Foundation of Cloud IoT Edge ML
Professor Doctor Rajiv Mishra
Department of Computer Science and Engineering
Indian Institute of Technology, Patna
Lecture 02
Introduction to Cloud

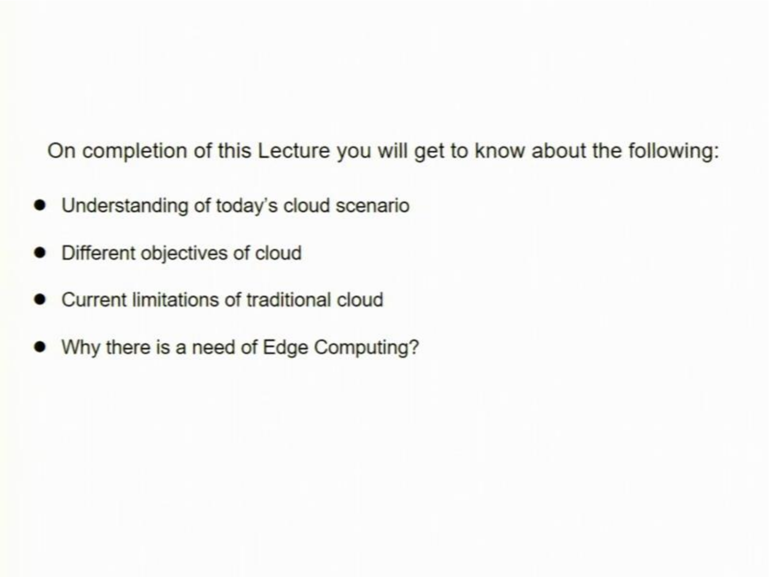
(Refer Slide Time: 00:20)

Introduction to Cloud

Dr. Rajiv Misra
Professor, Dept. of Computer Science & Engg. Indian Institute of Technology Patna rajivm@iitp.ac.in

Good morning, myself Doctor Rajiv Mishra, I am from IIT Patna. So, today's lecture topic is introduction to cloud.

(Refer Slide Time: 00:25)

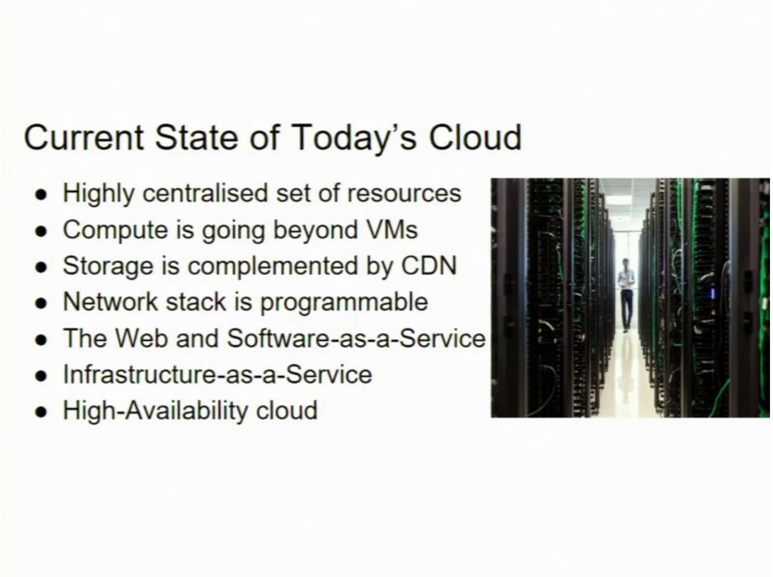


On completion of this Lecture you will get to know about the following:

- Understanding of today's cloud scenario
- Different objectives of cloud
- Current limitations of traditional cloud
- Why there is a need of Edge Computing?

So, this particular lecture after completion you will be knowing the following. First is to understand of today's cloud scenario. Second is you will have the knowledge of different objectives of cloud. Third one is that you will be knowing about the current limitations of the classical cloud and then after that, you will be knowing to know why there is a need of edge computing.

(Refer Slide Time: 00:58)



Current State of Today's Cloud

- Highly centralised set of resources
- Compute is going beyond VMs
- Storage is complemented by CDN
- Network stack is programmable
- The Web and Software-as-a-Service
- Infrastructure-as-a-Service
- High-Availability cloud

So, let us get started. So, in this particular scenario of today's cloud, we will now explore the current state of the classic cloud. So, in that, you will be knowing what do you mean by the highly centralized set of resources, which is the characteristics of today's cloud. The second important concept of today's cloud is that, that the compute which is these days called as what your machines now, how this compute is going beyond what your machines we will know about that in today's cloud scenario.

Third thing we have to know that, that the cloud which is also used for the storage, so, now, what is the newer development in that context is that the storage which cloud provides is complemented by the CDN that is content delivery network. So, this also we will talk about briefly in today's cloud important thing about the today's cloud is the network stack. So, this network stack is now being made as programmable.

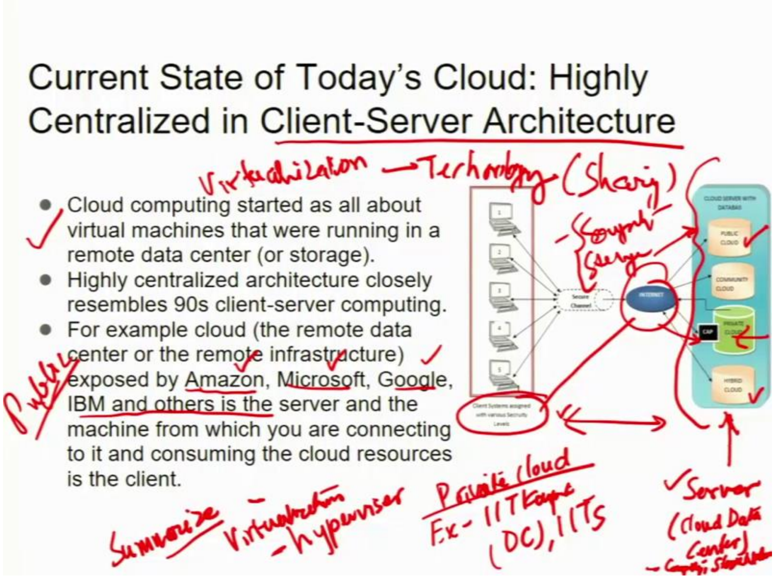
So, how that is all done, what is the technology which is used in the growth of today's cloud we will see this also in great details, then the web and software as a service, which is an application of today's cloud, then Infrastructure as a Service is also the model in which the cloud is being offered by various cloud providers or vendors, AI in the highly available cloud is the resultant of all these new innovations.

So, here in this figure, you can see that this is the place which is called as a data center and this data center has 100s and 1000s of computer nodes. These computing nodes are then collected with the networking, which is called a data center networking i in this particular data center, every cloud vendor has these particular data centers located at different geographic positions across the globe, which are in turn connected by the different data center networking.

So, inter intra data center networking is that called networking of a cloud, which is there. So, we will talk about that how the cloud is providing you the services what is the model. Now, as far as earlier cloud is seen as highly centralized set of resources. So, you can see here in the cloud data center, if you can see that hundreds and 1000s of computing nodes are connected via the network.

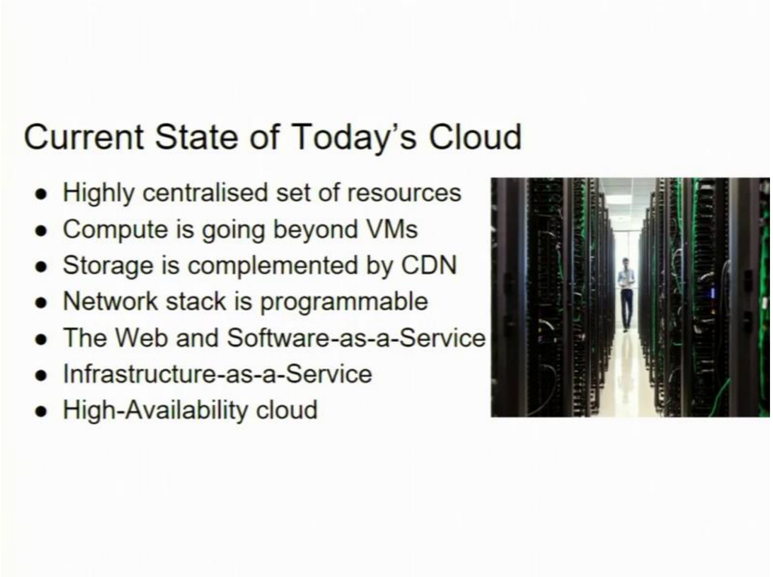
So, this particular model, which is called Data Center will contain the highly centralized set of resources such as compute, storage, and the networking. So, all these three different segments will be treated as highly centralized set of resources, which today's cloud will provide to differ different applications to run on it. Now, let us see what you mean by highly centralized set of resources, which is being supported by these data centers.

(Refer Slide Time: 04:44)



Current State of Today's Cloud: Highly Centralized in Client-Server Architecture

- Cloud computing started as all about virtual machines that were running in a remote data center (or storage).
- Highly centralized architecture closely resembles 90s client-server computing.
- For example cloud (the remote data center or the remote infrastructure) exposed by Amazon, Microsoft, Google, IBM and others is the server and the machine from which you are connecting to it and consuming the cloud resources is the client.



Current State of Today's Cloud

- Highly centralised set of resources
- Compute is going beyond VMs
- Storage is complemented by CDN
- Network stack is programmable
- The Web and Software-as-a-Service
- Infrastructure-as-a-Service
- High-Availability cloud

So, the current state of today's cloud which is also known as highly centralized set of resources follows an architecture which is called a client-server architecture. So, the cloud computing is started at is all about this virtual machine and that we're running in the remote data center or let us say storage system. So, that we have shown in the previous slide that you can see here, this is called a data center.

So, this highly centralized architecture closely resembles 90s client-server computing. So, client-server architecture has a client and a server. So, that particular data center which you have seen is mimics the server. Whereas, the client system which is the machines which the client used to connect through the network or through the internet to the cloud data center is called a client.

So, client and the server is the model, or is the architecture which cloud provides the services in this way. Now, for example, you can think of that the cloud which is also the remote data center or a remote infrastructure. And this remote infrastructure is exposed by these different providers such as Amazon, Microsoft, Google, IBM, and many others is nothing but a server in this architecture, and the machines from which you are connecting to it and consuming the computing resources or cloud resources is called as the client.

So, in this example, you can see that this is the server which is now also provided with the help of a cloud data center, which comprises of the resources such as the computing storage and the networking. Now, as far as this particular server or in this architecture, which is called a client-server architecture, in which this today's cloud is operating, is being supported by provided by different vendors such as Amazon, Microsoft, Google, IBM, and many others.

Now, there are different models of this architecture to be supported and they are called as a public cloud, they let us say that this entire data center all the data center or the computing infrastructure or resources, if it is provided by these vendors, then they are called as a public cloud whereas, if it is supported by that organization itself, not taking the help of these external vendors like Amazon, Google Microsoft, if it is provided internally, let us say for example, let us say IIT Kanpur can have these kinds of data center of its own and running this architecture that is client-server architecture then that cloud is called the private cloud.

So, this is an example of a private cloud that is called IIT, Kanpur, or some other IITs, who runs their own data center. And public cloud means that, if you use this computing resources as a service, which comes from either Microsoft image and Google IBM, then it is called a public cloud and the hybrid is the combination of both that is the organization's which uses their data center as well as they also extended so, that it uses the public infrastructure or the public cloud also, if both the clouds are there, then it is called the hybrid cloud.

To summarize, what we have seen here in this particular slide is that, that today's cloud which provides highly centralized resources, such as compute storage and the networking is often is supported by an architecture which is called a client-server architecture. So, the soda machines are the users if we are trying to get this access to the highly centralized computing and storage resources, then, we will become the client machine because our machines are not having that kind of we are going to access the highly centralized resources to the internet.

So, so these machines which connects to the cloud, whether it is a public cloud or it is a private cloud then this through the internet, then it is called the client machine. Whereas on the other side, these highly centralized resources of computing and storage is provided through the place which is called the data center. And in this client-server architecture, we call it as a server.

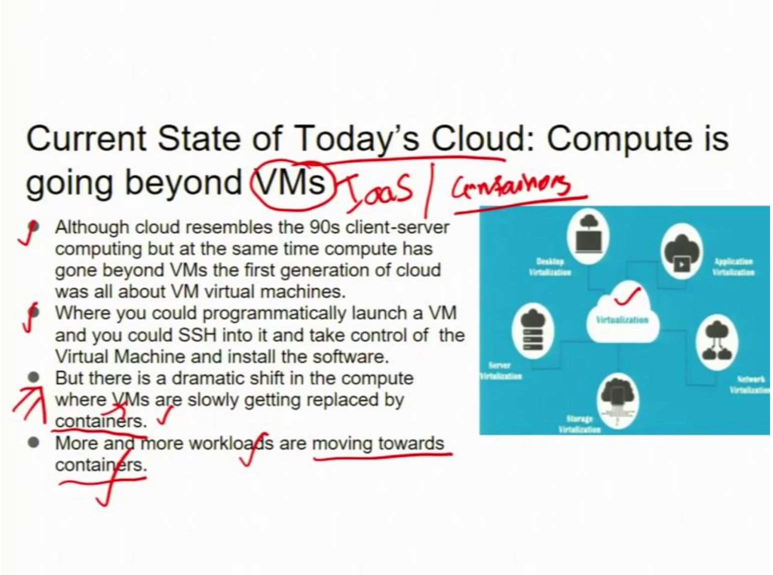
So, this client-server architecture is example, which is shown over here and it provides the cloud computing to support the various applications. Now, this particular server or a cloud which is having which is comprised of 100s and 1000s of machines, which are housed in a place which is called a data center. Now, how that is all being provided as a service. So, the technology which is called the virtualization, so, what you ideation is the key technology what you ideation is the key technology for this particular today's cloud.

So, this particular technology will allow sharing. So, virtualization is nothing but it allows the sharing of computing storage, computing storage, and networking. So, that particular technology uses the concept which is called the virtualization. And in this virtualization, it often applies the concept which is called the hypervisors.

So, a server with the help of hypervisor is virtualized or is virtualizing its compute so that different users can simultaneously use that computing as the service that is the same server is now virtualized will become the virtual server or a virtual machine. So, virtual machines comprises of all CPU, or let us say what your CPU that is called compute storage and the networking is all virtualized.

So, multiple users can coexist on multiple users' application can coexist, use the help of this technology which is called the virtualization AI in this virtualization often uses will allow the sharing of these resources the model to access this sharing is called a client-server architecture and that is all we have already explains in today's cloud.

(Refer Slide Time: 12:46)



Current State of Today's Cloud: Compute is going beyond VMs IaaS | Containers

- Although cloud resembles the 90s client-server computing but at the same time compute has gone beyond VMs the first generation of cloud was all about VM virtual machines.
- Where you could programmatically launch a VM and you could SSH into it and take control of the Virtual Machine and install the software.
- But there is a dramatic shift in the compute where VMs are slowly getting replaced by containers.
- More and more workloads are moving towards containers.

Now, next thing is that what is the new development so, far you have seen the concept of the virtualization which allows the sharing in the form of virtual machines I in this virtual machine is the machines or the infrastructure as a service called infrastructure as a service model. So, infrastructure as a service means that physical machine which comprises of CPU memory and the network is now virtualized so that it gives the concept of a virtual machine.

So, what you will machines are being given as a service to different vendors or to the different clients. So, these clients will connect with the help of internet and gain the access of virtual machines. So, the compute So, far means, classical cloud was supported by these virtual machines. Now, let us see that this compute is not confined to the virtual machine is going beyond that we will understand in this particular slide.

So, although the compute or the cloud resembles the it is client-server computing, but at the same time compute has gone beyond what your machines. So, the fourth generation of cloud was all about the virtual machines that we have already discussed and seen in the previous slides.

Now, in this concept of these virtual machines that you can programmatically launch a virtual machine and you could do using SSH log into it, and take the complete control of a virtual machine for what purpose you can install your own software applications that is and you run that application on these virtual machines. Now, but there is a dramatic shift in compute, where the virtual machines are slowly getting replaced by the container.

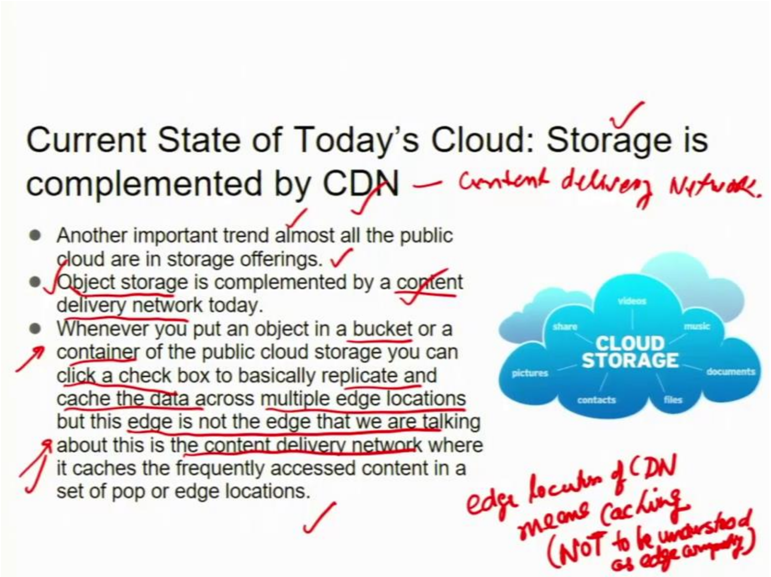
So, this is the newer development which we are which we have to now see and understand. So, that means, this technology of getting this computing or the cloud access, which is called virtual machines are getting replaced by the new technology which is called containers. So, there are advantages of containers that you might have already known in the cloud. So, what we are now tracking or tracing back is the current state of cloud from the classical cloud.

So, classical cloud uses what your machines that we have already seen, and that particular virtual machines uses that technology which is called the virtualization eye, and now that virtual machines are now slowly replaced by the new technique, which is called the containers. So, containers, you can also understand it's a lightweight version of virtual machines.

And once it is lightweight, then it is more efficient to use the computing engine and the cloud as a service. Now, here, with the help of this new technique, which is called the containers, more and more workloads are now moving towards the containers. So, now, you have a choice whether to go for whether to go with the virtual machines or to go with the container. So, we will see that the newer application why they are moving towards containers.

So, containers is the technique by which you can access the cloud resources in contrast to the virtual machines, so, this is there, the trend of today's cloud is that the computing is going beyond the virtual machines, that is nowadays it is used in the form of compute is used in the form of containers which we have already explained. So, more detail about the container we will understand separately in another lecture.

(Refer Slide Time: 16:41)



Current State of Today’s Cloud: Storage is complemented by CDN — Content delivery Network.

- Another important trend almost all the public cloud are in storage offerings.
- Object storage is complemented by a content delivery network today.
- Whenever you put an object in a bucket or a container of the public cloud storage you can click a check box to basically replicate and cache the data across multiple edge locations but this edge is not the edge that we are talking about this is the content delivery network where it caches the frequently accessed content in a set of pop or edge locations.

edge location of CDN means caching (NOT to be understood as edge computing)

Now, another state of today's cloud is in terms of the storage. Now, what we will understand here in this particular session or the slide is that this storage is often complemented by the content delivery network CDN means that now, let us understand that how they storage is complemented by content delivery network and what is the use of it, and why today's cloud or is adopting this storage, which is to be complemented by content delivery network.

So, another important trend almost in all the public clouds are in the form of storage offering not only compute, but storage is also one of the important resource which most of the applications are using out of the cloud as a service. So, this is also often called as a object storage. So, this object storage is now complemented by the technology which is called a content delivery network. So, we will understand what do you mean by CDN as of today.

So, to understand this, let us take this example that whenever you put an object into a bucket or a container of a public cloud storage, you then if you do a click on a checkbox to basically replicate and cache the data across multiple Edge locations. So, that is nothing but the caching at multiple places have that same object or storage, that possibility is now there, provided the client wants it.

So, that it will be that object storage is not only confined sent centralized storage, but it will be replicated at a cross and cash across various data locations that is called Edge locations, but this edge is quite different. This edge is not that the edge computing which we will be talking about later in this particular slide. So, this edge is the caching this edge refers to the caching.

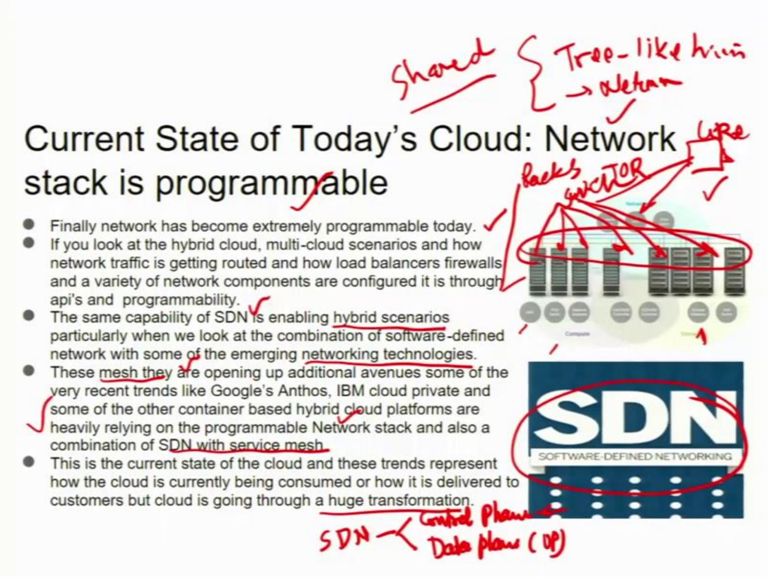
So, the Edge locations of CDN means the caching and it's not to be understood as edge computing. So, that is what we are understood that this object storage is not only stored at one place, if the client wants that this particular object which is stored is to be accessed very quickly or efficiently then caching has to be done at more than one places and those locations where that particular cache or it is the object or the storage is replicated.

Then they are called the edge location. So, this edge is not the edge computing but it is the caching it means the caching, and this caching is done through the technique which is called Content Delivery Network. So, all the public cloud vendors are providers, they complement this storage with this concept called Content Delivery Network.

For example, if let us say you want to watch a movie, and the movie is a file, let us say that a video file and this particular movie if you want to see if more than one people want to see simultaneously, then if you store at one place, then what will happen is that this content access will be done through one single place and it may be slowed down. So, if you instead cache at multiple Edge locations using content delivery network, if you operate, then this serving that particular content or accessing the content will become very fast.

So, this kind of sharing, file sharing video sharing, or image file sharing all these are nowadays and norms in various applications, let us say that Facebook or LinkedIn are many other kinds of services which you see that the users together they cooperatively shape and the content has to be accessed very fast. So, this particular storage is to be complemented by content delivery network, this is another trend of today's cloud.

(Refer Slide Time: 21:29)



Shared
Tree-like hier
→ Network
Racks
switch TOR
SDN
SOFTWARE-DEFINED NETWORKING
SDN
Control Plane
Data plane (DP)

Current State of Today's Cloud: Network stack is programmable

- Finally network has become extremely programmable today.
- If you look at the hybrid cloud, multi-cloud scenarios and how network traffic is getting routed and how load balancers firewalls and a variety of network components are configured it is through api's and programmability.
- The same capability of SDN is enabling hybrid scenarios particularly when we look at the combination of software-defined network with some of the emerging networking technologies.
- These mesh they are opening up additional avenues some of the very recent trends like Google's Anthos, IBM cloud private and some of the other container based hybrid cloud platforms are heavily relying on the programmable Network stack and also a combination of SDN with service mesh.
- This is the current state of the cloud and these trends represent how the cloud is currently being consumed or how it is delivered to customers but cloud is going through a huge transformation.

Now, another important thing is that the networking you know that if the data center or the today's cloud, which is nothing also called as a server in the client-server architecture. Now, these particular so many number of nodes 100 1000s of computing nodes, they are to be networked together and this network now when you say that what you will machine so, what you will machines means that the compute storage and the network, so, network is also shared across multiple clients simultaneously.

So, this particular network, if let us say that if it is programmable, then this virtual machine access will become quite flexible and efficient also, for example, you might have seen that although your internet or getting access to any other application is slower, but sometimes if you connect to the Google that access to the Google is faster. So, that becomes the sharing or a possibility of sharing by the Google let us say that network.

So, let us understand about this current state of today's cloud that is called the network stack, why this is programmable, and this is the newer development that is there in today's cloud. So, finally, the network has become extremely programmable these days, I have explained to you just now about that. So, if you look at the hybrid cloud multi-cloud scenarios, there can see that how the network traffic is getting routed, and how the load balancers firewall, and a variety of network components are configured it is through API's and programmability.

So, therefore, this particular capability of programmability of today's network is enabled with the help of a technology which is called a software-defined networking called SD ins. So, this SDN is enabling this hybrid scenarios, particularly when we look at the combination of a software-defined network with some of the emerging networking technologies. So, this particular mesh, why it is called a mesh because inside data center where you have hundreds and 1000s of computing nodes which are connected with each other forming a mesh.

So, these mesh, they are opening up additional avenues. And some of these very recent trends you can see in the offering such as clouds and costs and IBM, cloud private and some other container-based hybrid cloud platforms are heavily relying on programmable network stack AI and also a combination of SDN with the service mesh.

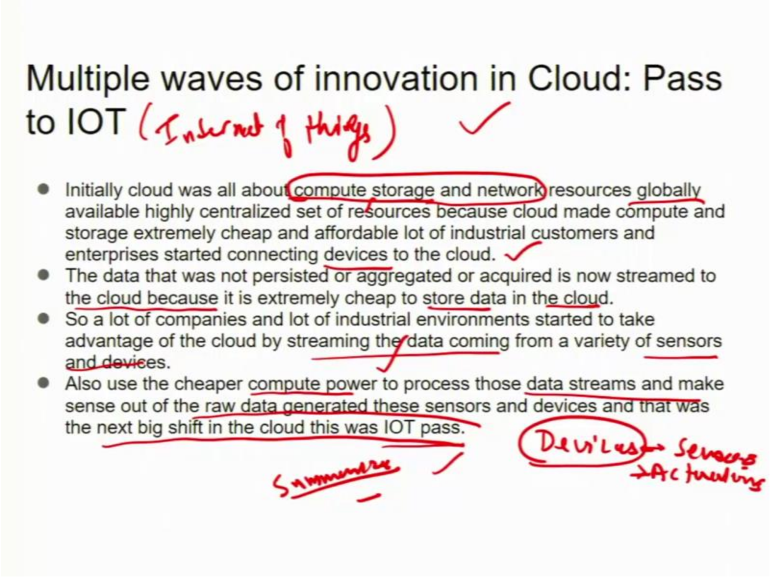
So, this is the current state of the cloud and these trends represents how the cloud is currently being consumed or how it is being delivered to the customers, but the cloud is going through the huge transformation that is not only confined as the highly centralized set of resources. So, you can also understand through this particular figure that you have so many number of computing node and storage nodes, often they are connected by the network.

So, these are called racks. And every rack on top of it, there is a switch and network switch. This is called top of the rack switch it is called these top of the rack switch in turn they are connected again by another level of switch called a core switch. And these are the example of a core switch, this is all the networking on this is the tree-like hierarchy in which this networking is there.

Now, you can see that networking has to be programmable that is if multiple This is to be shared this networking if you want to be shared or virtualized, then it has to be programmable and this sharing of the networking is if it is programmable and the concept which is being brought by the help of software-defined network, so, as DNS when they virtualize the network, they divide the network into two parts.

One is called control plane the other is called the data plane or user plane also. So, this particular data plane that is means that this control plane means that you can program this particular network so that this particular network is stack is programmable based on different vendors or the provided. So, this is another current trend of today's cloud.

(Refer Slide Time: 26:23)



Multiple waves of innovation in Cloud: Pass to IOT (Internet of things) ✓

- Initially cloud was all about compute storage and network resources globally available highly centralized set of resources because cloud made compute and storage extremely cheap and affordable lot of industrial customers and enterprises started connecting devices to the cloud. ✓
- The data that was not persisted or aggregated or acquired is now streamed to the cloud because it is extremely cheap to store data in the cloud.
- So a lot of companies and lot of industrial environments started to take advantage of the cloud by streaming the data coming from a variety of sensors and devices.
- Also use the cheaper compute power to process those data streams and make sense out of the raw data generated these sensors and devices and that was the next big shift in the cloud this was IOT pass.

Devices → Sensors
→ Actuating

Summarize ✓

Now, you can see that there are multiple waves of innovation which is happening inside the cloud AI and we will trace back in this course, the innovation in the cloud which is nothing but a path to the IoT. So, what do we mean by the past to IoT or IoT paths, how this cloud is now transforming because of the support to the Internet of Things or IoT, IoT means internet of things.

So, let us see that initially, the cloud was all about compute, storage, and network resources. So, three different resources. Earlier the cloud was talking about, which is to be globally I will label and highly centralized set of resources. Now, because the cloud made the compute and storage, extremely cheap and affordable, a lot of industrial customers and enterprises started connecting their devices to the cloud.

So, the data that was not persisted or aggregated, or acquired is now streamed to the cloud, because it is extremely cheap to store the data into the cloud. Now, lots of companies and lots of industrial environment, they started to take that advantage of cloud by streaming the data coming out from the variety of devices and the sensors.

Now, as you know, that it is also use the cheaper because of this cheaper compute power to process these data streams and make the sense out of the raw data generated these sensors and devices and that was the next big shift in the cloud and that was an IoT pass. So, let me summarize. What we mean by the IoT pass is that now, since this compute storage and network resources which were earlier provided globally, they have become so cheaper that most of these organizations are called industrial customers, they start connecting their devices to the cloud.

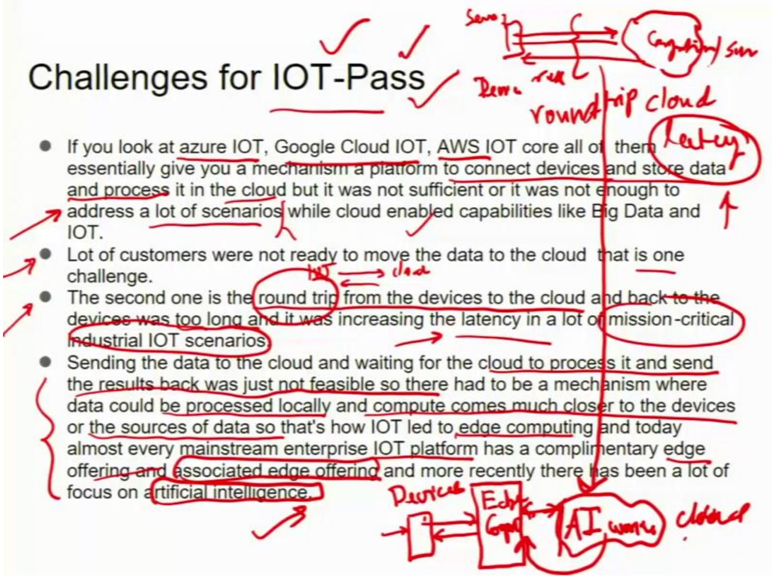
So, devices are means that the sensors, and the actuators when you say devices, you mean sensors, which will be the source of the environment or the machine data or the industrial environment data, and the actuators are kind of switches on off switch and means some actions that has to be now initiated out of these.

So, there are two types of devices and these devices are the industrial devices, they come from the industrial scenario. So, these particular devices are often connected to the cloud. And now since the cloud resources are which is nothing but a compute storage, and the network becomes so cheaper, so lots of devices, now start connecting to the cloud.

So, by means that start connecting, what they will do is that these devices will stream their data, which is coming out of these sensors, and they will be now using this particular not only for storage, these kinds of data, but they will use the enormous cheaper computing power to process these data streams and make the sense of these raw data, which is generated out of the sensors and devices out of this computing power of the cloud, and this particular trend of connecting these devices, lot of industrial devices to the cloud, this also called as IoT paths.

So, this is what is the innovation, which is having that why this so many industrial devices are now being started connecting to the cloud in wake of the situation that the compute storage and the networks are cheaper.

(Refer Slide Time: 30:34)



Challenges for IOT-Pass

- If you look at azure IOT, Google Cloud IOT, AWS IOT core all of them essentially give you a mechanism a platform to connect devices and store data and process it in the cloud but it was not sufficient or it was not enough to address a lot of scenarios while cloud enabled capabilities like Big Data and IOT.
- Lot of customers were not ready to move the data to the cloud that is one challenge.
- The second one is the round trip from the devices to the cloud and back to the devices was too long and it was increasing the latency in a lot of mission-critical industrial IOT scenarios.
- Sending the data to the cloud and waiting for the cloud to process it and send the results back was just not feasible so there had to be a mechanism where data could be processed locally and compute comes much closer to the devices or the sources of data so that's how IOT led to edge computing and today almost every mainstream enterprise IOT platform has a complimentary edge offering and associated edge offering and more recently there has been a lot of focus on artificial intelligence.

Nowadays, see that the if the so many devices are connecting to the cloud. So, what are the challenges for the cloud, and for this new path, which is called an IoT paths? So if you look at these particular cloud providers, like Microsoft, Google, and imagine, so all these cloud providers are now supporting a platform for these IoT paths, and they are called as Azure IoT, Google Cloud IoT and AWS IoT Core.

So, all of these kind of platform which supports this IoT pass, so they will give you the mechanism to connect these devices is an historical data and process it into the cloud. But it was not sufficient or it was not enough to address a lot of situations, which the industrial environment is now facing. So, although it is connecting, it is enabling these platforms to connect the devices, but they are not sufficient, because a lot of situations or scenarios are there, which are not able to exploit these IoT paths.

So, why this while this cloud-enabled capabilities are still there, like big data and IoT, but it's still there are a lot of situations which are still unaddressed, even after providing these kind of platforms, which allows to connect the devices to the cloud and store and process the data out of these IoT devices. So, let us understand this. What are the challenges is still remains part of the IoT paths.

So, now you can see that a lot of customers who are not ready to move the data to the cloud. And because of these challenges, so this has become a challenge why, although the cloud is there, and the platforms which are being provided by different vendors, like Azure IoT, or a Google IoT and AWS IoT is still why the customers are not ready to move their data to the cloud, through these kind of IoT paths or IoT platform.

So, we will understand about this challenges. And what are the way out this course will tell you about all that into cases. The second challenge for the IoT paths is that this round trip from the device to the cloud. So, for example, if the device is connected to the cloud. So, there are two ways in which the data has to move first, the data has to be carried out from the sensors or the IoT devices to the cloud, and perform the computation on public storage, and then return back the result.

So, this particular time is called round trip time, this time is called a round trip. So, round trip means that two trips, one is going from IoT to the cloud, the other is that cloud will give the results. So, it has to come back to the IoT This is called a round trip. So, this round trip from the devices to the cloud and back to the devices was too long, too long in the sense that application cannot wait this much of time for the computation from the cloud.

So, therefore, this is the challenge that many applications which these industries are the which the industrial environment needs to be computed by the cloud, but not during not this too much of time. It can wait for that. So, it was increasingly or increasing the latency. And in a lot of mission-critical applications, industrial scenarios, these kind of latency is not at all acceptable. And that has become a challenge for an IoT pass.

So, sending the data to the cloud and waiting for the cloud to process it and send back the result was not feasible. So, there had to be a mechanism where the data could be processed locally, and compute comes much closer to the devices or the sources of the data. So, that so that's how the IoT led to an edge computing. And today almost every mainstream organization or enterprise IoT platform has a complementary edge offering and associated edge offering and more recently, there has been a lot of focus on artificial intelligence.

So, let us see this particular scenario. To support the IoT paths, these are the devices and IP directly they are being sent to the cloud it means going to add a lot of latency. Now to overcome from this particular section, which says that between the device and the cloud, let us add a layer of technology which is called an edge computing. So, the devices may not have to directly do the computation from the cloud, rather the cloud will now do an offering of an edge.

So, these the compute are the services which cloud was giving which is can also be given by the edge if that is the case then the device very deep with the device very close by there is an edge layer or edge computing. So, edge computing gives the response of devices, data computation, and the latency can be reduced to a much lower level and which is being acceptable by most of these applications.

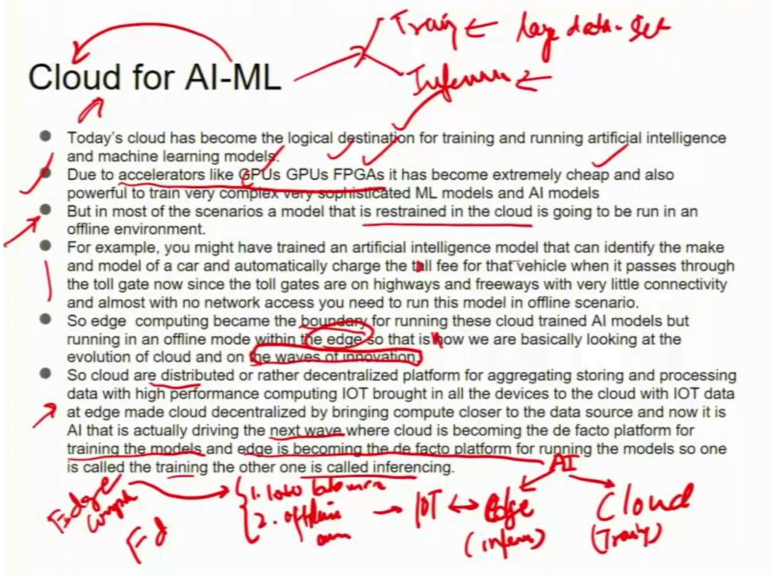
Now, for this you have to understand what for the devices are sending the data to the cloud mostly they are sending the data to the cloud to do the AI, computations AI workloads. Now, if the AI workloads are to be supported by the edge offering, then these devices can be satisfied with their response at a very lower time.

So, therefore, edge computing and associated edge computing with the help of then AI which can be now work along with the cloud and edge together to come up with the AI iron, so that devices can support or an IoT paths can be supported in this particular model. So, we will see in this particular course, how this edge computing comes and solves this particular problem.

And how this IoT pass can become successfully overcoming from these challenges, which is known here in two terms one is that the devices which are generating their data that is the industrial use cases are becoming a challenge, because of that the data has to move to the cloud and that computation has to come a long way that is a huge amount of round trip time that is being overcome with the help of a technology which is called the edge computing and edge computing and this particular Artificial intelligence has to be now being carried out at the level of the edge.

So, whatever earlier cloud was doing now edge can also do that with the same capabilities with the (())(38:43) with the cloud that is what we are going to see here in this part of the course.

(Refer Slide Time: 38:47)



Cloud for AI-ML

- Today's cloud has become the logical destination for training and running artificial intelligence and machine learning models.
- Due to accelerators like GPUs GPUs FPGAs it has become extremely cheap and also powerful to train very complex very sophisticated ML models and AI models
- But in most of the scenarios a model that is restrained in the cloud is going to be run in an offline environment.
- For example, you might have trained an artificial intelligence model that can identify the make and model of a car and automatically charge the toll fee for that vehicle when it passes through the toll gate now since the toll gates are on highways and freeways with very little connectivity and almost with no network access you need to run this model in offline scenario.
- So edge computing became the boundary for running these cloud trained AI models but running in an offline mode within the edge so that is how we are basically looking at the evolution of cloud and on the waves of innovation.
- So cloud are distributed or rather decentralized platform for aggregating storing and processing data with high performance computing IOT brought in all the devices to the cloud with IOT data at edge made cloud decentralized by bringing compute closer to the data source and now it is AI that is actually driving the next wave where cloud is becoming the de facto platform for training the models and edge is becoming the de facto platform for running the models so one is called the training the other one is called inferencing.

Now, we will see that the cloud why the cloud is used for AI that is artificial intelligence, and machine learning. So, today's Cloud has become the logical destination for artificial intelligence and machine learning. So, artificial intelligence and machine learning if you see has two different phases, one is called the training which requires a huge amount of large data set which is the poor training the artificial intelligence model and machine learning model.

The other part is called running this model and called inferencing. So, a new data when it comes the other part that the model is trained and given that is called the inferencing. So, artificial intelligence or the machine learning process is divided into two parts one is called the training the model the other is called inferencing through the model and for training and inference for training, you require accelerators that is why the data or artificial intelligence is now depending upon the cloud.

So, the artificial intelligence and machine learning to do this kind of training efficiently requires accelerators like GPUs, CPU and FPGAs, CPUs, GPUs, FPGAs, and so on. And it has become extremely cheap, if you do through the cloud, and also very powerful to train a very, very complex or very sophisticated machine learning and AI models.

And because of that, this particular cloud is a very popular means of doing AI and ML at a much cheaper way and also very powerful way to train very complex and sophisticated machine learning and AI models. But in most of the situations, this particular model is restrained in the cloud that it is going to run in an offline environment, although the training the model is done in the cloud, but where this application is going to be used, it is used in an offline environment.

So, therefore, many of these offline environment are not able to exploit this because if let us say internet connectivity is now then the cloud is no longer connected with that offline environment. So, if the environment is offline where it is going to be useful. So, this particular model directly using the cloud for AI in ML is becoming a challenge to support in all situations.

For example, you might have trained an artificial intelligence model that can identify the make and model of a car automatically and automatically charge that toll fee for that vehicle, when it passes through the toll gate. Now, since the toll gates are on highways and freeways are very little connectivity and almost with a no network access, you need to run this model in the offline situation this is an example or this is the scenario of one such use case, we are directly connecting connected with the cloud is not all the time possible, yet, how are you going to use that AI NML.

So, the edge computing comes becomes a boundary for running these particular cloud rain AI model but running in an offline mode as well within the edge. So, that now, you're basically looking at the evolution of the cloud and on the waves of innovation. So, what this course is about that how to understand these waves of this innovation, where even if your application is running in offline mode, how you are going to leverage this cloud for AI and ML and that is nothing but the edge computing.

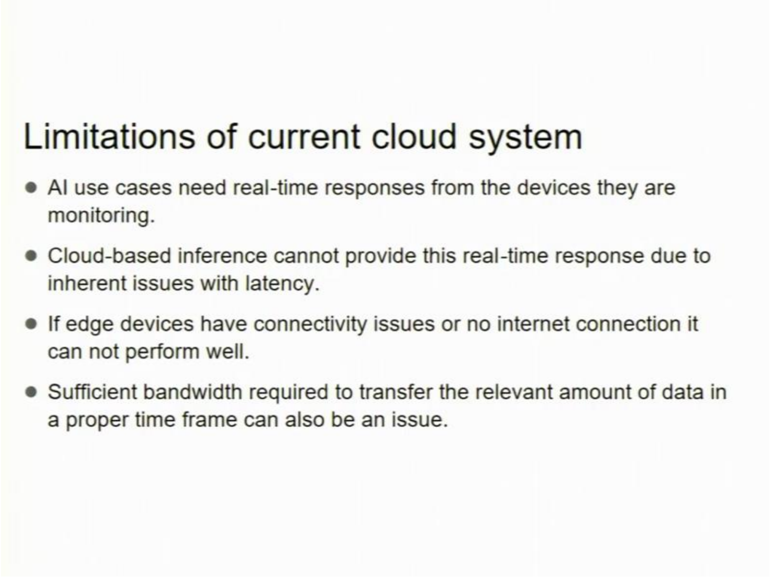
So, the clouds are now earlier it was highly centralized. So, the clouds with the help of edge computing, now they are distributed or other decentralized platform for aggregating that storing, and processing the data with a high-performance computing. So, so, it has brought to all the devices to the cloud with IoT data at the edge made the cloud decentralized by bringing the compute very closer to the data source.

And now, it is the AI that is actually driving the next wave we are the cloud is becoming the de facto standard for training the models and the edge is becoming the de facto standard for running the AI models. So, one is called as the training the other is called as the inferencing. So, if you see the cloud edge together use for AI. So, the training is done at the cloud and the inferencing part of the AI can be done at the edge, and here are the IoT devices.

And now these IoT devices can get can so can be served in this particular model quite efficiently with a low latency and also in an offline mode. So, the advantage of this particular model is it will serve with a low latency and even it can serve in an offline environment. And this is possible with the help of a technique which is called the edge computing.

So, edge computing will bring both the advantages that is it will bring the low latency that is it will reduce that round-trip time to an affordable latency which is prescribed in the applications. Second is that this particular challenge that still to solve the applications being in the offline environment that also become doable and possible with this new approach, which is the edge computing and the current AI and ML.

(Refer Slide Time: 45:13)



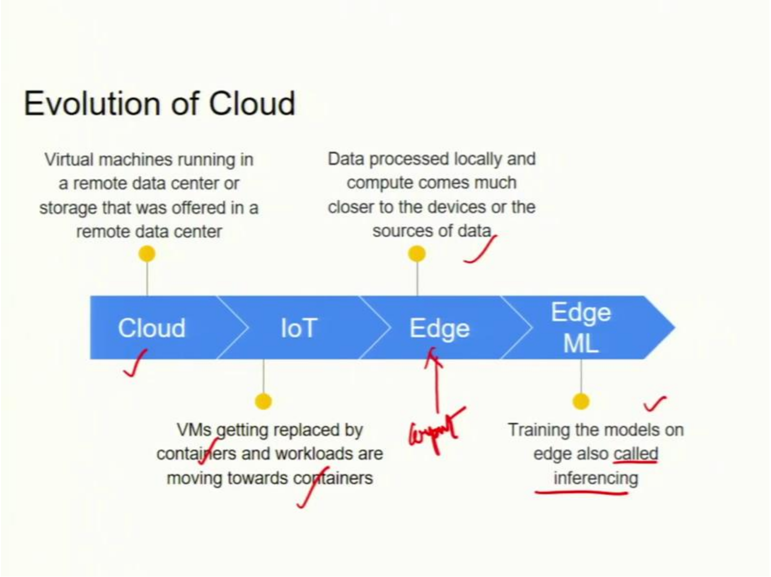
Limitations of current cloud system

- AI use cases need real-time responses from the devices they are monitoring.
- Cloud-based inference cannot provide this real-time response due to inherent issues with latency.
- If edge devices have connectivity issues or no internet connection it can not perform well.
- Sufficient bandwidth required to transfer the relevant amount of data in a proper time frame can also be an issue.

So, therefore, the current cloud, which is now going through various levels of innovations, let us summarize the limitations and how the new cloud is now gearing up for all that system. So, AI use cases need the real-time responses from the devices which they are monitoring. Similarly, the cloud-based inference cannot provide this real-time response due to the inherent issues with the latency that we have explained with the help of toll example.

That if the toll is to be solved with the help of AI and ML in an offline mode, the cloud cannot do that. So, if the edge devices have this particular connectivity issues and no internet connection, it cannot perform well. Therefore, this is the limitations of the earlier classical cloud AI in the sufficient bandwidth also is required to transfer the relevant data in the proper timeframe and also be an issue.

(Refer Slide Time: 46:14)



Evolution of Cloud

Virtual machines running in a remote data center or storage that was offered in a remote data center

Data processed locally and compute comes much closer to the devices or the sources of data

VMs getting replaced by containers and workloads are moving towards containers

Training the models on edge also called inferencing

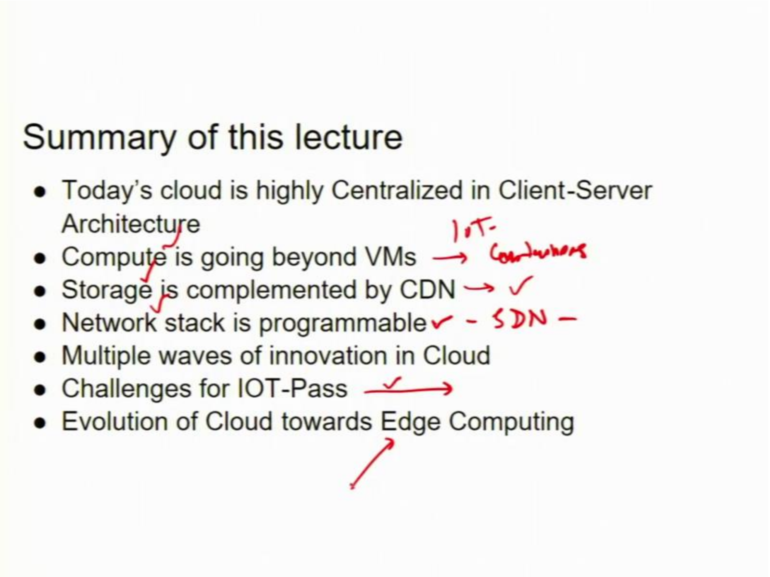
So, if you see these classical cloud then the waves of innovation and the evolution of today's cloud, what you can see is that let us summarize that the classical cloud was being sold with the help of virtual machines running in a remote data center or the storage that was offered through a remote data center.

Then with the IoT pass, these virtual machines are now replaced with the help of containers, because of that particular region that they are not be efficient to support so many number of devices which are start connecting to the cloud. So, virtual machines are now getting replaced with the help of containers, and the workloads of these devices are now moving towards these containers.

Then, we have also seen that this to support these IoT paths with the AI and ML models of computation of these data which these devices IoT devices generates these data need to be processed locally not have to be sent to the cloud due to the round-trip delay. So, therefore, data is to be processed locally, and that compute comes much closer to the devices or the sources and that technology is called an edge on edge computing.

That we will discuss later part of this particular course, now comes the AI and ML. So, that particular AI and ML, which you can now run at the edge, which is very close to the devices, that is called ml or edge machine learning. So, there what you can do, you can still train the model on the cloud. And now with the edge having lot of computing power sometimes the training also you can do at the edge and but as far as the other part that is influencing you can do at the edge and this particular capability is called edge ML.

(Refer Slide Time: 48:18)



Summary of this lecture

- Today's cloud is highly Centralized in Client-Server Architecture
- Compute is going beyond VMs → IoT- Containers
- Storage is complemented by CDN → ✓
- Network stack is programmable ✓ - SDN -
- Multiple waves of innovation in Cloud
- Challenges for IOT-Pass ✓
- Evolution of Cloud towards Edge Computing

So, let us summarize this particular lecture. So, here we have understood about today's cloud, which is classical cloud, which is highly centralized and a set of resources in terms of compute, storage, and the network and often is categorized as the client-server architecture of classical cloud. Now, then, we have seen that the compute which is being provided with the help of the technique, which is called the virtual machine in the classical cloud is going beyond that, what is that beyond is called the containers we have seen that beyond what you will machine is the containers.

So, nowadays to support this IoT pass a new technique other than this virtual machines the containers are used. Now, another besides compute there is another important resource of a cloud is called storage. Now, storage is also complemented by a content delivery network. So, what do we mean by content delivery network that is storage, if it is required to be accessed at more than one places.

So, therefore, this storage is not only stored at a centralized highly centralized place, but it's to be replicated at various Edge locations, and that is being supported by a technique called Content Delivery Network. So, storage is now complemented by content delivery network and this is also a next wave of innovation that is being supported inside the cloud services.

Now, third important thing is the resource which is called the network, we have also seen that this particular virtual machines is being shared across multiple vendors or the users maybe also with the help of containers, so, how this particular network also can be shared and this sharing can coexist across multiple applications.

So, therefore, networking stack has to be made programmable. So, we have seen the technique called SDN software-defined network is the technology which provides this network sharing across multiple applications. And therefore, the network stack has to become programmable that we have already seen here in that as a development of today's cloud. Finally, we have seen that the multiple waves of innovation are happening in the cloud.

Why to support a new kind of use case where the devices from different industrial enterprises, many devices, industrial enterprises start connecting to the cloud and various public cloud providers they provide platform such as, as your IoT. Similarly, AWS IoT and Google Cloud IoT, so these platforms are available but still It has to go through a lot of challenges called IoT paths.

So, we have seen that these challenges to overcome from these challenges, there is an evolution or the wave of innovation of a cloud and this is called edge computing. So, edge computing has to be brought into to go forward this IoT path, and the current cloud how which is highly centralized becomes a decentralized here in this manner.

So, in the next class, we are going to discuss more about this new way or new innovation into the cloud that is called edge computing. So, thank you very much for listening to this particular lecture. Thank you all. Thank you very much.