# Foundation of Cloud IoT Edge ML

**Professor Rajiv Misra**

**Department of Computer Science and Engineering**

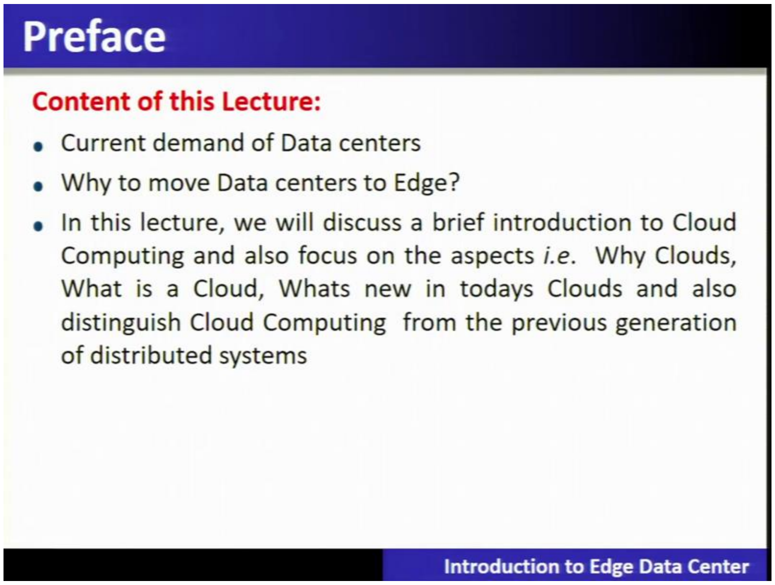
**Indian Institute of Technology, Patna**

**Lecture 18**

## Introduction to Edge Data Center for IoT Platform

Introduction to Edge Data Center for IoT Platform. I am Doctor Rajiv Misra from IIT Patna.

(Refer Slide Time: 00:21)



Preface

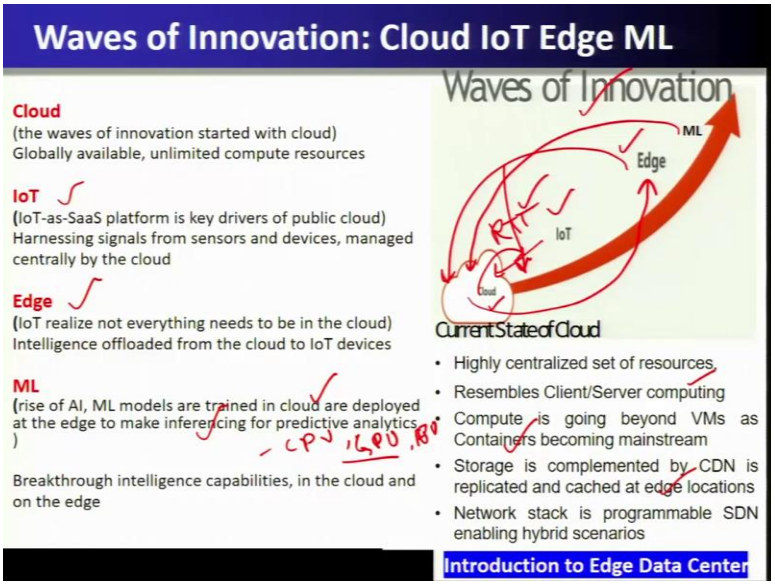
Content of this Lecture:

- Current demand of Data centers
- Why to move Data centers to Edge?
- In this lecture, we will discuss a brief introduction to Cloud Computing and also focus on the aspects *i.e.* Why Clouds, What is a Cloud, Whats new in todays Clouds and also distinguish Cloud Computing from the previous generation of distributed systems

Introduction to Edge Data Center

So, the content of this lecture is as follows. So, we will introduce you to the current demand of data centers and then we will also talk about the data centers moving to the edge. So, paving away to a new kind of data center called as edge data center. In this lecture we will also discuss the brief introduction to the cloud computing and its focus on various aspects such as why Cloud, what is a Cloud, what is new in today's Cloud and also what distinguish the Cloud Computing from previous generations and now introduce about the edge computing uses.

(Refer Slide Time: 01:00)



## Waves of Innovation: Cloud IoT Edge ML

**Cloud**
(the waves of innovation started with cloud)
Globally available, unlimited compute resources

**IoT** ✓
(IoT-as-SaaS platform is key drivers of public cloud)
Harnessing signals from sensors and devices, managed centrally by the cloud

**Edge** ✓
(IoT realize not everything needs to be in the cloud)
Intelligence offloaded from the cloud to IoT devices

**ML**
(rise of AI, ML models are trained in cloud are deployed at the edge to make inferencing for predictive analytics
)

- CPU, GPU, TPU

Breakthrough intelligence capabilities, in the cloud and on the edge

Waves of Innovation

Current State of Cloud

- Highly centralized set of resources
- Resembles Client/Server computing
- Compute is going beyond VMs as Containers becoming mainstream
- Storage is complemented by CDN is replicated and cached at edge locations
- Network stack is programmable SDN enabling hybrid scenarios

Introduction to Edge Data Center

So, let us start with the motivation of this particular lecture of edge data centers. Now, what we have seen is the evolution of the cloud and also the waves of the innovation and that we will see in the cloud IoT edge and ML scenarios. So, what you see is that this cloud has become a highly available centralized set of resources with an unlimited compute resources and the storage and network resources.

Now, this particular development where the cloud is considered to be highly centralized but globally available resources often required new innovations in the cloud due to the fact that new kind of applications such as Internet of Things is also supported by the public cloud. Lot of IoT devices start connecting to the cloud for sending their data called, streaming their data for doing the processing storage and other aggregations. So, these are all managed by centralized cloud.

So, if you see the current state of the cloud that is nothing but highly centralized set of resources. Often, this cloud is resembling the client server computing or model and the compute earlier which was considered in the cloud in the form of virtual machine is going beyond that and now, it is introducing or innovation is in the form of containers and which becomes the mainstream in the cloud.

Storage in the cloud is also being changed. That is complemented by content distribution network which are replicated and cached at various edge locations for efficient accessing this particular web resources through the web. So, network stack also is programmable and software defined network is the technology which enabling the hybrid kind of scenarios.

So, with this current state of cloud and allowing this lot of IoT devices to connect to the public cloud for computing and processing and also these, harnessing the signals from sensors and IoT data which is fully managed by the cloud. But this IoT, after so much of use, it has realized that not everything need to be done at the cloud, thereby realizing a way of, for the edge computing.

So, edge computing is a wave of innovation which is evolving around the cloud which is also around development around the IoT use case. So, IoT requires the data which is to be streamed to the cloud earlier for doing the computation analysis and taking up the insights. So, what is being found by IoT applications is that not every data is required to be sent to the cloud because of several reasons that it is going to add the round trip delay when the data is sent to the cloud for making the computations.

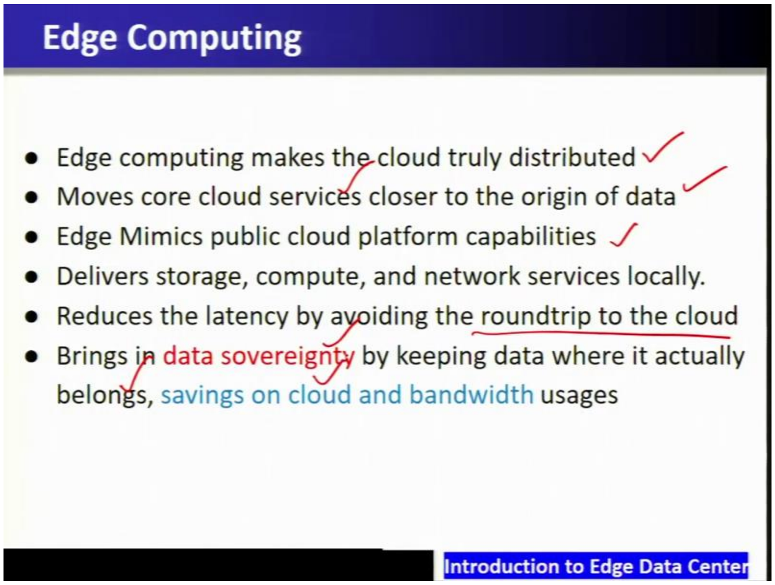
So, therefore many applications which is a real time applications which cannot be supported in an IoT cloud model, thereby it requires that IoT data need to be processed very close to the device and that introduces the edge computing. With the advent or with the rise of artificial intelligence what is being found is that the machine learning models often trained in the cloud due to the several regions such as the availability of CPU, GPU and FPU, there are many other processing units, which are available at the cloud at a lower cheaper cost.

So, this becomes an ideal destination to perform the machine learning model training in the cloud. Now, once the models are trained then for inferencing part it can be now deployed to the edge for inferencing, thereby able to process or doing the inferencing very close to the device by the edge. So, this is the development which is also we are termed as the waves of innovation.

So, the development of the cloud has paved into the application or the Industries or manufacturing unit to use the cloud for streaming the data I am doing the analytics. Now, the data has to be sent to the cloud, thereby round-trip time is being added and lot of bandwidth is also used in this particular process. IoT application, they realized the need of edge between the cloud and the IoT edge. So, when you say edge that is the network edge.

So, that is out of the cloud if all these compute resources or computing capabilities which cloud used to do if it is available at the edge that is very close to the IoT devices that will now pave a way of reducing this latency, low latency applications can be supported, and also performing the machine learning on the cloud was the standard because of the availability of CPU GPU and other processors to do the machine learning training. Once the model is trained, then it can be deployed in the edge for the computing, so for the inferencing. So, thereby, lot of innovations, you can see, are paving the way of a new technology which is called an edge computing, which is the evolution of the cloud.

(Refer Slide Time: 07:00)



## Edge Computing

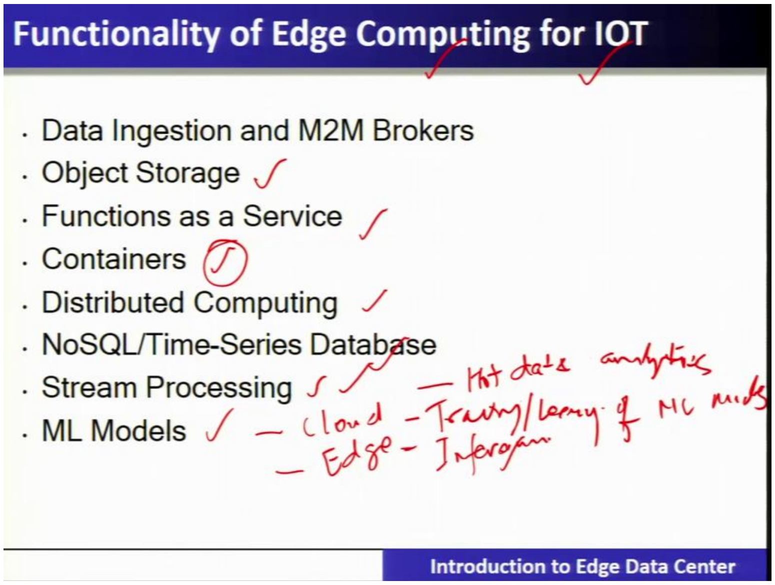
- Edge computing makes the cloud truly distributed ✓
- Moves core cloud services closer to the origin of data ✓
- Edge Mimics public cloud platform capabilities ✓
- Delivers storage, compute, and network services locally.
- Reduces the latency by avoiding the roundtrip to the cloud
- Brings in data sovereignty by keeping data where it actually belongs, savings on cloud and bandwidth usages

Introduction to Edge Data Center

So, let us go ahead and see the edge computing. So, edge computing makes the cloud a truly distributed and then most of the core will be now of the cloud services now be available at the origin of the data. Now, this edge mimics the public cloud platform capabilities and thereby it will deliver the storage compute and network services locally.

It also reduces the latency by avoiding the round trip to the cloud for performing the computation and inferencing, this brings into data sovereignty by keeping the data where it belongs without any breach of privacy. Similarly, it will also save on the cloud and bandwidth usage if the data is not streamed to the cloud but it is processed very close to the cloud, that is at the edge, so giving away to the as a computing.

(Refer Slide Time: 08:00)



## Functionality of Edge Computing for IOT

- Data Ingestion and M2M Brokers
- Object Storage ✓
- Functions as a Service ✓
- Containers ✓
- Distributed Computing ✓
- NoSQL/Time-Series Database
- Stream Processing ✓ ✓ — Hot data analytics
- ML Models ✓ — Cloud — Training/Learning of ML models
  — Edge — Inferencing

Introduction to Edge Data Center

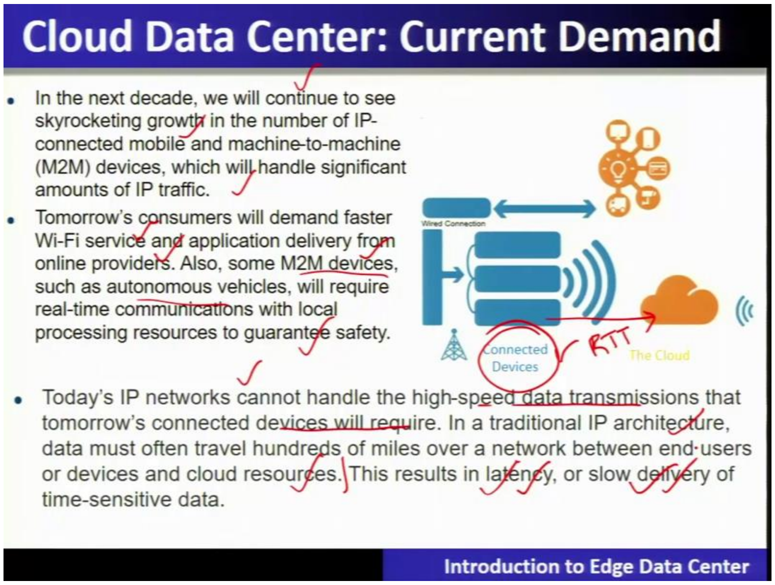
Now, let us see what are the functionalities required for edge computing to be enabled. So, the data ingestion and machine-to-machine brokers for data ingestion from the IoT devices which was earlier done by the cloud, now these functionalities are available at the edge, thereby paving away this functionality or mimic this functionality from the cloud to the edge.

Similarly, the data storage which was earlier used to send to the cloud, these IoT sensors now, for storage, now, that is also possible to be done with the help of edge computing, whatever is the data which is available. So, function as a service is also, that is the compute is also available at the edge, in the form of containers. Machine learning model inferencing can be deployed, distributed computing model can be created in this using the edge computing.

And also, various databases like NoSQL and time-series databases can also be supported with edge computing. Stream processing is nothing but it is also called a hot data analytics. So, this kind of functionality which was earlier there with the cloud is now possible to be brought into the edge and performing the applications of this kind of analytics at the edge, very close to the source.

Machine learning models uses normally this kind of specialized processor CPU, GPU, FPU and so on, and thereby the cloud was the de facto destination for performing the training, or it is called the learning part of the machine learning models. Once the models are trained, that is machine learning models are trained on the cloud, then they can be deployed at the edge with the help of containers for making the inferencing. So, the edge computing, once the IoT data, once the IoT device sends the data to the edge using this model, that is machine learning trained model, it will be able to do the inferencing and thereby, it will be now used for business intelligence and many other functionalities.

(Refer Slide Time: 10:37)



## Cloud Data Center: Current Demand

- In the next decade, we will continue to see skyrocketing growth in the number of IP-connected mobile and machine-to-machine (M2M) devices, which will handle significant amounts of IP traffic.
- Tomorrow's consumers will demand faster Wi-Fi service and application delivery from online providers. Also, some M2M devices, such as autonomous vehicles, will require real-time communications with local processing resources to guarantee safety.

- Today's IP networks cannot handle the high-speed data transmissions that tomorrow's connected devices will require. In a traditional IP architecture, data must often travel hundreds of miles over a network between end-users or devices and cloud resources. This results in latency, or slow delivery of time-sensitive data.

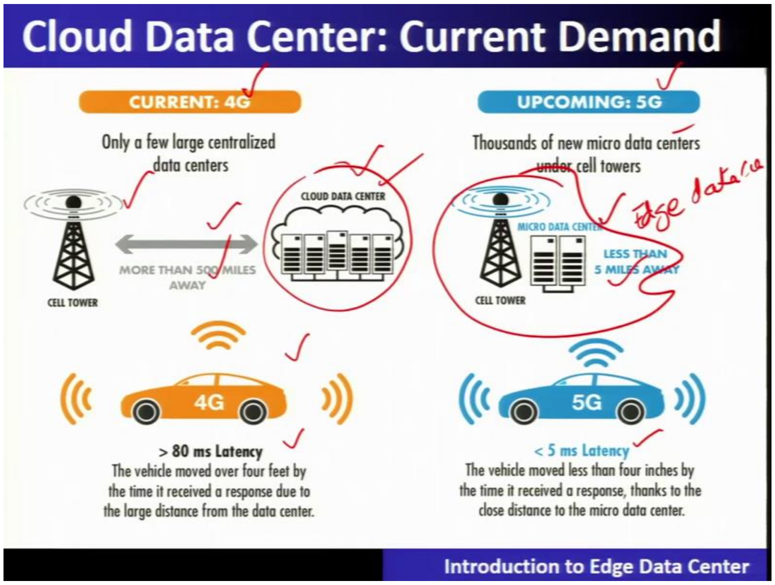
Introduction to Edge Data Center

So, let us go ahead and see the development in this particular direction, that is how the cloud is being provided services with the help of cloud data center and how the edge will provide this functionality with another type of data center that is called edge data center. Let us see all these development in this particular lecture. So, in the next decade we will continue to see the growth in the number of IP connected mobiles and machine-to-machine devices which will handle significant amount of this IP traffic.

Now, tomorrow's consumers will demand the faster services and the application delivery for online providers. So, there are some machine-to-machine devices such as autonomous vehicles, they require the real time communication with the local processing resources to guarantee the safety applications. Here, that is all shown over here, that is these connected devices often are connecting to the cloud for doing various processing.

For that it requires these machine-to-machine devices, they are enabled. So, this IP network cannot handle the high-speed data transmission that tomorrow's connected devices is needed. Thereby, a traditional IP architecture where data must travel often the hundreds of miles over the network end users or the devices is no longer be the appropriate model because this results into the latency and slow delivery of time sensitive data. So, this particular model is deviating that IoT is sending, these connected devices are sending their data to the cloud for streaming, and it, because of round trip time it is giving the latency and slow delivery of time sensitive data.

(Refer Slide Time: 12:43)



Cloud Data Center: Current Demand

> 80 ms Latency
The vehicle moved over four feet by the time it received a response due to the large distance from the data center.

< 5 ms Latency
The vehicle moved less than four inches by the time it received a response, thanks to the close distance to the micro data center.

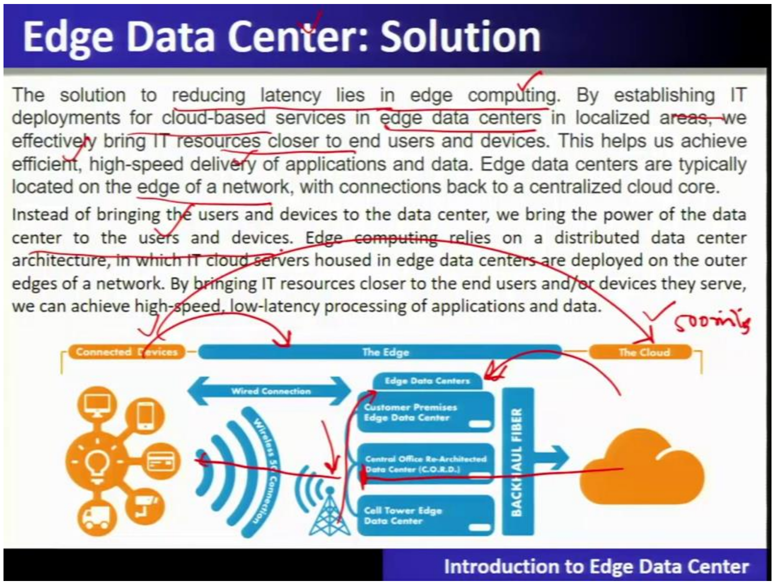
Introduction to Edge Data Center

So, thereby let us see the cloud data centers which are developed around the current demands of the 4G. So, 4G is the technology, communication cellular technology which uses often, the cloud data centers and the cloud data centers are often developed in support of these kind of 4G technologies.

But these applications, this kind of model where all the processing is done at the centralized cloud data center is not able to support the latency sensitive applications, for example self-driving car, where the latency is more than 800 milliseconds because it is 400 miles away, that cloud data center where the computing or processing has to be happened. Now, thereby a new kind of cellular networks are coming up they are called 5G. So, 5G requires these, a new kind of data centers which are called micro data centers, sometimes they are called edge data centers. So, these edge data centers are very close to the source, you can see.

Thereby, the data need not have to travel 500 miles, but it is only has to travel 5 miles, and thereby this latency sensitive applications are able to deliver the compute results very efficiently, very fast, and the latency is reduced up to less than 1 millisecond in this particular scenario, where the data centers are brought very close to the source. They are called edge data centers or a micro data centers.

(Refer Slide Time: 14:30)



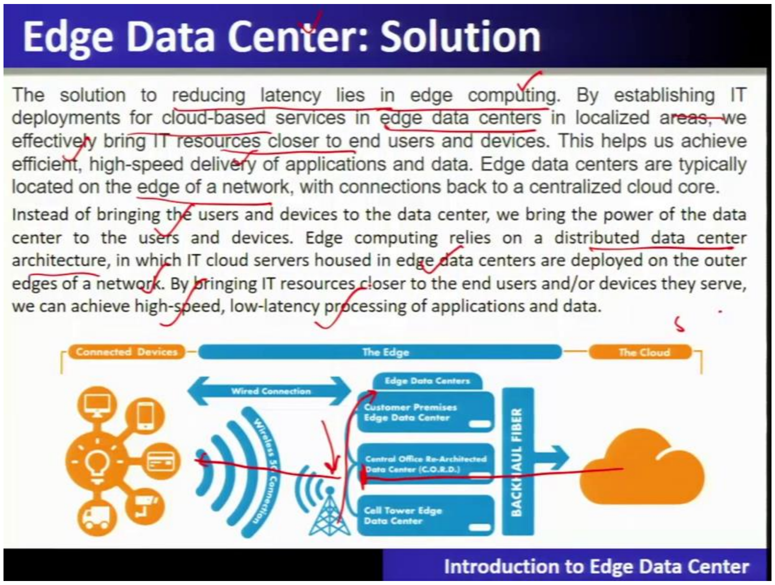
## Edge Data Center: Solution

The solution to reducing latency lies in edge computing. By establishing IT deployments for cloud-based services in edge data centers in localized ~~areas, we~~ effectively bring IT resources closer to end users and devices. This helps us achieve efficient, high-speed delivery of applications and data. Edge data centers are typically located on the edge of a network, with connections back to a centralized cloud core.

Instead of bringing the users and devices to the data center, we bring the power of the data center to the users and devices. Edge computing relies on a distributed data center architecture, in which IT cloud servers housed in edge data centers are deployed on the outer edges of a network. By bringing IT resources closer to the end users and/or devices they serve, we can achieve high-speed, low-latency processing of applications and data.

500 mils
Connected Devices
The Edge
The Cloud
Wired Connection
Wireless 5G Connection
Edge Data Centers
Customer Premises Edge Data Center
Central Office Re-Architected Data Center (C.O.R.D.)
Cell Tower Edge Data Center
BACKHAUL FIBER

Introduction to Edge Data Center



## Edge Data Center: Solution

The solution to reducing latency lies in edge computing. By establishing IT deployments for cloud-based services in edge data centers in localized ~~areas, we~~ effectively bring IT resources closer to end users and devices. This helps us achieve efficient, high-speed delivery of applications and data. Edge data centers are typically located on the edge of a network, with connections back to a centralized cloud core.

Instead of bringing the users and devices to the data center, we bring the power of the data center to the users and devices. Edge computing relies on a distributed data center architecture, in which IT cloud servers housed in edge data centers are deployed on the outer edges of a network. By bringing IT resources closer to the end users and/or devices they serve, we can achieve high-speed, low-latency processing of applications and data.

Connected Devices
The Edge
The Cloud
Wired Connection
Wireless 5G Connection
Edge Data Centers
Customer Premises Edge Data Center
Central Office Re-Architected Data Center (C.O.R.D.)
Cell Tower Edge Data Center
BACKHAUL FIBER

Introduction to Edge Data Center

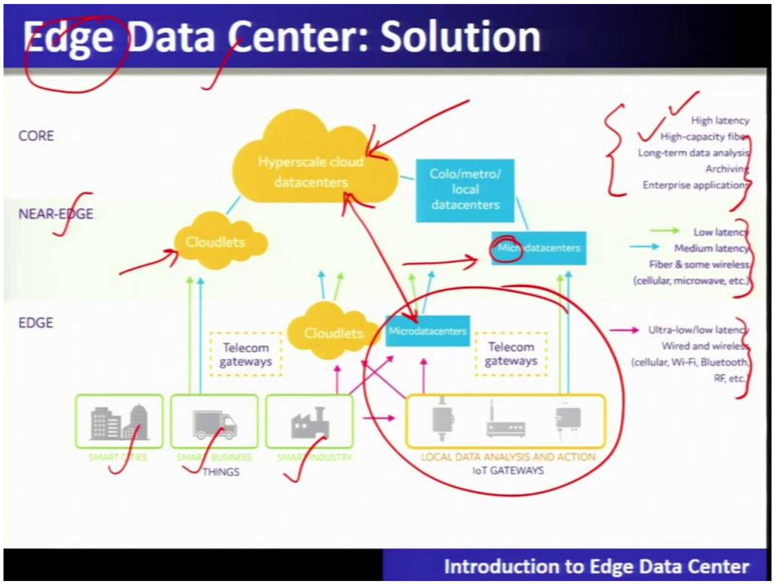
So, let us see the edge data center, what are the solutions. So, the solution to reducing the latency lies in the edge computing. And thereby, you can see in this particular example that these connected devices, earlier has to send the data to the cloud, which is far away, that is 500 miles away, but now these connected devices will send the data very close to the device. They are called the edge.

So, with the help of edge data center this is possible to do the computation very close to the source, that is, the devices and this can be done by establishing this IT deployments for cloud-based services in the edge data centers. So, whatever cloud capabilities were there earlier, now is being enabled at the edge, in the edge data centers as well. So, this particular, this way these IT resources in the form of compute storage and the network is becoming more closer to the end user in this technology that is called the edge computing. And this is enabled with the help of edge data centers. So, this helps us achieve efficient, high speed delivery of the application in the data.

These edge data centers are typically located on the edge of the network. When you say edge of the network that is out of this cloud and before it connects to the device, that is called the edge. So, this particular example shows that this is the 5G tower. Very close to it, there will be an edge data center serving. So, these devices has to only serve, being served by the edge data center which is very close to this particular device.

So, you can see that instead of bringing the user and the devices to the data center, that is the cloud data center, we bring the power of data center to the users and the devices through the edge computing. So, the edge computing relies on a distributed data center architecture in which the cloud services housed in an edge data centers and are deployed on the outer edge of the network. So, by bringing these IT resources closer to the end users or the devices they serve what you can achieve is high speed low latency processing of the data for different applications.

(Refer Slide Time: 17:05)



Edge Data Center: Solution

Introduction to Edge Data Center

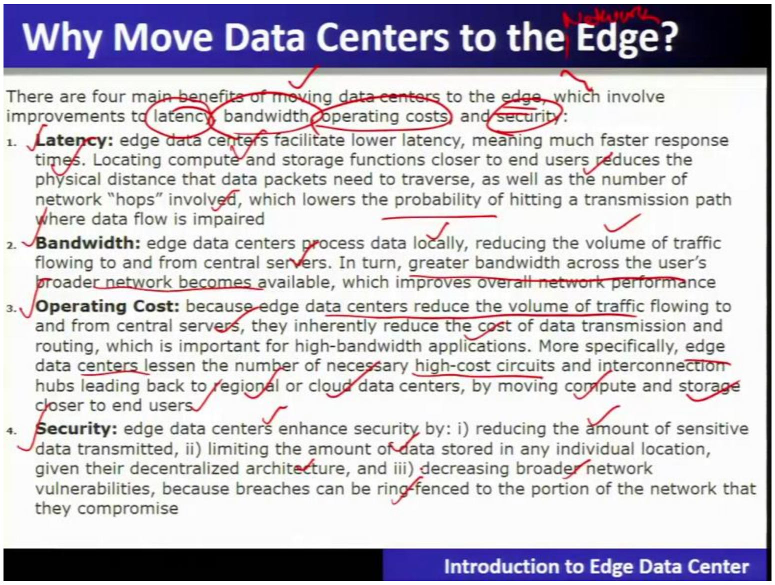
So, let us see the edge data center details. So, what you will find is that these data center, that is a cloud data centers are also called as a hyper scale cloud data centers which is a highly centralized set of resources. And they are called as cloud data centers. But these cloud data centers provides enormous amount of compute storage and network capabilities.

But suffers, if it is used for computation, it suffers from high latency high capacity fiber long term data analysis archival. For this, it is good, but high latency or latency sensitive applications are not good to be computed in this hyperscale cloud data centers. Thereby, the edge data centers, you can see, is out of the cloud and they are very close to these particular devices.

So, they are called the edge of the network. So, the data centers away from the cloud if they are available at the edge of these kind of network then they are called micro data centers or it is called the edge data centers, we name it. So, micro data centers are small sized data centers but having the same capabilities as the cloud data centers, and thereby, they are able to serve the devices at the local, very close by.

So, this way of serving the devices with the help of edge data centers, what you will find is that the latency is reduced or maybe latency is reduced to a ultra low latency sensitive applications in this manner. So, new kind of applications, such as smart cities applications and smart business or smart industries, they are possible in this particular scenario that is with the help of edge data centers.

(Refer Slide Time: 19:10)



## Why Move Data Centers to the Edge?

There are four main benefits of moving data centers to the edge, which involve improvements to latency, bandwidth, operating costs and security:

1. **Latency:** edge data centers facilitate lower latency, meaning much faster response times. Locating compute and storage functions closer to end users reduces the physical distance that data packets need to traverse, as well as the number of network "hops" involved, which lowers the probability of hitting a transmission path where data flow is impaired
2. **Bandwidth:** edge data centers process data locally, reducing the volume of traffic flowing to and from central servers. In turn, greater bandwidth across the user's broader network becomes available, which improves overall network performance
3. **Operating Cost:** because edge data centers reduce the volume of traffic flowing to and from central servers, they inherently reduce the cost of data transmission and routing, which is important for high-bandwidth applications. More specifically, edge data centers lessen the number of necessary high-cost circuits and interconnection hubs leading back to regional or cloud data centers, by moving compute and storage closer to end users
4. **Security:** edge data centers enhance security by: i) reducing the amount of sensitive data transmitted, ii) limiting the amount of data stored in any individual location, given their decentralized architecture, and iii) decreasing broader network vulnerabilities, because breaches can be ring-fenced to the portion of the network that they compromise

Introduction to Edge Data Center

Now, the question is why move this cloud data centers at the edge or at the network edge? So, there are four benefits of moving this centralized cloud data centers to the edge which involve the improvement in terms of latency, latency will be reduced, bandwidth utilization, bandwidth will be saved and operating cost also will be reduced because not every time this cloud is going to be used, and also security will be improved. Why? Because the computation is to be done locally without taking away the data from the local premise.

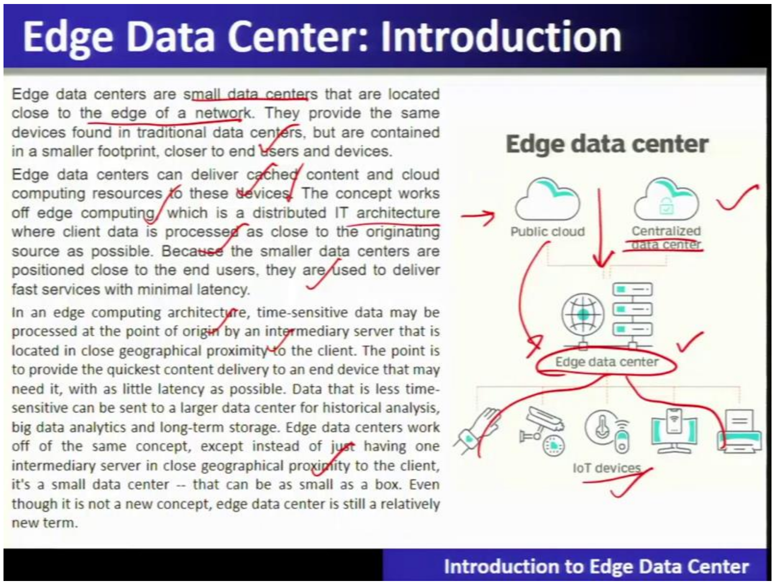
So, thereby, let us see one by one what are the benefits if you move the services from the cloud data center to the edge data center. First important development or improvement or innovation is in the terms of latency. So, edge data centers will facilitate low latency meaning that fast response time is there. And for a time, sensitive databases or data applications, this is a very much need. So, locating the compute and the storage functions very close to the end users will reduce the physical distance that the data packet need to traverse and the number of hops involved will be lower, and thereby the data flow is impaired.

So, second importance or second advantage is in terms of bandwidth saving. So, edge data center will process the data locally, thereby reducing the traffic which is exposed on the internet. So, reducing the volume of the traffic flowing to and from the central or a cloud data centers, thereby the greater bandwidth across the users broader network becomes available and it will also improve the network performance.

Third important thing is the operating cost because the edge data center reduce the volume of the traffic flowing from the cloud data centers. They inherently reduce the cost of data transmission routing involved in performing the computation from this cloud. So, this particular way, that is called edge data centers is required to see as the high cost saving and also becoming the hub and, which are often giving the services like a cloud, thereby moving the compute and storage very close to the user, and the operating cost also will be reduced in this particular way of running the applications for enabling the computation at the edge.

Security. So, edge data centers will enhance the security thereby reducing the amount of sensitive data which is to be transmitted, limiting the amount of data which is stored in any location given their decentralized architecture. Now, third important advantage of the security is that it will decrease the border network vulnerabilities. Why? Because the data will be stored locally, thereby this breach can be ring-fenced to portion of the network that they compromise.

(Refer Slide Time: 22:23)



Edge Data Center: Introduction

Edge data centers are small data centers that are located close to the edge of a network. They provide the same devices found in traditional data centers, but are contained in a smaller footprint, closer to end users and devices.

Edge data centers can deliver cached content and cloud computing resources to these devices. The concept works off edge computing, which is a distributed IT architecture where client data is processed as close to the originating source as possible. Because the smaller data centers are positioned close to the end users, they are used to deliver fast services with minimal latency.

In an edge computing architecture, time-sensitive data may be processed at the point of origin by an intermediary server that is located in close geographical proximity to the client. The point is to provide the quickest content delivery to an end device that may need it, with as little latency as possible. Data that is less time-sensitive can be sent to a larger data center for historical analysis, big data analytics and long-term storage. Edge data centers work off of the same concept, except instead of just having one intermediary server in close geographical proximity to the client, it's a small data center -- that can be as small as a box. Even though it is not a new concept, edge data center is still a relatively new term.

Introduction to Edge Data Center

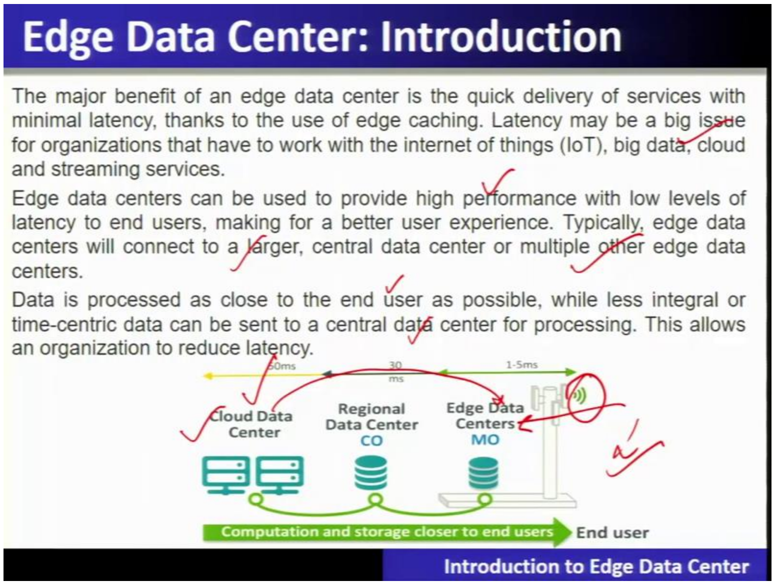
So, let us introduce the edge data center. So, edge data centers are the small data centers. They are located very close to the network edge. So, in this example, in this figure you can see on one end there is a public cloud which is nothing but provided with the help of centralized data center. Now, if you move away from this cloud you will find the network edge, and at the network edge you place a data center, they are called edge data center.

These edge data center will serve these devices locally. So, you can see that local devices in a particular premise will require a smaller compute capability, not that equivalent to the cloud. But those cloud services if they are available at the edge or at the network edge, then these devices can be served at the edge data center itself.

So, they provide the devices and the applications, which is or otherwise was processed by the centralized cloud data centers. So, edge data centers can deliver also the cashed contents and the cloud computing resources to these particular devices will be provided. So, the concepts of edge computing which is nothing but a distributed IT architecture where the client data is processed as close to the originating source as possible because the smaller data centers are positioned close to the end users, and they used to deliver the services very fast with a minimal latency.

So, in the edge computing architecture, time sensitive data may be processed at the point of origin by intermediary servers, that is located close to the geographical proximity and thereby providing the content delivery at the end devices with a very little latency which is possible. So, data that is less time sensitive can be sent to the cloud data center for historical analysis such as big data analytics for the long-term storage, whereas the edge data centers work offline means off in the same concept except of just having one server very close to the geographical proximity. And that is the edge data center used to be a small data center.

(Refer Slide Time: 24:58)



## Edge Data Center: Introduction

The major benefit of an edge data center is the quick delivery of services with minimal latency, thanks to the use of edge caching. Latency may be a big issue for organizations that have to work with the internet of things (IoT), big data, cloud and streaming services.

Edge data centers can be used to provide high performance with low levels of latency to end users, making for a better user experience. Typically, edge data centers will connect to a larger, central data center or multiple other edge data centers.

Data is processed as close to the end user as possible, while less integral or time-centric data can be sent to a central data center for processing. This allows an organization to reduce latency.

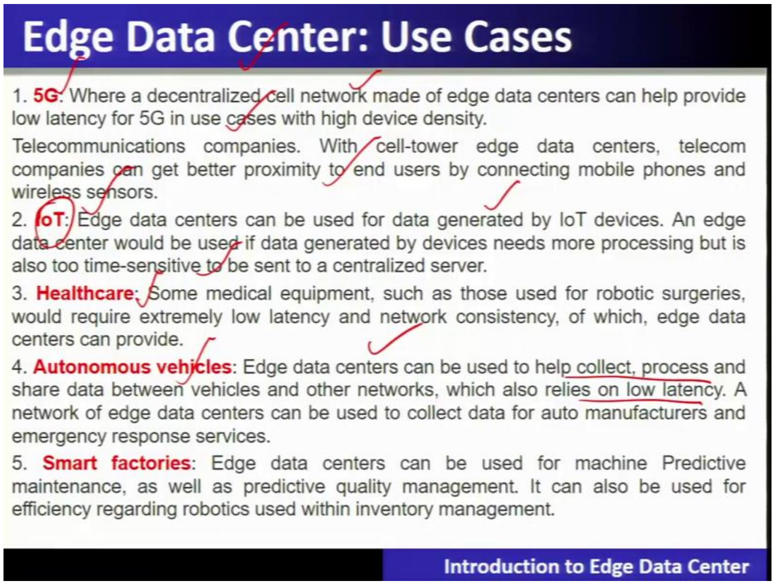
Introduction to Edge Data Center

So, this is a new concept that is called edge data center is a new innovation wave around the evolution of the cloud computing. So, edge data center, let us introduce. Some major benefit is the quick delivery that we have already seen and thereby the latency may be a big issue for various organization who are using the internet of things and for various compute purposes, such as streaming services and analytics or analysis or inferencing services of a machine learning or AI model.

So, edge data centers can be used to provide high performance low level of latency to these end users making a better experience. So, these edge data center will connect to a larger, that is the centralized or cloud data centers or maybe a multiple of other edge data centers. So, this will form a complete distributed environment. So, the cloud will turn into a distributed platform with the help of this edge data center.

So, in this particular architecture, data is processed as close to the end user as possible while it is less integral to the time sensitive data that is sent to the central data center processing. So, you can see here in this particular illustration, there is a cloud data center but as far as the devices are concerned, devices can directly be reached to the edge data centers and those kind of capabilities of a cloud are mimicked in the edge data center. These edge data center will be connected close by with the other regional data centers and finally these regional data centers are connected in the form of, with the cloud data centers. So, you can see there is a distributed way these data centers are organized and thereby able to do these computations.

(Refer Slide Time: 26:52)



## Edge Data Center: Use Cases

1. **5G**: Where a decentralized cell network made of edge data centers can help provide low latency for 5G in use cases with high device density.

Telecommunications companies. With cell-tower edge data centers, telecom companies can get better proximity to end users by connecting mobile phones and wireless sensors.

2. **IoT**: Edge data centers can be used for data generated by IoT devices. An edge data center would be used if data generated by devices needs more processing but is also too time-sensitive to be sent to a centralized server.

3. **Healthcare**: Some medical equipment, such as those used for robotic surgeries, would require extremely low latency and network consistency, of which, edge data centers can provide.

4. **Autonomous vehicles**: Edge data centers can be used to help collect, process and share data between vehicles and other networks, which also relies on low latency. A network of edge data centers can be used to collect data for auto manufacturers and emergency response services.

5. **Smart factories**: Edge data centers can be used for machine Predictive maintenance, as well as predictive quality management. It can also be used for efficiency regarding robotics used within inventory management.

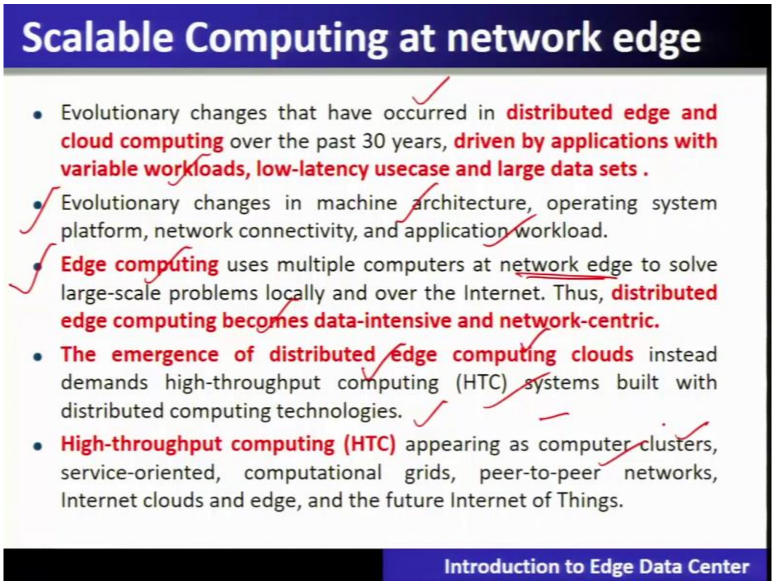
Introduction to Edge Data Center

The use cases of edge data centers is, first of all, in the 5G. So, 5G is having a decentralized cellular network made of the edge data centers, and provide low latency use cases in 5G. So, these are the telecom companies which requires this edge data center usage. Second is called IoT application. So, edge data centers can be useful for computing these IoT devices.

So, IoT devices are now able to connect to the edge data centers instead of directly connecting to the cloud data centers, thereby not only it is having low latency but also it will improve the scalability and it will turn the cloud which is highly centralized into the distributed environment for doing this processing. So, IoT will be major beneficiary of this edge data centers.

Healthcare is also other applications. Autonomous vehicle is a new kind of applications which is only possible to be supported with the help of edge data centers because the edge data centers is used to collect process and share the data between the vehicle and other networks, and if the network is at the edge so this kind of computing which is required to run the autonomous vehicle will be done very close by.

(Refer Slide Time: 28:18)



# Scalable Computing at network edge

- Evolutionary changes that have occurred in **distributed edge and cloud computing** over the past 30 years, **driven by applications with variable workloads, low-latency usecase and large data sets .**
- Evolutionary changes in machine architecture, operating system platform, network connectivity, and application workload.
- **Edge computing** uses multiple computers at network edge to solve large-scale problems locally and over the Internet. Thus, **distributed edge computing becomes data-intensive and network-centric.**
- **The emergence of distributed edge computing clouds** instead demands high-throughput computing (HTC) systems built with distributed computing technologies.
- **High-throughput computing (HTC)** appearing as computer clusters, service-oriented, computational grids, peer-to-peer networks, Internet clouds and edge, and the future Internet of Things.

Introduction to Edge Data Center

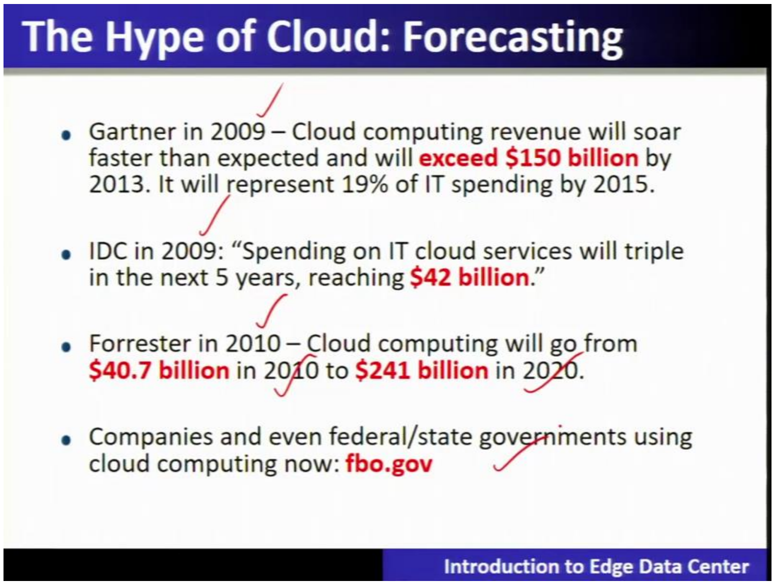
Now, scalable computing at the network edge. So, what we have seen is that evolutionary changes have occurred in the distributed edge and the cloud computing over the last 30 years, that is called the evolution and innovation. So, this is driven by these applications which requires, which is of variable workloads and sometimes requires the low latency use cases and often considered with the large data sets, for example artificial intelligence use cases.

So, these evolutionary changes are very much needed in the architecture of the cloud which is highly centralized to support these kind of workloads through the network connectivity operating system platform and the machine architecture. So, thereby development that is converting the cloud which is highly centralized into the decentralized or distributed, so it introduces or the concept of edge computing.

So, edge computing uses multiple computers at the network edge to solve the large-scale problem locally over the internet. Thus, distributed edge computing becomes data-centric and network-centric. The emergence of distributed edge computing clouds demands high throughput computing built with the distributed computing technologies.

So, this is possible with the innovation in the chip manufacturers such as the chip manufacturer companies which are manufacturing CPU, GPU and FPU, various other type of accelerator processors at a lower cost and with a small size which can be now equipped at the edge data centers or edge cloud. Now, another thing is called high throughput computing clusters which supports the peer-to-peer network, network edge and also to support the future Internet of Things.

(Refer Slide Time: 30:20)



## The Hype of Cloud: Forecasting

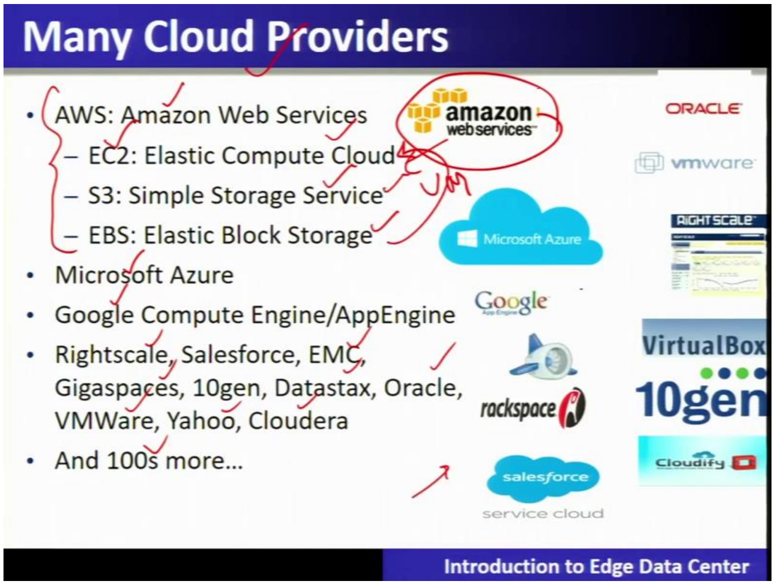
- Gartner in 2009 – Cloud computing revenue will soar faster than expected and will **exceed $150 billion** by 2013. It will represent 19% of IT spending by 2015.
- IDC in 2009: "Spending on IT cloud services will triple in the next 5 years, reaching **$42 billion**."
- Forrester in 2010 – Cloud computing will go from **$40.7 billion** in 2010 to **$241 billion** in 2020.
- Companies and even federal/state governments using cloud computing now: **fbo.gov**

Introduction to Edge Data Center

So, thereby, let us go ahead looking into the cloud model of computing. So, cloud computing started with very hype in 2009, if we trace back. Gartner had forecasted about the technology which is called a cloud computing. So, Gartner in 2009 predicted that cloud computing revenue will soar faster than expected by 2013, and it will represent 19 percent of entire IT spending by 2015.

IDC in 2009 also has hyped about the cloud computing saying that the spending on IT cloud services will triple in the next five years. Forester in 2010 also hyped about the cloud computing that it will go from 2010 to 2020 in many folds of increase. Companies, even the government organizations were also very hopeful about the usage of the cloud for their usage.

(Refer Slide Time: 31:20)



## Many Cloud Providers

- AWS: Amazon Web Services
  - EC2: Elastic Compute Cloud
  - S3: Simple Storage Service
  - EBS: Elastic Block Storage
- Microsoft Azure
- Google Compute Engine/AppEngine
- Rightscale, Salesforce, EMC, Gigaspaces, 10gen, Datastax, Oracle, VMWare, Yahoo, Cloudera
- And 100s more…

VM
amazon web services
ORACLE
vmware
RIGHTSCALE
Microsoft Azure
Google App Engine
VirtualBox
10gen
rackspace
Cloudify
salesforce
service cloud

Introduction to Edge Data Center

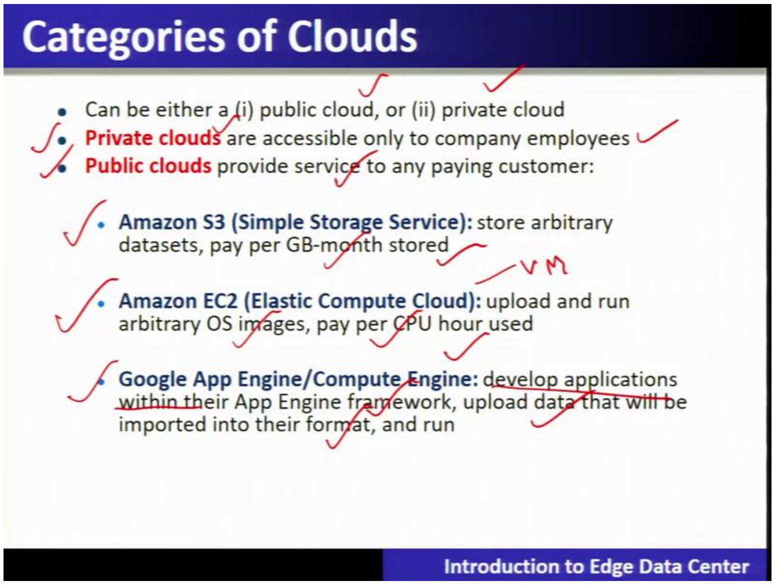
So, thereby it has given the rise to many cloud providers to come in this particular business and take over this kind of opportunities of innovation. So, the companies like Amazon, Google, Microsoft, Salesforce, EMC, Gigaspaces, 10gen, Datastax, Oracle, VMWare, Yahoo, Cloudera and many more companies whose icons are shown over here, they getting started building their cloud and providing these services.

Now, how the cloud is giving the services is in a following manner, we will discuss one such use case that is AWS, that is Amazon Web Service. All other cloud providers are also following the similar kind of offering from the cloud. So, AWS provides three type of services. The first one is called EC2, that is elastic compute cloud. It is nothing but in the form of virtual machines.

So, virtual machines are offered to the users on pay-as-you-go by, pay-as-you-use. So, its not like physical machine where you have to purchase upfront whether you use it or not but in the virtual machine that is the cloud will also provide you the same virtual machine that is with a processor for compute, memory for storage and the network for communication. So, that is called elastic compute cloud service, that is called virtual machines.

AWS also provides a storage service which is called simple storage service, called S3. So, S3 is used for object storage. So, storage service is being provided. For example, the amount of storage which you use, you have to pay in this particular service. Another type of service which AWS provides is called elastic block storage. This is the storage which is required with the virtual machines or with the EC2 instances. So, these are the different type of cloud providers and different type of offerings which are being given by these companies.

(Refer Slide Time: 33:45)



## Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- **Private clouds** are accessible only to company employees
- **Public clouds** provide service to any paying customer:
  - **Amazon S3 (Simple Storage Service):** store arbitrary datasets, pay per GB-month stored
  - **Amazon EC2 (Elastic Compute Cloud):** upload and run arbitrary OS images, pay per CPU hour used
  - **Google App Engine/Compute Engine:** develop applications within their App Engine framework, upload data that will be imported into their format, and run

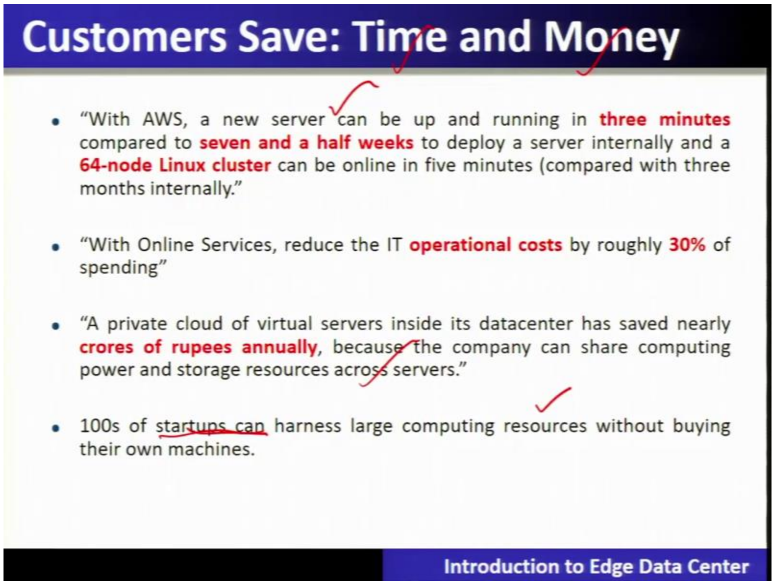
Introduction to Edge Data Center

So, thereby, all these cloud services is divided into two different types, one is called public cloud which is being provided by the companies like Amazon, Google, Microsoft for anyone. The other type of cloud is called a private cloud which is being set up in the similar manner at the premise owned by the company and it is only used by the company employee, and it is not public, it is private.

So, there are two types of cloud which is possible one is called public cloud which we have seen in the previous slide and the private cloud. Now, public cloud provides the service to any customer who wants to use it and for that he has ready to pay the usage. So, the examples which we have are telling you about Amazon S3, that is simple storage service. So, it provides the storage services where the users can use the storage to store the data set and the user has to pay per GB per month of storage which it is using through this Amazon S3 service.

Similarly, Amazon EC2, that is elastic compute cloud is also another compute services which is possible in the form of virtual machines. So, wherein the user can load their operating system images and they have to pay per CPU hour usage. Third type of service which is provided by Google app engine, that is also called a compute engine, you are not given the virtual machines but the entire platform you are given and you have to use the programming language like python or some other programming to build your application and upload your data and it will allow you to run. That is called the platforms.

(Refer Slide Time: 35:51)



## Customers Save: Time and Money

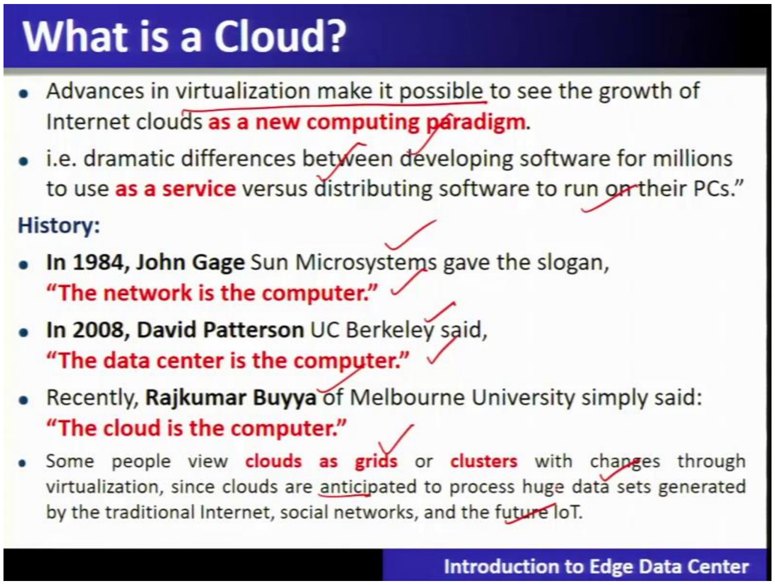
- "With AWS, a new server can be up and running in **three minutes** compared to **seven and a half weeks** to deploy a server internally and a **64-node Linux cluster** can be online in five minutes (compared with three months internally."
- "With Online Services, reduce the IT **operational costs** by roughly **30%** of spending"
- "A private cloud of virtual servers inside its datacenter has saved nearly **crores of rupees annually**, because the company can share computing power and storage resources across servers."
- 100s of startups can harness large computing resources without buying their own machines.

Introduction to Edge Data Center

Now, using, why the cloud is becoming popular is that the customers are saving both time and money. So, let us see how. For example, if you, instead of buying upfront a new machine, this takes lot of time in market survey giving the purchase order and spending money upfront. So, what you can do is you can save the time by taking the services that is called virtual machines, that is through EC2 instances you can take that can be available instantly and also, so thereby, you are saving the time.

And also, money is also saved. For example, if you are not going to use for a longer time, that physical machine than that particular service or service which is being taken by the cloud in the form of a virtual machine will be saving the money. So, lot of startup companies, they start their businesses by harnessing the large computing resources without buying these kind of machines or hardware up front, but rather they are being supported by these cloud offerings of compute storage and network.

(Refer Slide Time: 37:07)



## What is a Cloud?

- Advances in virtualization make it possible to see the growth of Internet clouds **as a new computing paradigm**.
- i.e. dramatic differences between developing software for millions to use **as a service** versus distributing software to run on their PCs."

**History:**

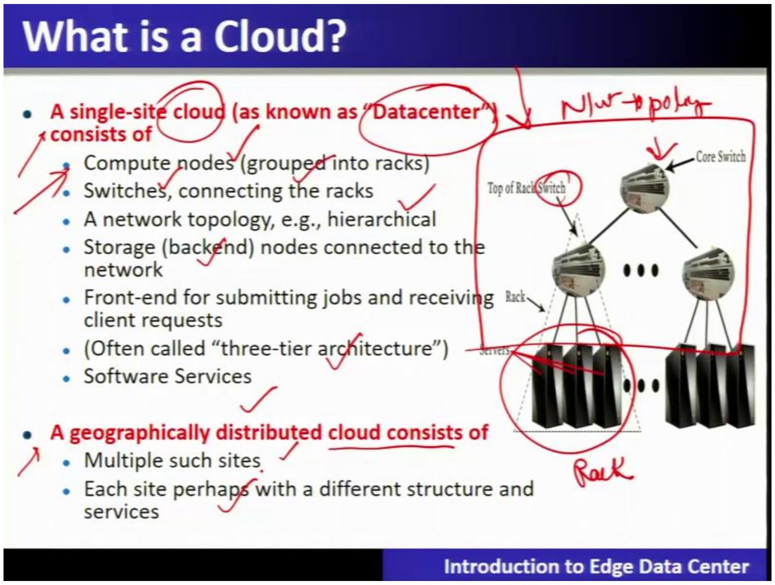
- **In 1984, John Gage** Sun Microsystems gave the slogan, **"The network is the computer."**
- **In 2008, David Patterson** UC Berkeley said, **"The data center is the computer."**
- Recently, **Rajkumar Buyya** of Melbourne University simply said: **"The cloud is the computer."**
- Some people view **clouds as grids** or **clusters** with changes through virtualization, since clouds are anticipated to process huge data sets generated by the traditional Internet, social networks, and the future IoT.

Introduction to Edge Data Center

So, let us see what is a cloud. So, advances in the virtualization has made it possible the growth of these clouds that is called internet cloud that is called a new computing paradigm. So, thereby this virtualization allows the shearing across multiple clients or the users of the same infrastructure in the form of a virtual machine. So, this particular technology which is called the virtualization technology that is also has made the dramatic difference between developing the software for the millions to use as a service versus distributing the software to run on their PCs.

So, let us see that about what is the cloud, let us trace back the history. In 1984, long back, the Sun Microsystems gave the slogan that network is the computer. So, and in 2008, David Patterson from UC Berkeley says that data set is the computer. And recently, Rajkumar Buyya from Melbourne, then they say that the cloud is the computer. So, some people view the cloud as the grid or the cluster which changes through the virtualization. These clouds are anticipated to process huge data set generated by traditional internet or the Internet of Things.

(Refer Slide Time: 38:36)



## What is a Cloud?

- A single-site cloud (as known as "Datacenter") consists of
  - Compute nodes (grouped into racks)
  - Switches, connecting the racks
  - A network topology, e.g., hierarchical
  - Storage (backend) nodes connected to the network
  - Front-end for submitting jobs and receiving client requests
  - (Often called "three-tier architecture")
  - Software Services
- A geographically distributed cloud consists of
  - Multiple such sites
  - Each site perhaps with a different structure and services

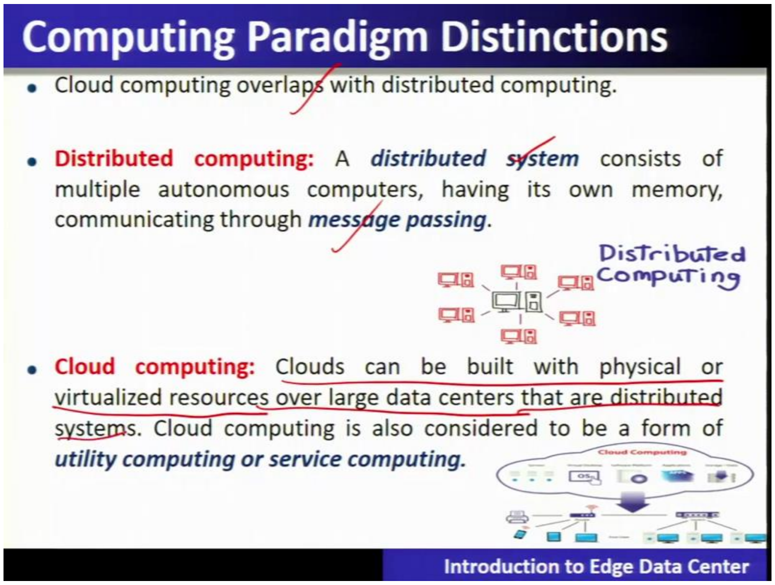
Introduction to Edge Data Center

Now, what is there inside cloud how that is all being provided these services. So, this cloud, if you go inside the cloud you will see there are two types of technologies which is used. One is called single-site cloud which is nothing but also called as a data center and this particular data center houses lot of compute nodes which are grouped into the racks which are shown over here. This is a rack.

These racks are now connected together with the networking which is called the switch. And also, these racks using the switch they are connected with the top of, with the core switch. So, this particular network topology which connects these compute nodes is a tree like structure or say hierarchical network topology. Now, as far as the storage is concerned which is at the back end of the nodes they are connected to the network here in this case.

And front end will be in the form of submitting the jobs and receiving the client requests. So, this is a three-tier architecture which runs the software services on it. So, this is called the data center. Now, another way of looking the cloud is that it is a geographically distributed cloud which consists of multiple such sites, that is multiple data centers which are geographically distributed in form of cloud. And these sites perhaps with the different structure and services are being organized.

(Refer Slide Time: 40:21)



**Computing Paradigm Distinctions**

- Cloud computing overlaps with distributed computing.
- **Distributed computing:** A ***distributed system*** consists of multiple autonomous computers, having its own memory, communicating through ***message passing***.

Distributed Computing

- **Cloud computing:** Clouds can be built with physical or virtualized resources over large data centers that are distributed systems. Cloud computing is also considered to be a form of ***utility computing or service computing.***

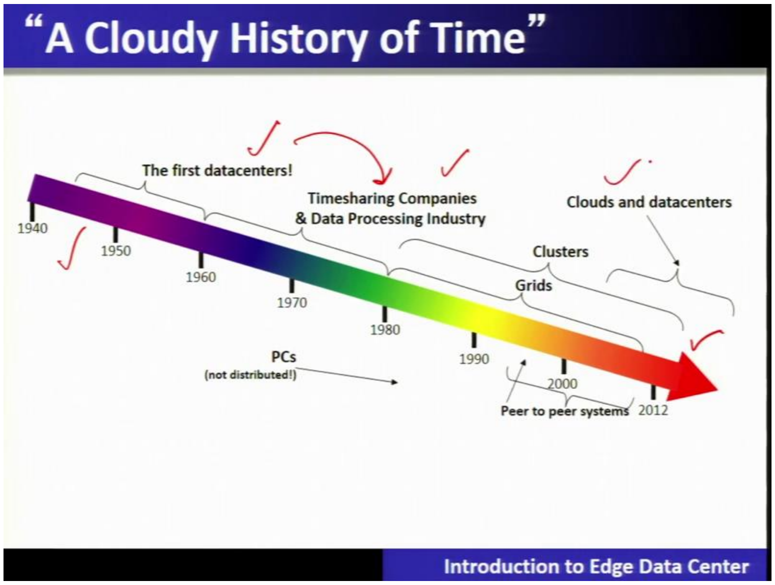
Introduction to Edge Data Center

So, computing paradigm distinctions is that if you see this particular cloud computing, it often overlaps with the distributed computing concepts like you have seen the geo-distributed cloud or in-site cloud also you can see that the hundreds and thousands of compute nodes, they are connected to the together with the network forming a full-fledged distributed system.

So, cloud is nothing but an example or a model of a distributed system. And distributed system consists of multiple autonomous computers having their own memory and they communicate using the message passing. So, the cloud computing, if you compare with the distributed computing what you find is that clouds are built with the physical or the virtualized resources over the large data centers that are distributed systems.

So, cloud computing is also considered to be a form of utility computing or the service computing. Before we go ahead let me repeat this particular definition of a cloud computing. So, clouds are built with the physical or a virtualized resources, like we have seen, over the large data centers and they are the distributed systems.

(Refer Slide Time: 41:37)



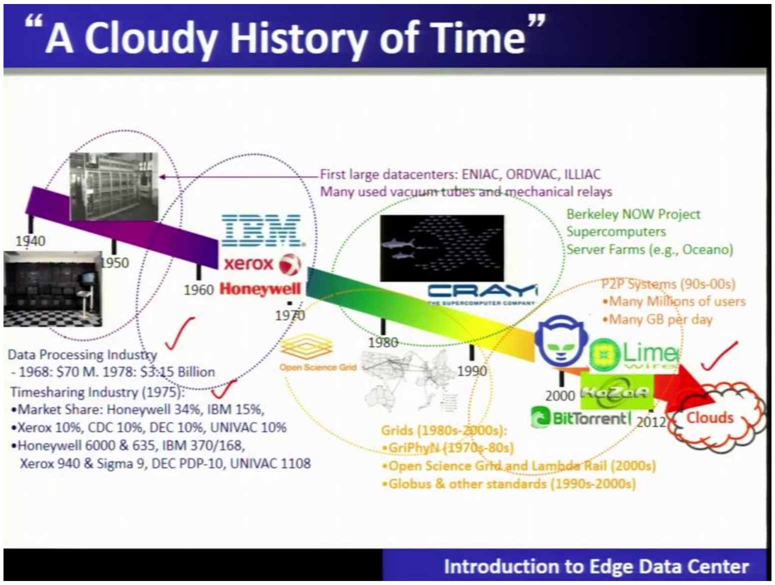
“A Cloudy History of Time”

Introduction to Edge Data Center

So, let us trace back the history of the development of the cloud over the time. Very beginning days, let us say in 1940s to 1960s, what you find is also there was a data center where a very big computers were housed and they were called a data center. Then the time shift came and the trend shifted.

There was a paradigm shift from data centers of 1940s to becoming the time-shared computer systems, and that was the trend in the data processing industries which came. After the 1980s these time-shared machines also have made a paradigm shift into the clusters and the grids. And after 2012 what you will find is that this is being matured in the form of cloud and the data centers with the help of peer-to-peer technologies.

(Refer Slide Time: 42:39)



## "A Cloudy History of Time"

First large datacenters: ENIAC, ORDVAC, ILLIAC
Many used vacuum tubes and mechanical relays

Berkeley NOW Project
Supercomputers
Server Farms (e.g., Oceano)

P2P Systems (90s-00s)
- Many Millions of users
- Many GB per day

Data Processing Industry
- 1968: $70 M. 1978: $3.15 Billion

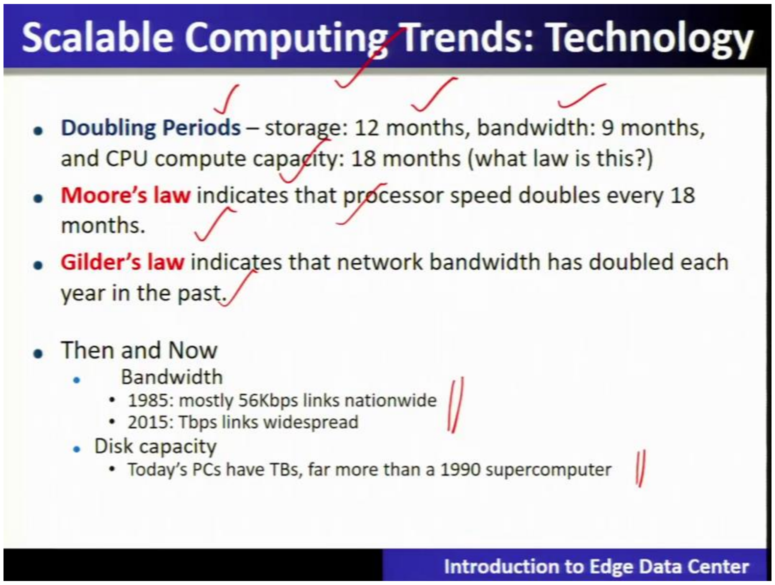
Timesharing Industry (1975):
- Market Share: Honeywell 34%, IBM 15%,
- Xerox 10%, CDC 10%, DEC 10%, UNIVAC 10%
- Honeywell 6000 & 635, IBM 370/168, Xerox 940 & Sigma 9, DEC PDP-10, UNIVAC 1108

Grids (1980s-2000s):
- GriPhyN (1970s-80s)
- Open Science Grid and Lambda Rail (2000s)
- Globus & other standards (1990s-2000s)

Introduction to Edge Data Center

So, if you see those kind of industries, that is called data processing industry was there in the beginning and which requires this kind of data centers of a huge mainframe machines. And then came the time-shared industries where the machines were through the operating system they were shared for doing the computations. And then these kind of technology changes and 2012, you see the advent of data centers and the cloud data centers. After that, which was precursor to the grid computing or a cluster computing.

(Refer Slide Time: 43:15)



## Scalable Computing Trends: Technology

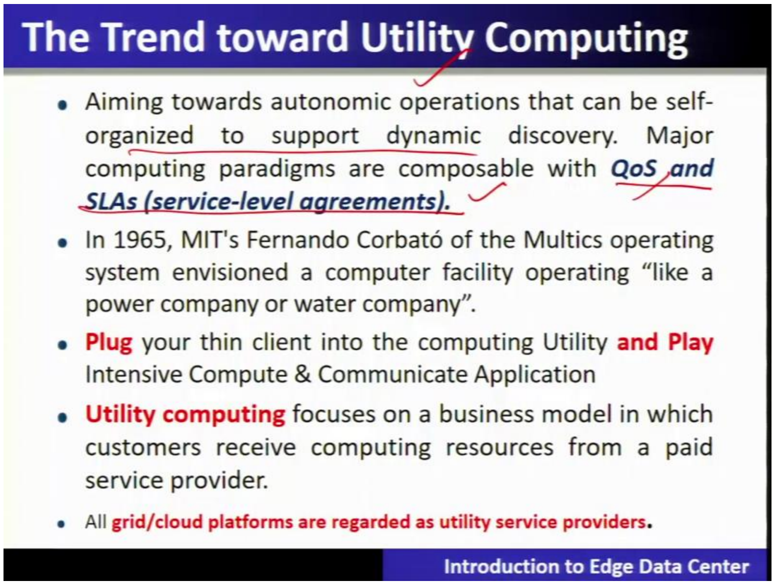
- **Doubling Periods** – storage: 12 months, bandwidth: 9 months, and CPU compute capacity: 18 months (what law is this?)
- **Moore's law** indicates that processor speed doubles every 18 months.
- **Gilder's law** indicates that network bandwidth has doubled each year in the past.
- Then and Now
  - Bandwidth
    - 1985: mostly 56Kbps links nationwide
    - 2015: Tbps links widespread
  - Disk capacity
    - Today's PCs have TBs, far more than a 1990 supercomputer

Introduction to Edge Data Center

So, if you see this scalable computing trend technologies, what you will find is that there is a pattern of doubling period. That is the storage after every 12-month doubles, that is from the of the same cost the storage after every 12-month doubles, that is the development or innovation, the technology. Similarly, the bandwidth after every 9-month doubles. And CPU compute capacity, after every 18 months, they double. And this law is called Moore's Law, which indicates that the processor speeds doubles every 18 months.

There is another law in this direction called Gilder’s law indicates the network bandwidth has doubled each year in the past. So, therefore the bandwidth, you can see, that it is keeps on growing because of the current demand. And the storage capacity also is going. Today, the PCs have terabytes and so on, which was, which used to happen with the super computer long back.

(Refer Slide Time: 44:24)



## The Trend toward Utility Computing

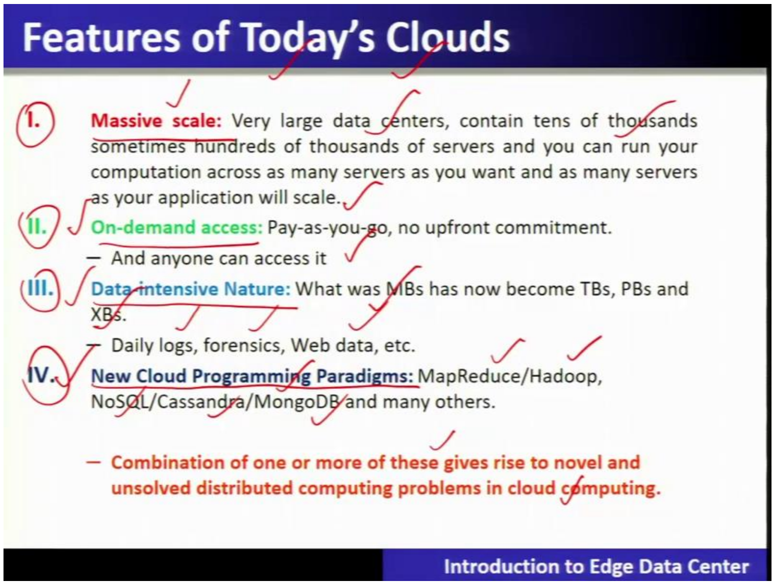
- Aiming towards autonomic operations that can be self-organized to support dynamic discovery. Major computing paradigms are composable with **QoS and SLAs (service-level agreements).**
- In 1965, MIT's Fernando Corbató of the Multics operating system envisioned a computer facility operating “like a power company or water company”.
- **Plug** your thin client into the computing Utility **and Play** Intensive Compute & Communicate Application
- **Utility computing** focuses on a business model in which customers receive computing resources from a paid service provider.
- All **grid/cloud platforms are regarded as utility service providers.**

Introduction to Edge Data Center

So, let us see about these particular trends is that most of these operations, when you talk about hundreds and thousands of computing node put into a data centers, then most of these management related operations and management becomes automatic. They are called self-organized, support dynamic discovery and so on. Therefore, these major computing paradigms are composable, with quality of service and service level agreements.

So, the cloud, if it is providing the utility computing towards utility computing it has to govern with the quality of service and service level agreement. Then only the customers will be using these kind of services with these kind of guarantees. So, let us see that these are the newer developments. So, let us see what are the features of today's cloud.

(Refer Slide Time: 45:15)



## Features of Today's Clouds

I. **Massive scale:** Very large data centers, contain tens of thousands sometimes hundreds of thousands of servers and you can run your computation across as many servers as you want and as many servers as your application will scale.

II. **On-demand access:** Pay-as-you-go, no upfront commitment.

– And anyone can access it

III. **Data-intensive Nature:** What was MBs has now become TBs, PBs and XBs.

– Daily logs, forensics, Web data, etc.

IV. **New Cloud Programming Paradigms:** MapReduce/Hadoop, NoSQL/Cassandra/MongoDB and many others.

– **Combination of one or more of these gives rise to novel and unsolved distributed computing problems in cloud computing.**

Introduction to Edge Data Center

So, there are four features which are now there in today's cloud. The first one is called the massive scale. So, massive scale means that it is cloud, today's cloud is to be seen as the highly centralized set of computing and storage and networking resources. And often, they are called very large data centers. So, these large data centers often house tens to thousands of servers. And to connect these servers you require a huge amount of networking and storage and all these particular things.

And therefore, your application will also scale because these number of servers you can keep on adding and therefore this size is called the massive scale. And that is the characteristics of today's cloud. The second thing is that these highly centralized resources, that is compute storage and the network, often is given on-demand access. So, that means pay as you use. There is no upfront cost of usage.

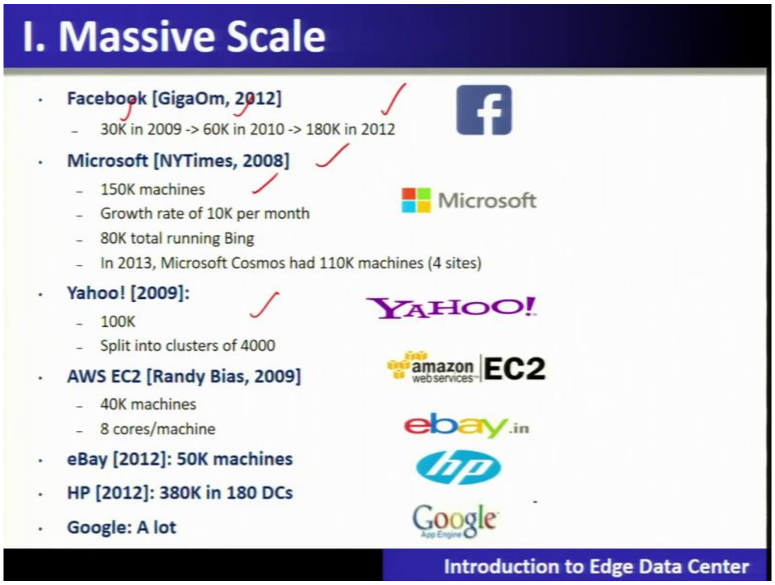
So, on-demand access is to be provided with the help of technology which is called the virtualization. And we will see, this is another characteristics of today's cloud. Third important characteristics of today's cloud is that it supports data intensive nature of computing. So, what you will see is that the workloads which was earlier in the megabytes now is turned into the terabytes, petabyte and zettabytes.

In the form of, let us say, that if you are doing this daily log analysis or a forensic analysis or a web data analytics then this kind of workload is categorized as data-intensive nature of computing. There is a paradigm shift from the earlier way of computing which was more on compute-intensive nature of a computing where super computers and all that things was earlier the need, but now it is a data-intensive nature of computing. And that is the feature of today's cloud.

And also, the cloud programming paradigms also supports the programming of, programming in this particular model, that is in the form of the data centers and to support on demand access and data intensive nature of computing, you require, you have to see a new kind of cloud programming paradigms such as MapReduce, Hadoop, NoSQL, Cassandra, MongoDB and so on.

So, what we have seen so far is that there are four features, I have discussed. They are massive scale, on-demand access, data-intensive nature of computing and new cloud programming paradigm. Now, if one or more of these combinations are available, then it is called the cloud computing problem. And if they are missing, then it is not a cloud computing problem, it is a normal computing problem. So, we will go and see the details of each and every feature which is present in today's cloud.

(Refer Slide Time: 48:47)



## I. Massive Scale

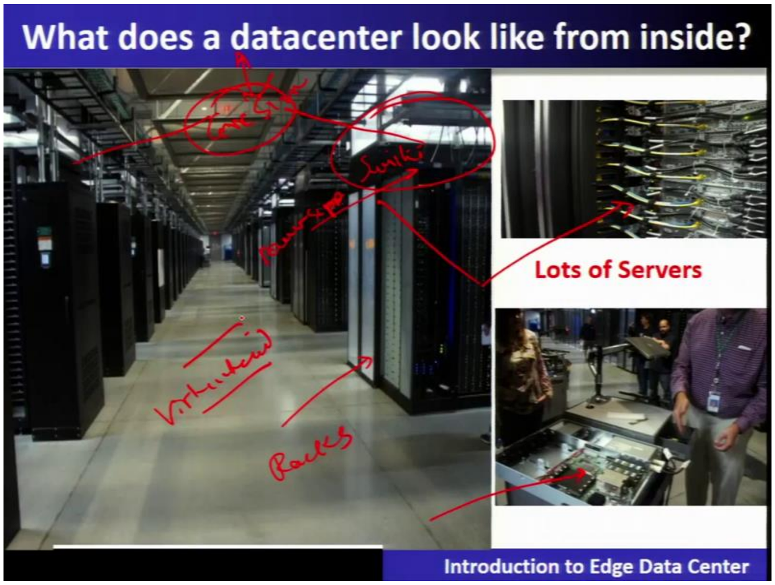
- Facebook [GigaOm, 2012]
  - 30K in 2009 -> 60K in 2010 -> 180K in 2012
- Microsoft [NYTimes, 2008]
  - 150K machines
  - Growth rate of 10K per month
  - 80K total running Bing
  - In 2013, Microsoft Cosmos had 110K machines (4 sites)
- Yahoo! [2009]:
  - 100K
  - Split into clusters of 4000
- AWS EC2 [Randy Bias, 2009]
  - 40K machines
  - 8 cores/machine
- eBay [2012]: 50K machines
- HP [2012]: 380K in 180 DCs
- Google: A lot

Introduction to Edge Data Center

So, let us see the feature which is called the massive scale. In that massive scale, what is seen is that the number of servers in 2009 in the Facebook cloud was 30,000, and which was grown to 60,000, just double after one year. And in 2012, it was 180,000 servers were there. So, that is the scale which is called a massive scale. Microsoft uses 150,000 machines and which are growing at the rate of 10,000 machines per month.

Yahoo, if you see, in 2009, it used 100,000 machines which are split into the clusters of 4,000. Amazon EC2 uses 40,000 machines in 2009 and eBay was using 50,000 machines 2012. And HP was using 380,000 machines in 180 different data centers. Google is using a much, a lot of this machine, that is called massive scale.

(Refer Slide Time: 49:48)



What does a datacenter look like from inside?

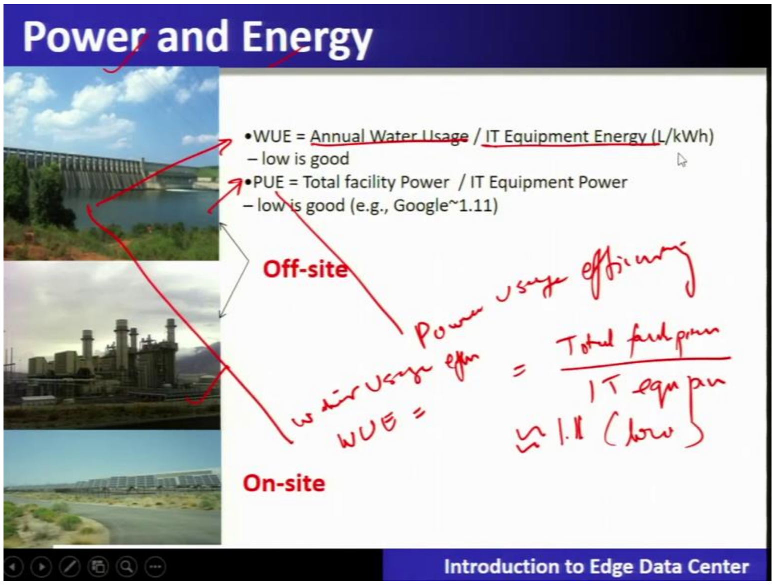
Lots of Servers

Introduction to Edge Data Center

So, if you see inside a data center, you will see a lot of racks and the racks are having lot of servers running within it. So, this particular data center which is storing or which is housing, these are called racks. And inside racks, you will see, these are the servers. And each server will be, will look like in this manner. That is called the rack server. So, this particular rack is having this network, that is called top of the rack switch.

This is called, they are sharing a common switch and also the common UPS or the power supply of a rack, is being shared across these lot of servers. And this top of the rack switch is now further connected with, together they are connected with a complete core switch, and core switch is now connected to the internet so that the client request can be received and being computed over here. So, this entire infrastructure is virtualized so that this can be shared by different clients in this particular manner.

(Refer Slide Time: 51:16)



## Power and Energy

- WUE = Annual Water Usage / IT Equipment Energy (L/kWh)
  - low is good
- PUE = Total facility Power / IT Equipment Power
  - low is good (e.g., Google~1.11)

Off-site

On-site

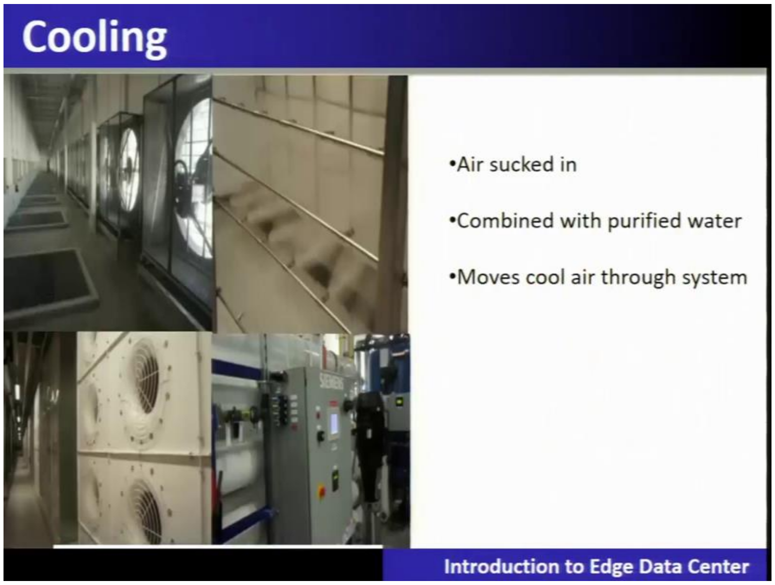
Power usage efficiency = $\frac{\text{Total facil power}}{\text{IT eqn pwr}}$

≃ 1.11 (low)

water usage effn

WUE =

Introduction to Edge Data Center



## Cooling

- Air sucked in
- Combined with purified water
- Moves cool air through system

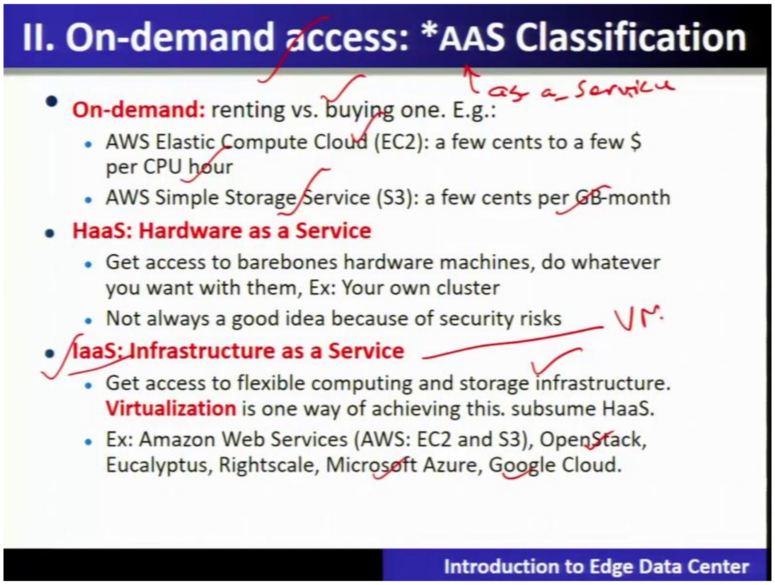
Introduction to Edge Data Center

So, if you see this huge data center which is having hundreds and thousands of machines which requires a cooling and this kind of power supply infrastructure. So, power and energy are the two important major requirement for running this kind of data centers. So, let us see that the power comes from the power house, that is thermal power stations are supplying this particular power.

So, there is a metric which will measure whether the data center is effectively using it or not. So, that is called power usage efficiency. Power usage efficiency is a formula which says that total facility power divided by IT equipment power. And if it is close to 1.11, very close to one then if it is, then it is called a low and it is very good. If it is more, then it is not that good. So, this is the metric which will check whether the power is efficiently used in the data center or not.

Now similarly, the so much of power, if it is used in the machines that is in the form of servers, then lot of heat is being generated. How to cool this? Water is used to cool this particular heat. So, there is another way of measuring whether the power is efficiently or effectively used or not, that is called a water usage efficiency or WUE. So, WUE says that annual water usage divided by IT equipment energy. And if it is low, then it is good. So, these are the basic requirements to run the data center where power and the water is very much effectively required to be used. And these are the metric which we have already discussed.

(Refer Slide Time: 53:25)



## II. On-demand access: *AAS Classification

as a Service

- **On-demand:** renting vs. buying one. E.g.:
  - AWS Elastic Compute Cloud (EC2): a few cents to a few $ per CPU hour
  - AWS Simple Storage Service (S3): a few cents per GB-month
- **HaaS: Hardware as a Service**
  - Get access to barebones hardware machines, do whatever you want with them, Ex: Your own cluster
  - Not always a good idea because of security risks
- **IaaS: Infrastructure as a Service**
  - Get access to flexible computing and storage infrastructure. **Virtualization** is one way of achieving this. subsume HaaS.
  - Ex: Amazon Web Services (AWS: EC2 and S3), OpenStack, Eucalyptus, Rightscale, Microsoft Azure, Google Cloud.

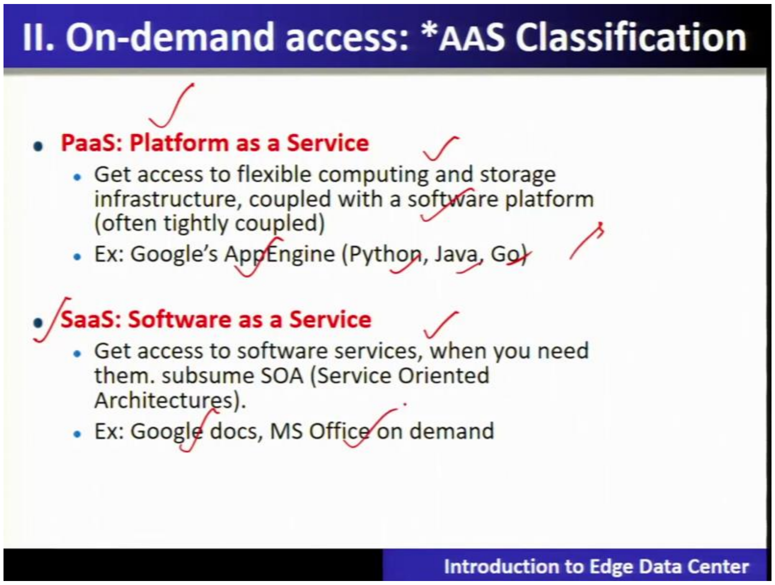
VM.

Introduction to Edge Data Center

Second feature of today's cloud is called On-Demand access. It is also called as a service, called as a service classification. So, let us not go in more detail. So, on-demand access means that we are buying the compute resources out of the cloud offering. For example, if you want to use a virtual machine then you can use the elastic compute cloud feature by paying few cents per hour usage of that particular virtual machine. Similarly, if you are using only the storage then storage cost also requires a few cents per GB per month.

Now this particular on-demand service is being categorized in different ways. The first one is called infrastructure-as-a-service. So, infrastructure-as-a-service is a way of on-demand access and is being provided by major cloud vendors which we have seen, whether it is Microsoft Azure, Microsoft, Google, Right force, they are all in this kind of business that is called infrastructure-as-a-service. So, infrastructure-as-a-service is being facilitated with the help of the technology which is called the virtualization technology and this particular infrastructure-as-a-service is providing the virtual machines.

(Refer Slide Time: 55:02)



## II. On-demand access: *AAS Classification

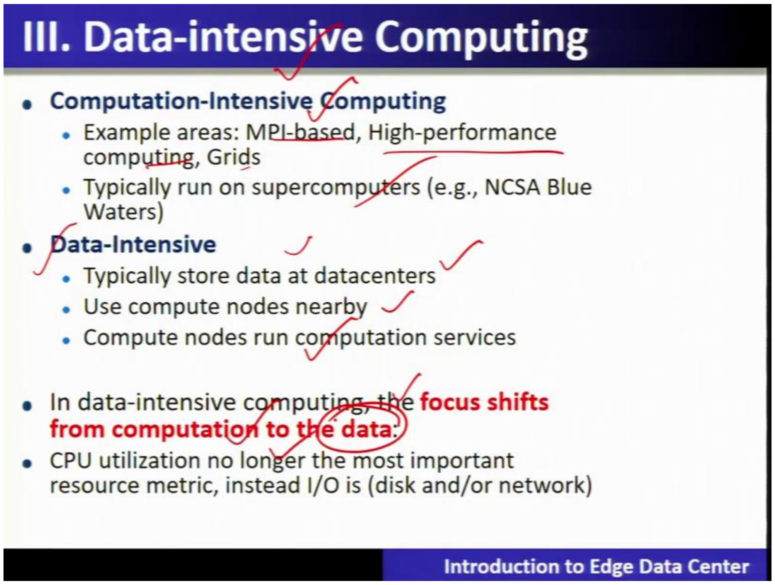
- **PaaS: Platform as a Service**
  - Get access to flexible computing and storage infrastructure, coupled with a software platform (often tightly coupled)
  - Ex: Google's AppEngine (Python, Java, Go)
- **SaaS: Software as a Service**
  - Get access to software services, when you need them. subsume SOA (Service Oriented Architectures).
  - Ex: Google docs, MS Office on demand

Introduction to Edge Data Center

So, other type of on demand access to the cloud resources is called platform-as-a-service. So, here the virtual machine is not given, but rather it is given a flexible computing and storage infrastructure where the users using a Python or a Java program, they can write the programs, they can develop the program and upload their data and this program will run on that particular data.

And so, this kind of thing is called platform-as-a-service. The example is Google's app engine. Third type of on-demand access which is being provided by Major cloud providers they are called software-as-a-service. So, you can directly get access to the services which is run by the software such as you might have used Google docs, MS Office on demand and so on.

(Refer Slide Time: 55:55)



## III. Data-intensive Computing

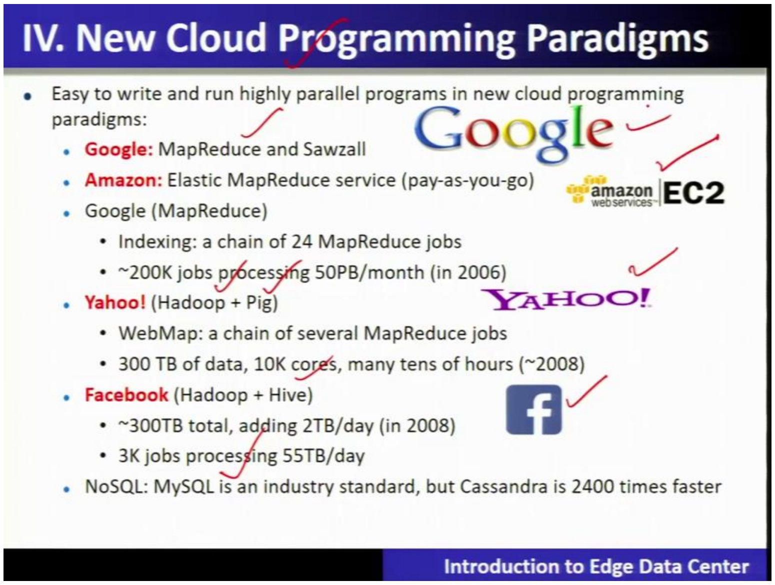
- **Computation-Intensive Computing**
  - Example areas: MPI-based, High-performance computing, Grids
  - Typically run on supercomputers (e.g., NCSA Blue Waters)
- **Data-Intensive**
  - Typically store data at datacenters
  - Use compute nodes nearby
  - Compute nodes run computation services
- In data-intensive computing, the **focus shifts from computation to the data**:
- CPU utilization no longer the most important resource metric, instead I/O is (disk and/or network)

Introduction to Edge Data Center

Third type of feature which today's cloud is having is called data-intensive nature of computing. So, data-intensive nature of computing is quite deviation from the earlier paradigm of computing called compute-intensive computing which uses high performance computing super computer and message passing interface. But nowadays, it is more on data-intensive nature of computing where you have a large data set which you need to store. For example, IoT sensors are sending their data to the data center to store and use the data for training the machine learning models.

So, therefore if the data is stored at the data center, then you have to now bring the compute nodes where the data is stored for performing the computation. So, compute this type of way things is where the compute is to be done very close to where the data is stored called data-intensive nature of computation in contrast to the compute-intensive nature of computation where the data is small and it goes to that particular compute in the super computing model. So, in the data intensive nature of computing the focus is shipped from computation to the data intensive. So, here the CPU utilization is no longer important. The only thing is that input output thing has to be reduced, Io, load has to be reduced.

(Refer Slide Time: 57:25)



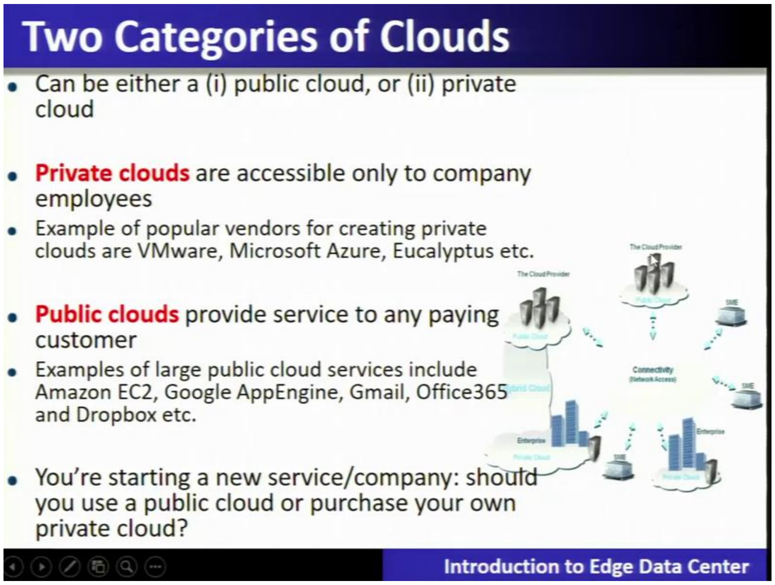
## IV. New Cloud Programming Paradigms

- Easy to write and run highly parallel programs in new cloud programming paradigms:
  - **Google:** MapReduce and Sawzall
  - **Amazon:** Elastic MapReduce service (pay-as-you-go)
  - Google (MapReduce)
    - Indexing: a chain of 24 MapReduce jobs
    - ~200K jobs processing 50PB/month (in 2006)
  - **Yahoo!** (Hadoop + Pig)
    - WebMap: a chain of several MapReduce jobs
    - 300 TB of data, 10K cores, many tens of hours (~2008)
  - **Facebook** (Hadoop + Hive)
    - ~300TB total, adding 2TB/day (in 2008)
    - 3K jobs processing 55TB/day
  - NoSQL: MySQL is an industry standard, but Cassandra is 2400 times faster

Introduction to Edge Data Center

So, now the new cloud programming paradigm is there to develop the program and to use this particular model that is the cloud data center. So, the programming languages such as MapReduce, Elastic MapReduce, Hadoop, Pig, Hive, they are all developed by different, NoSQL, they are all developed by different companies such as Facebook has developed a Hadoop and Hive, Yahoo has developed Hadoop and Pig, EC, Amazon has developed Elastic MapReduce, Google has developed the MapReduce. They are all different new cloud programming paradigms.

(Refer Slide Time: 58:05)



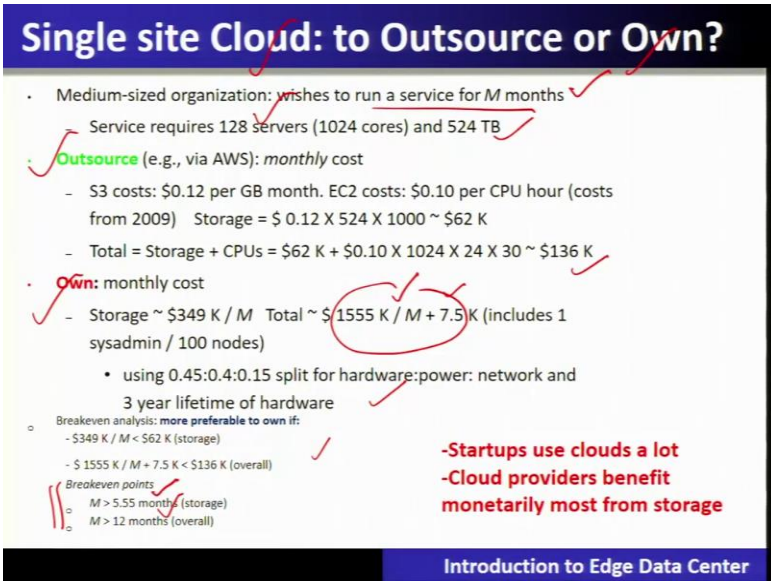
## Two Categories of Clouds

- Can be either a (i) public cloud, or (ii) private cloud
- **Private clouds** are accessible only to company employees
- Example of popular vendors for creating private clouds are VMware, Microsoft Azure, Eucalyptus etc.
- **Public clouds** provide service to any paying customer
- Examples of large public cloud services include Amazon EC2, Google AppEngine, Gmail, Office365 and Dropbox etc.
- You're starting a new service/company: should you use a public cloud or purchase your own private cloud?

Introduction to Edge Data Center

So, this is a new category that is a private cloud and the public cloud. These are the offering.

(Refer Slide Time: 58:16)



## Single site Cloud: to Outsource or Own?

- Medium-sized organization: wishes to run a service for *M* months
  - Service requires 128 servers (1024 cores) and 524 TB
- Outsource (e.g., via AWS): *monthly* cost
  - S3 costs: $0.12 per GB month. EC2 costs: $0.10 per CPU hour (costs from 2009) Storage = $ 0.12 X 524 X 1000 ~ $62 K
  - Total = Storage + CPUs = $62 K + $0.10 X 1024 X 24 X 30 ~ $136 K
- Own: monthly cost
  - Storage ~ $349 K / *M* Total ~ $ 1555 K / *M* + 7.5 K (includes 1 sysadmin / 100 nodes)
    - using 0.45:0.4:0.15 split for hardware:power: network and 3 year lifetime of hardware
- Breakeven analysis: **more preferable to own if:**
  - $349 K / *M* < $62 K (storage)
  - $ 1555 K / *M* + 7.5 K < $136 K (overall)
  - *Breakeven points*
    - *M* > 5.55 months (storage)
    - *M* > 12 months (overall)

-Startups use clouds a lot
-Cloud providers benefit monetarily most from storage

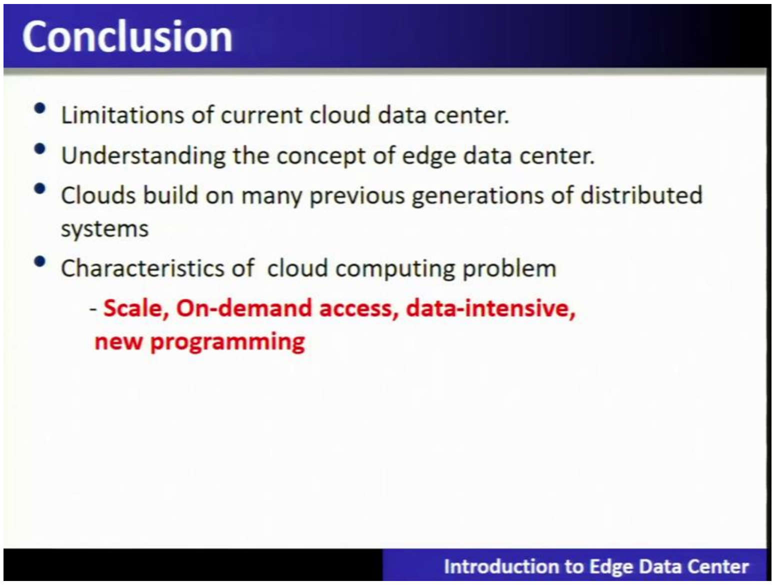
Introduction to Edge Data Center

If you see that whether you use the single-site cloud or you own to have the private cloud. If you want to decide then economies of the scale can be calculated in this. For example, if you are running, if starting a startup company and you want to run your service for let us say M months but your service requires let us say that to purchase 128 servers with 524 terabytes of storage to run your organization’s IT infrastructure.

So, one thing is that, so if you purchase this infrastructure, it will cost something in this particular amount. And if you want to run for M months, so this particular cost will be divided by M here in this case. So, it comes out to be this value how if you want to lease the services from the cloud and you want to calculate for over M month what you will find is that if this particular cost of outsourcing is less than owning, then you have to go for the.

So, what is being found in most of the cases on an average that if you require the storage for more than five months and if you require the overall infrastructure more than 12 months, then you can own. It if it is less than that, that is why the startup companies are looking for this cloud services.

(Refer Slide Time: 59:50)



## Conclusion

- Limitations of current cloud data center.
- Understanding the concept of edge data center.
- Clouds build on many previous generations of distributed systems
- Characteristics of cloud computing problem
  - **Scale, On-demand access, data-intensive, new programming**

Introduction to Edge Data Center

So, let us conclude. In the summary, let me tell you that in this lecture, what we have brought you is about the limitation of the current cloud data centers, and we have also given you the concept of new innovation in the cloud data center, that is called edge data centers. And we have also told you about the cloud which are how they are being developed as, with the help of the concepts of a distributed systems. And therefore, we have also seen the characteristics of cloud computing that is the scale on-demand access, data-intensive nature of computing and new programming paradigm. With this, let us conclude this lecture. Thank you.