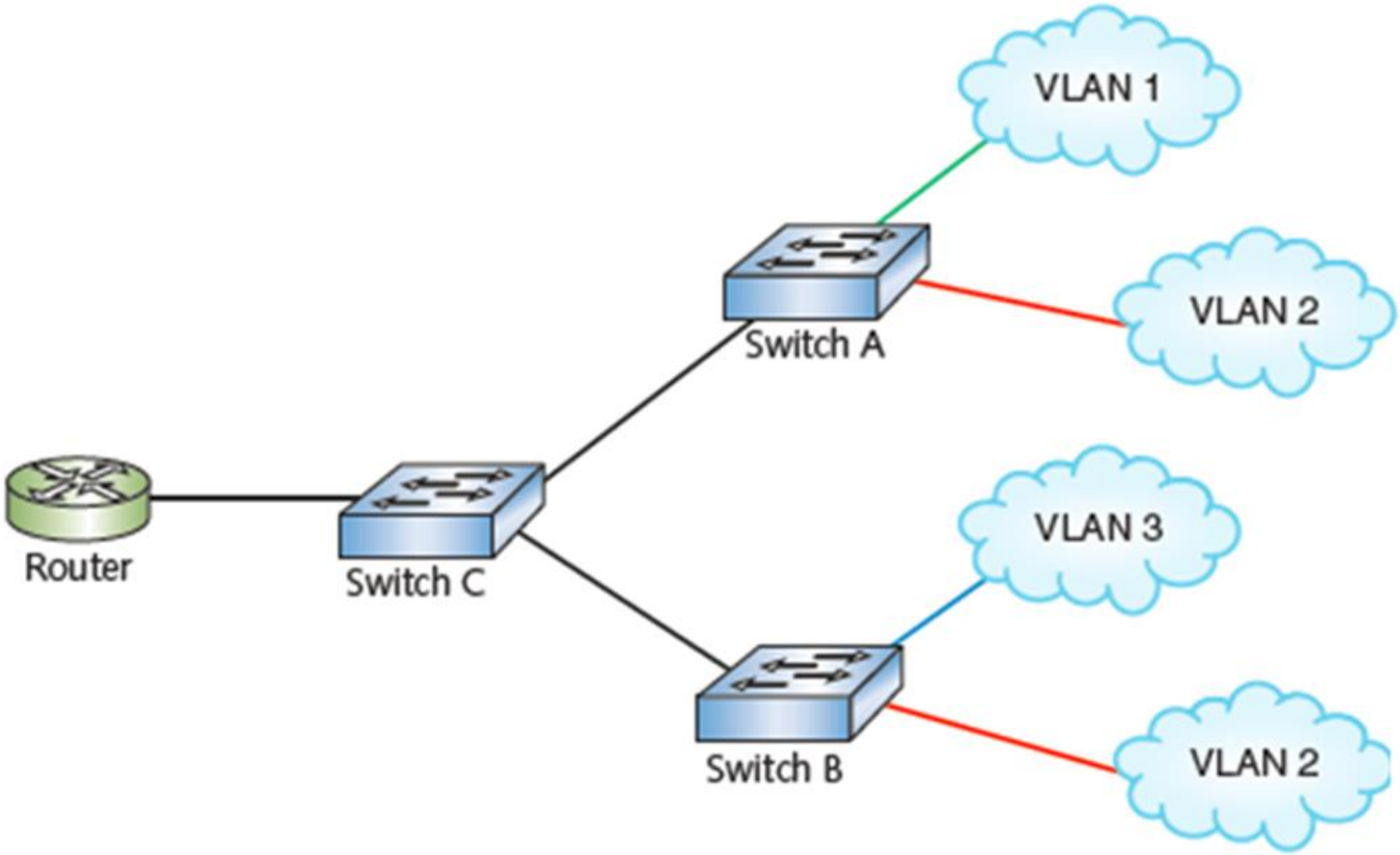
BPP

Virtual appliances and virtual network services

- Alternative to test servers for new software
- Virtual appliance includes:
  - Image of operating system, software, hardware specifications, and application configuration
- Most commonly virtual servers
- Popular functions
  - Firewall
  - Network management
  - E-mail solutions
  - Remote access

BPP

Virtual Local Area Networks



A single VLAN can be managed by multiple programmable switches

Virtual Local Area Networks

- Reasons for using VLANs:
  - Separating groups of users who need special security or network functions
  - Isolating connections with heavy or unpredictable traffic patterns
  - Identifying groups of devices whose data should be given priority handling
  - Containing groups of devices that rely on legacy protocols incompatible with the majority of the network's traffic
  - Separating a large network into smaller subnets

BPP

VLANs and trunking

- Trunk
  - ➢A single physical connection between switches through which many logical VLANs can transmit and receive data
- A port on a switch is configured as either an access port or a trunk port
  - ➢Access port - used for connecting a single node
  - ➢Trunk port - capable of managing traffic among multiple VLANs

BPP

VLANs vs subnetting

- VLANs are ideal for controlling broadcast traffic, reducing network congestion
- **VLANs = Layer 2**
- Subnetting is essential for managing IP addresses and improving routing efficiency.
- **Subnetting = Layer 3**
- VLAN can align with a subnet, but can also be idependently addressed using VLAN ID

BPP

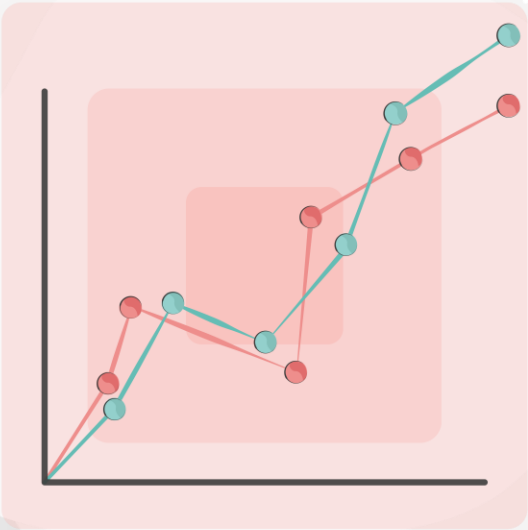
Policy-based VLANs

- VLANs can also be assigned based on policies such as device type, user roles, or security requirements.

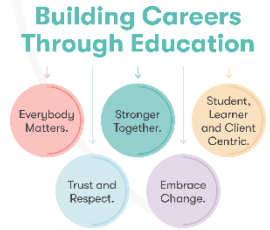
- Example: All VoIP phones might be assigned to VLAN 40 regardless of their physical location.

BPP

Network costs and sustainability



Understanding network costs



Logical cluster architecture

- Capital Expenditure (CapEx): Initial costs for network hardware like routers, switches, servers, and cabling.
- Operational Expenditure (OpEx): Ongoing costs including power consumption, maintenance, and network management.

Submit your responses to the chat!

Question: How do these costs impact budget planning and overall financial strategy for businesses?

BPP

TCO

Total Cost of Ownership encompasses all the direct and indirect costs associated with the acquisition, deployment, operation, and maintenance of network resources over their lifecycle.

- **Hardware and Software Acquisition:** Costs of purchasing network equipment and software licenses.
- **Installation and Configuration:** Expenses involved in setting up and configuring network hardware and software.
- **Maintenance and Upgrades:** Ongoing costs for maintaining, updating, and upgrading network systems to ensure efficiency and security.
- **Training and Support:** Costs related to training staff to operate and manage network systems and ongoing technical support.
- **Energy Consumption:** Operational costs for electricity which can be significant in data centers and network operations.

BPP

Other cost factors

- **Software Licenses**: Expenses related to network management and security software.

- **Manpower:** Costs of hiring qualified IT staff for network setup, management, and troubleshooting.

- **Downtime and Risk Costs:** Financial impact associated with network downtimes, including loss of productivity and potential breach risks.

BPP

Strategies to reduce cost

- **Energy Efficiency:** Investing in energy-efficient hardware to reduce power consumption.
- **Virtualisation:** Using virtual machines and software-defined networking to optimise resource usage.
- **Cloud Services:** Outsourcing certain network functions to cloud providers to reduce on-premises hardware needs.
- **Preventative Maintenance:** Regular maintenance to prevent costly downtime and repairs.

BPP

Data compression

- Reducing the size of data to save storage space and decrease transmission times.
    - **Lossless**: Compression where the original data can be perfectly reconstructed from the compressed data (e.g., ZIP files).
    - **Lossy**: Compression where some data is lost but the result is good enough for the purpose (e.g., multimedia files).
- **Benefits**: Reduced bandwidth usage, faster transmission speeds, and lower energy consumption.

BPP

Transfer optimisation

- **Caching:** Storing copies of frequently accessed data closer to the user to reduce data retrieval times and bandwidth usage.
- **Load Balancing:** Distributing network traffic across multiple servers to optimize resource use and minimise response times.
- **Traffic Shaping:** Prioritising certain types of data to ensure critical applications have the bandwidth they need.
- **Benefits:** Enhanced user experience, reduced server load, and minimized energy consumption due to efficient data handling.

BPP

Network sustainability - terms

- **Sustainability Concerns**: Environmental impact of network operations, particularly energy usage and electronic waste.

- **Green Networking:** practices and technologies that reduce the environmental footprint of network infrastructure.

- **Net-zero emissions:** the balance between the amount of greenhouse gas emissions produced and the amount removed from the atmosphere.

BPP

Case Study: Google's Sustainable Networking

Net Zero

- **Data Centre Efficiency:** Google has been improving its data centre energy efficiency through advanced cooling technologies and machine learning for optimum energy use.

- **Renewable Energy:** As of 2017, Google matches 100% of the energy consumed by their global operations with renewable energy and continues to do so.

- **Material Use:** Utilising recycled materials in hardware and committing to zero waste to landfill from their data centres.

BPP

Case Study: Google's Sustainable Networking

Net Zero

- **Video Compression**: Utilising advanced encoding techniques to reduce the size of video files without compromising quality.

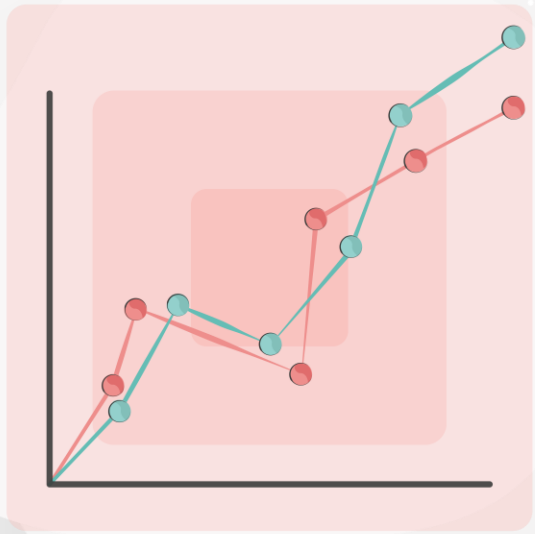
- **Dynamic Optimisation**: Adjusting the compression level based on the user's bandwidth and device capabilities to minimise data usage.

- **Content Delivery Network (CDN)**:

    - **Global Distribution**: Netflix deploys servers worldwide, storing and delivering content locally to reduce data transit distances and energy consumption.
    - **Edge Servers**: Using edge servers to bring data closer to the end-user, significantly reducing latency and energy use during data transmission.

BPP

Worksheet overview



Key Learning Summary

The key takeaways from this session are as follows:

- Understanding the concepts of logical networking, virtualisation, SDN, and NFV is essential for managing modern network infrastructures effectively.
- By mastering distributed data products and modern cluster management solutions you will be well-equipped to handle large datasets efficiently and reliably, ensuring optimal performance and reliability in real-world applications.
- Sustainable network practices are crucial in our collective effort to combat environmental challenges
- Understanding infrastructure costs is crucial for efficient budgeting, strategic planning, and making informed network investment decisions

BPP