

Cloud Computing

Prepared by Dr. Sherif Mostafa

IT381

Lecture 6



Physical Network for Virtualization

The Role of Networks in Cloud Computing

Cloud Deployment Models and the Network

Data center network architecture

Switch virtualization

Layer 2 and Layer 3 Switches

Layered Architecture of Cloud

Operating system based Virtualization

The Role of Networks in Cloud Computing

The confluence of technology advancements and business developments in Broadband Internet, Web services, computing systems, and application software over the past decade has created a perfect storm for cloud computing. The “cloud model” of delivering and consuming IT functions as services is poised to fundamentally transform the IT industry and rebalance the inter-relationships among end users, enterprise IT, software companies, and the service providers in the IT ecosystem

The Role of Networks in Cloud Computing

Networking is in the center of the cloud delivery and consumption model. The network serves as the linkage between the end users consuming cloud services and the provider's data centers providing the cloud services.

In addition, in large-scale cloud data centers, tens of thousands of compute and storage nodes are connected by a data center network to deliver a single-purpose cloud service.

Cloud Deployment Models and the Network

The IT industry is attracted by the simplicity and cost effectiveness represented by the cloud computing concept that IT capabilities are delivered as services in a scalable manner over the Internet to a massive number of remote users.

The IT industry views cloud computing – an emerging business model – as a new way to solve today's business challenges.

Business executives from different enterprises identified “reduce capital cost,” “reduce IT management cost,” “accelerate technology deployment,” and “accelerate business innovation” as the main business benefits for cloud computing.

Cloud Deployment Models and the Network

Despite the benefits promised by cloud computing, the IT industry also sees that significant innovation and improvement on technologies and operations governance are needed to enable broad adoption of cloud services. Chief concerns are security and performance issues.

Take security as an example, while it is acceptable for individual consumers to turn to Amazon Elastic Compute Cloud (EC2) and Simple Storage Services (S3) for on-demand compute resources and storage capacities, it is a different matter for a bank to store its customer information in a third-party owned cloud.

Based on the differences in the deployment model, cloud services can be delivered in three principal ways: Public cloud, Private cloud, and Hybrid cloud.

A **public cloud** refers to a cloud service delivery model in which a service provider makes massively scalable IT resources, such as CPU and storage capacities, or software

applications, available to the general public over the Internet.

Private cloud represents a deployment model where enterprises (typically large corporations with multi-location presence) offer cloud services over the corporate network (can be a virtual private network) to its own internal users behind a firewall-protected environment.

Hybrid cloud is a deployment model for cloud services where an organization provides cloud services and manages some supporting resources in-house and has others provided externally as blending the characteristics of public and private clouds

An Overview of Network Architectures for Clouds

There are three principal areas in which the network architecture is of importance to cloud computing:

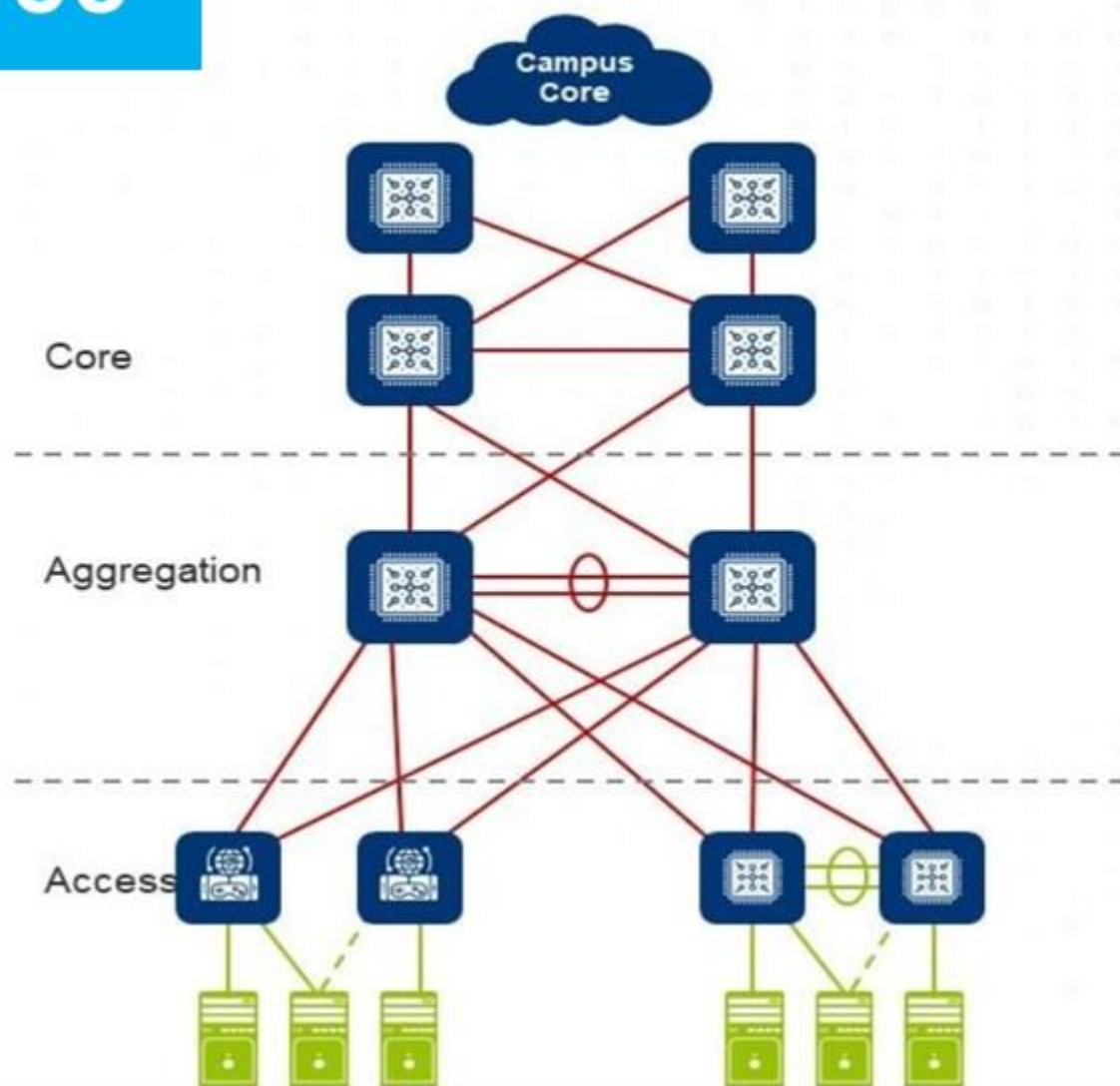
- (1) Data center network that interconnects the infrastructure resources (e.g. servers and storage devices) within a cloud service data center,
- (2) Data center interconnect network that connects multiple data centers in a private, public, or hybrid cloud to supporting the cloud services,
- (3) Public Internet that connect end users to the public cloud provider's data centers.

Data Center Network

Cloud providers offer scalable cloud services via massive data centers. In such massive-scale data centers, Data Center Network (DCN) is constructed to connect tens, sometimes hundreds, of thousands of servers to deliver massively scalable cloud services to the public.

Hierarchical network design is the most common architecture used in data center networks. The following Figure shows a conceptual view of a hierarchical data center network as well as an example of mapping the reference architecture to a physical data center deployment.

Data center network architecture



Fundamental basis of data center architecture is the layered approach, which aims to increase scalability, speed, adaptability, robustness, and maintenance

• **Core Layer**

- Offers a high-speed packet switching backplane for all data center traffic in and out.
- Connects several aggregation modules and creates a Layer 3 routed fabric that is robust and has no single point of failure.

• **Aggregation Layer Modules**

- Offers essential capabilities including service module integration, Layer 2 domain definitions, spanning tree processing, and default gateway redundancy.
- Server-to-server multi-tier traffic passes through the aggregation layer, where it may make use of services like firewalls and server load balancing to improve and protect applications.

• **Access Layer**

- Where the servers connect to the network physically.

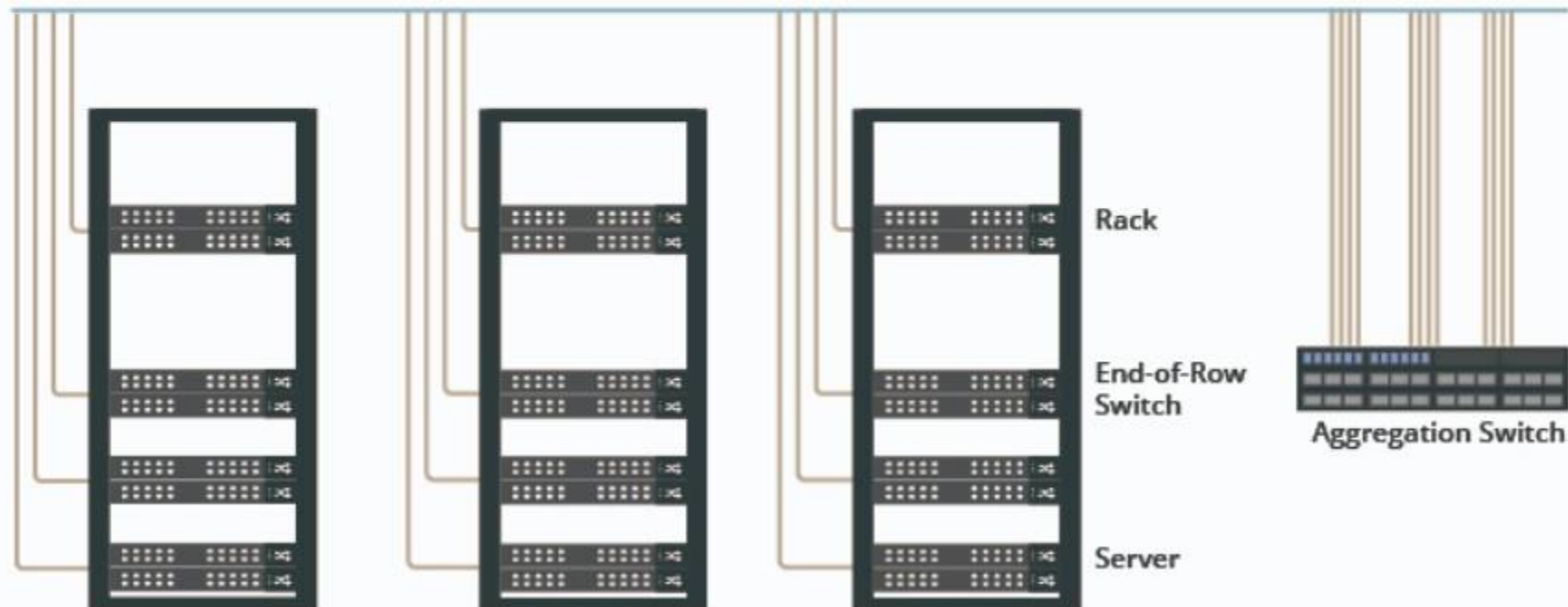
The access layer of a data center network provides connectivity for server resource pool residing in the data center. Design of the access layer is heavily influenced by the decision criteria such as server density, form factor, and server virtualization that can result in higher interface count requirements. The commonly used approaches for data center access layer connectivity are end-of-row (EoR) switch, top-of-rack (ToR) switch, and integrated switch (typically in the form of blade switches inside a modular blade server chassis). Another form of the integrated switch is the embedded software switch in a server end point. Each design approach has pros and cons, and is dictated by server hardware and application requirements.

In End-of-row EoR network design, there is a direct connection of each server in the rack with the end of row aggregation switch. This eliminates the need to connect servers directly with the in-rack switch.

Racks are normally arranged in such a way that they form a row, a cabinet or rack is positioned at the end of this row. This rack has the row aggregation switch, which provides network connectivity to servers mounted in individual racks.

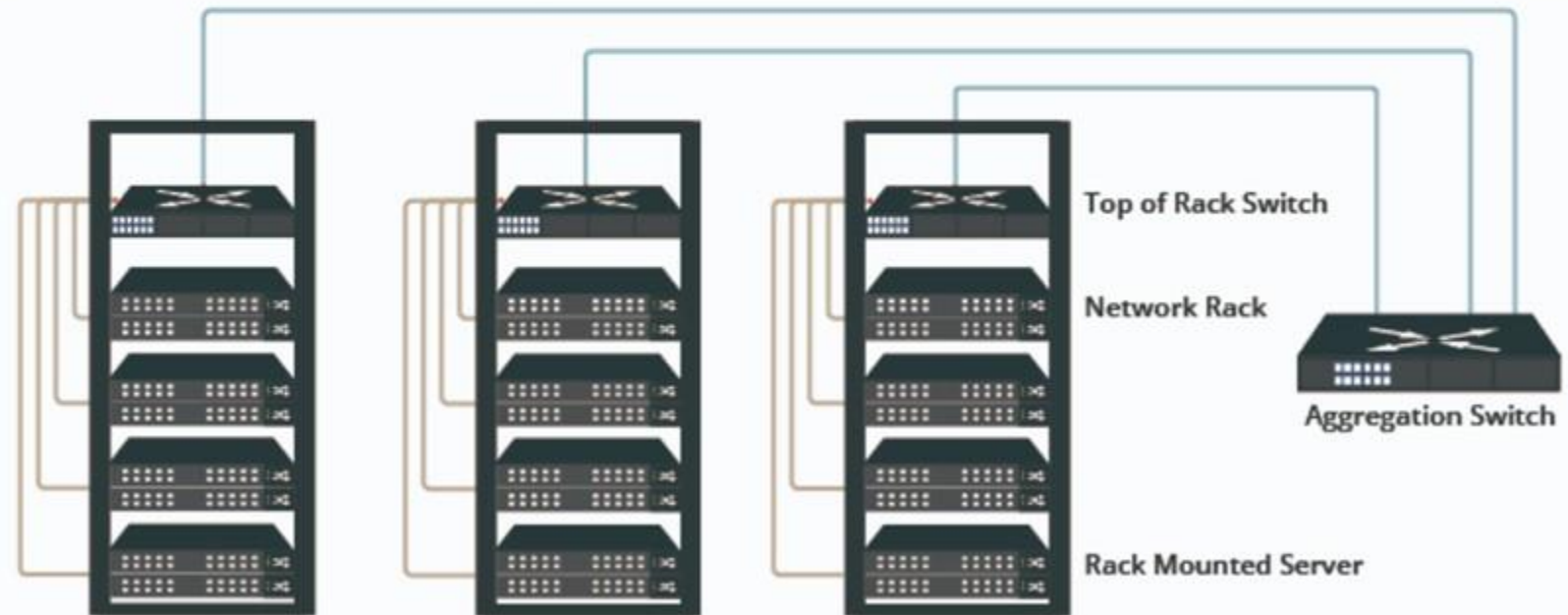
This switch sometimes supports hundreds of server connections. A large amount of cabling is required to support this architecture.

In ToR each rack is an independent unit whereas in EoR the whole row of servers acts as a group within the data center. Any issue with the row aggregation switch impacts the complete row of servers.



Top of rack (ToR) which is also known as In-Rack design. In this approach, the network access switch is placed on the top of the server rack; hence, servers are directly connected to the network access switch. This means that 1 or 2 Ethernet switches are directly installed inside the rack, therefore copper cables stay inside the rack. It is cost-effective because it reduces the number of copper cables between racks. The rack is linked to the data center network by an Ethernet switch, often through a fiber cable. This fiber cable is a direct link from the common aggregation area to the rack.

In the ToR approach, every rack in the data center network is a separate entity that eases its management. Any change, upgrade, or malfunction in the rack usually affects that rack only. Fewer cables mean that one can opt for better quality and higher bandwidth cables in the same budget.



End-of-Row Switches vs. Top-of-Rack Switches:

When choosing between End-of-Row Switches and Top-of-Rack Switches, it's essential to understand their distinct characteristics and use cases.

Deployment: End-of-Row Switches are ideal for large-scale data centers with extensive cabling needs, while Top-of-Rack Switches suit smaller setups with limited cabling.

Management: End-of-Row Switches offer centralized management, making them suitable for enterprises with complex networks. Top-of-Rack Switches provide localized control, beneficial for mid-market companies seeking simplicity.

The Aggregation layer

The Aggregation layer of the data center provides a consolidation point where access layer switches are connected providing connectivity between servers for multi-tier applications, as well as connectivity across the core of the network to the clients residing within the campus, WAN, or Internet.

The aggregation layer typically provides the boundary between Layer-3 routed links and Layer-2 Ethernet broadcast domains in the data center. The access switches are connected to the aggregation layer using 802.1Q VLAN trunks to provide the capability of connecting servers belonging to different VLANs and IP subnets to the same physical switch.

The primary function of **the core layer** in a data center network is to provide highly available, high-performance Layer-3 switching for IP traffic between the data center and the Internet backbone.

In some situations, multiple geographically distributed data centers owned by a cloud service provider may be connected via a private WAN or a Metropolitan Area Network (MAN). For such environments, expanding Layer 2 networks across multiple data centers is a better architecture design

In other situations, the traffic geographically distributed data centers has to be carried over the public Internet.

By configuring all links connecting to the network core as point-to point Layer-3 connections, rapid convergence around any link failure is provided, and the control plane of the core switches is not exposed to broadcast traffic from end node devices or required to participate in STP for Layer-2 network loop prevention.

Large-scale data centers

The evolution of networking technology to support large-scale data centers is most evident at the access layer due to rapid increase of number of servers in a data center. Some research work calls for a large Layer-2 domain with a flatter data center network architecture (2 layers vs. 3 layers).

While this approach may fit a homogenous, single purpose data center environment

A more prevalent approach is based on the concept of switch virtualization which allows the function of the logicalLayer-2 access layer to span across multiple physical devices. There are several architectural variations in implementing switch virtualization at the access layer. They include:

- Virtual Blade Switch (VBS)
- Fabric Extender
- Virtual Ethernet Switch technologies

Switch virtualization

The Virtual Blade Switch VBS approach allows multiple physical blade switches to share a common management and control plane by appearing as a single switching node.

The Fabric Extender approach allows a high-density, high-throughput, multi-interface access switch to work in conjunction with a set of fabric extenders serving as “remote I/O modules” extending the internal fabric of the access switches to a larger number of low-throughput server access ports.

The Virtual Ethernet Switch is typically software based access switch integrated inside a hypervisor at the server side.

Switch virtualization

These switch virtualization technologies allow the data center to support multi-tenant cloud services and provide flexible configurations to scale up and down the deployment capacities according to the level of workloads.

Data Center Interconnect Network

Data center interconnect networks (DCIN) are used to connect multiple data centers to support a seamless customer experience of cloud services. While a conventional, statically provisioned virtual private network can interconnect multiple data centers and offer secure communications, to meet the requirements of seamless user experience for cloud services (high-availability, dynamic server migration, application mobility), the DCIN for cloud services has emerged as a special class of networks based on the design principle of Layer 2 network extension across multiple data centers

Data Center Interconnect Network

For example, in the case of server migration (either in a planned data center maintenance scenario or in an unplanned dynamic application workload balancing scenario) when only part of the server pool is moved at any given time, maintaining the Layer 2 adjacency of the entire server pool across multiple data centers as opposed to renumbering IP addresses of servers is a much better solution. The Layer 2 network extension approach, on one hand, is a must from business-continuity perspective; on the other hand, is cost effective from the operations perspective because it maintains the same server configuration and operations policies.

Understanding the Differences Between Layer 2 and Layer 3 Switches

For decades there were switches, and then there were routers. It was obvious that switches performed the switching, and routers did the routing. But in the past few years, there has been the emergence of “Layer 3 switches,” which has raised questions for some about the difference between Layer 2 and Layer 3. i.e., the differences between switching and routing, and which type of switch should be deployed in which situations.

Understanding the Differences Between Layer 2 and Layer 3 Switches

Layer 2 and Layer 3. Layer 2 switches operate at the data link layer (layer 2) of the OSI model and forward data packets based on the MAC addresses of the devices on a LAN. Layer 3 switches also support routing capabilities at the network layer (layer 3) and can forward packets based on both MAC and IP addresses. A Layer 3 switch can perform both switching and routing, making them more versatile than a traditional Layer 2 switch. When choosing an Ethernet switch, it's essential to consider the needs of your network and the level of functionality required.

How Layer 2 works

Layer 2, called the “Data Link Layer” transfers data based on physical MAC (media access control) addresses. Every device that has a network interface controller (NIC) has a unique MAC address, assigned by the manufacturer of the device. You may wonder how every network-enabled device can have a unique MAC address, when there are at least billions of such devices in the world. Well, MAC addresses use this format:

D4-BE-D9-8D-46-9A

Allowing for over 281 trillion possible addresses.

How Layer 2 works

So a Layer 2 switch transfers data based on known MAC addresses. But how does it know the MAC addresses of the devices connected to it? Whenever the switch receives a data frame, it examines the frame contents for the source and destination MAC addresses. If the source address is new to the switch, it adds the address to a “MAC Address Table”. This table is usually very simple, only containing port numbers, MAC addresses, VLAN, and sometimes whether the address was learned dynamically or statically.

How Layer 2 works

If the destination MAC address of the data frame is in the switch's MAC address table, it forwards the packet to the corresponding port in the table. If the destination address is unknown to the switch, it sends the frame out all ports except the one on which it was received.

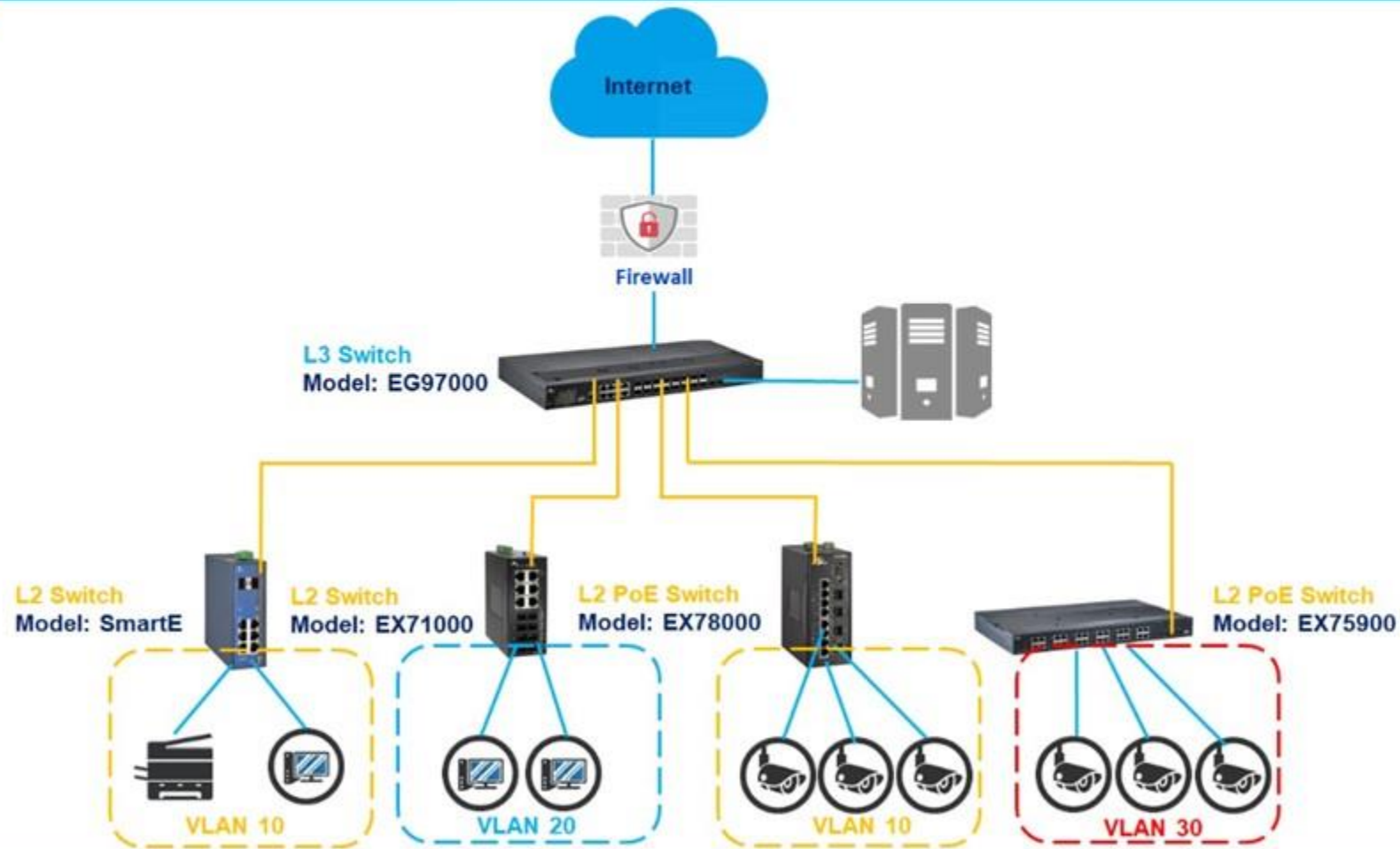
How Layer 2 works

Requiring no routing algorithm, and not needing IP addresses to forward data, Layer 2 switches are very fast, and cost less than routers. However, broadcast traffic, that is traffic that is sent to every device on the network, is not controlled by Layer 2 switches. This can lead to network congestion on large networks. Lastly, Layer 2 switches cannot pass data between different VLANs.

When to choose a Layer 2 or a Layer 3 switch

For small networks where the volume of data transmitted is not too great and there is no need to connect multiple VLANs, Layer 2 switches are ideal. For larger networks, networks needing VLAN connectivity, or in situations where enhanced security is required, a Layer 3 switch is the way to go. Most networks use a combination of Layer 2 and Layer 3 switches to optimize cost and performance.

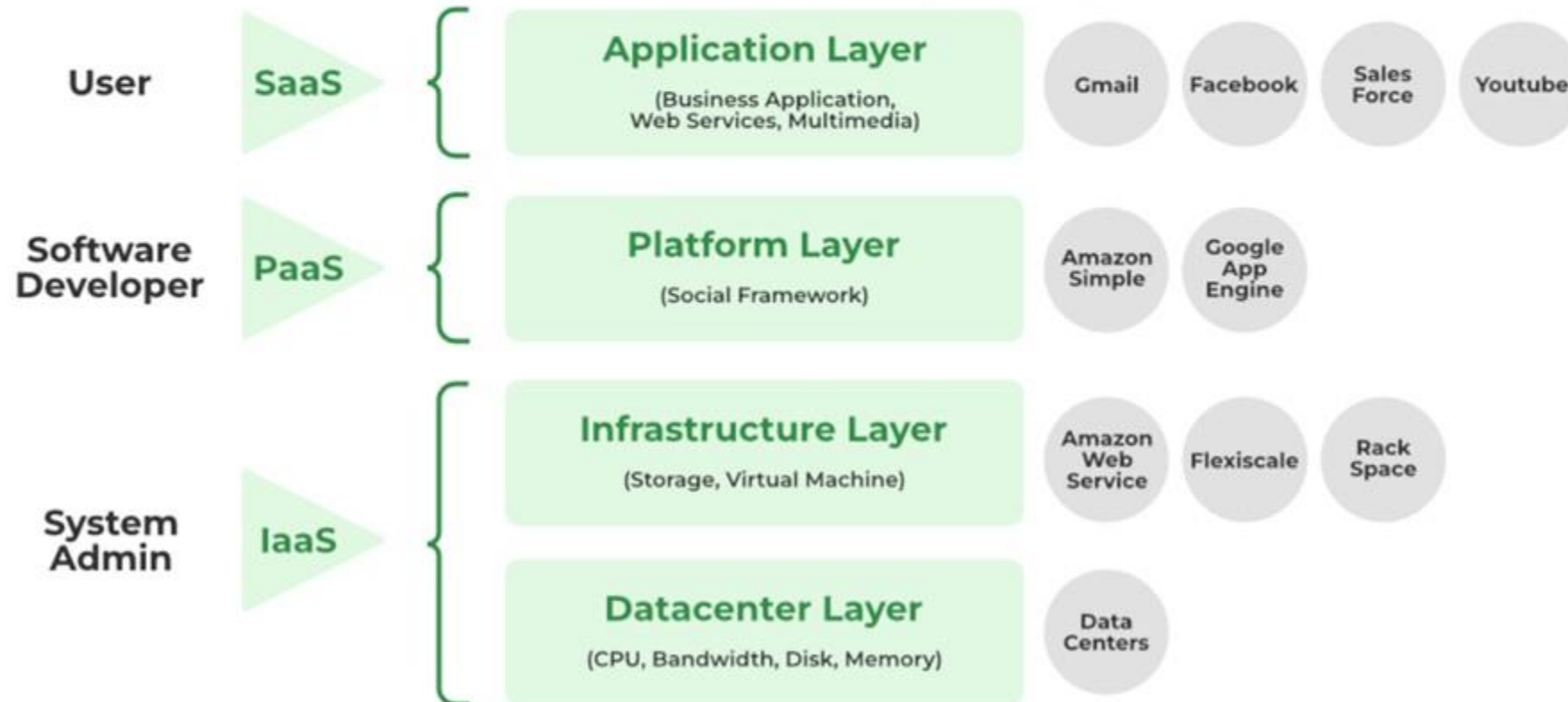
When to choose a Layer 2 or a Layer 3 switch



When to choose a Layer 2 or a Layer 3 switch

A Layer 2 managed switch is designed to forward traffic between network hosts within the same subnet, based on the entries in its MAC address table. On the other hand, a Layer 3 managed switch is capable of forwarding traffic between different subnets, using a map of the IP network maintained in its routing table. In other words, if you need to send traffic between devices on different subnets, you will need a Layer 3 switch. It's important to carefully consider your networking needs and choose the right type of managed switch to ensure smooth and efficient operation.

Cloud Computing Layers



Layered Architecture of Cloud

Application Layer

The application layer, which is at the top of the stack, is where the actual cloud apps are located. Cloud applications, as opposed to traditional applications, can take advantage of the automatic-scaling functionality to gain greater performance, availability, and lower operational costs.

This layer consists of different Cloud Services which are used by cloud users. Users can access these applications according to their needs. Applications are divided into Execution layers and Application layers.

Layered Architecture of Cloud

Application Layer

In order for an application to transfer data, the application layer determines whether communication partners are available. Whether enough cloud resources are accessible for the required communication is decided at the application layer. Applications must cooperate in order to communicate, and an application layer is in charge of this. The application layer, in particular, is responsible for processing IP traffic handling protocols like Telnet and FTP. Other examples of application layer systems include web browsers, SNMP protocols, HTTP protocols, or HTTPS, which is HTTP's successor protocol.

Layered Architecture of Cloud

Platform Layer

The operating system and application software make up this layer.

Users should be able to rely on the platform to provide them with Scalability, Dependability, and Security Protection which gives users a space to create their apps, test operational processes, and keep track of execution outcomes and performance. SaaS application implementation's application layer foundation.

Layered Architecture of Cloud

Platform Layer

The objective of this layer is to deploy applications directly on virtual machines.

Operating systems and application frameworks make up the platform layer, which is built on top of the infrastructure layer. The platform layer's goal is to lessen the difficulty of deploying programmers directly into VM containers.

By way of illustration, Google App Engine functions at the platform layer to provide API support for implementing storage, databases, and business logic of ordinary web apps.

Layered Architecture of Cloud

Infrastructure Layer

It is a layer of virtualization where physical resources are divided into a collection of virtual resources using virtualization technologies like Xen, KVM, and VMware.

This layer serves as the Central Hub of the Cloud Environment, where resources are constantly added utilizing a variety of virtualization techniques.

A base upon which to create the platform layer. constructed using the virtualized network, storage, and computing resources. Give users the flexibility they want.

Automated resource provisioning is made possible by virtualization, which also improves infrastructure management.

Layered Architecture of Cloud

Infrastructure Layer

The infrastructure layer sometimes referred to as the virtualization layer, partitions the physical resources using virtualization technologies like Xen, KVM, Hyper-V, and VMware to create a pool of compute and storage resources.

The infrastructure layer is crucial to cloud computing since virtualization technologies are the only ones that can provide many vital capabilities, like dynamic resource assignment.

Layered Architecture of Cloud

Datacenter Layer

In a cloud environment, this layer is responsible for Managing Physical Resources such as servers, switches, routers, power supplies, and cooling systems.

Providing end users with services requires all resources to be available and managed in data centers.

Physical servers connect through high-speed devices such as routers and switches to the data center.

Layered Architecture of Cloud

Datacenter Layer

In software application designs, the division of business logic from the persistent data it manipulates is well-established. This is due to the fact that the same data cannot be incorporated into a single application because it can be used in numerous ways to support numerous use cases. The requirement for this data to become a service has arisen with the introduction of microservices.

Layered Architecture of Cloud

Datacenter Layer

A single database used by many microservices creates a very close coupling. As a result, it is hard to deploy new or emerging services separately if such services need database modifications that may have an impact on other services. A data layer containing many databases, each serving a single microservice or perhaps a few closely related microservices, is needed to break complex service interdependencies.

Physical Data Center Equipment

When discussing physical infrastructure, we are referring to hardware found in a data center, including servers and storage devices. However there are other core data center network components highlighted below.

Cabling

Cabling, such as copper and fiber optic cables, enables various devices to be connected within the data center. Similar to cabling found in homes, data center cabling acts no different and helps with high-speed data transmission between devices.

Power Supply

Power supply is of course a critical component to IT infrastructures, without it hardware simply couldn't operate. Power distribution units (PDUs) are responsible for controlling power, such as distributing electricity to racks of computers and networking equipment.

Given how important Uptime is for data centers, fallback power supply is needed if the primary source of electricity fails. With that in mind, uninterruptible power supply (UPS) systems and backup generators are useful as part of data center network architectures, as they ensure facilities remain functional during outages, by maintaining a stable power supply.

Servers

Servers are essentially computers that deliver data and resources to other devices over a network. In the context of data center network architecture design, servers are critical as they make it possible for databases and resource management to be hosted locally.

Storage Devices

Storage technologies unsurprisingly store and manage data, common examples include storage area networks (SANs), network attached storage (NAS), and direct attached storage (DAS). A critical component, storage devices ensure data is always available for transferring to other computer devices when required.

Physical Data Center Equipment

Racks, Enclosures, and Cabinets

Racks, enclosures and cabinets are different frameworks used to store and protect equipment within a data center.

An often overlooked aspect of data center network infrastructure design is how racks and cabinets are organized. Proper rack management is vital in ensuring optimal airflow between devices, which enhances the performance and longevity of data center equipment and encourages sustainable operations by increasing power use effectiveness (PUE).

Physical Data Center Equipment

Cooling Systems

Maintaining optimal temperature levels within the data center is essential for infrastructure devices to function properly and not overheat. Cooling systems, such as air conditioning and liquid cooling solutions (immersion and direct-to-chip cooling) prevent overheating, hardware failures, and eventual data loss.



Thank You