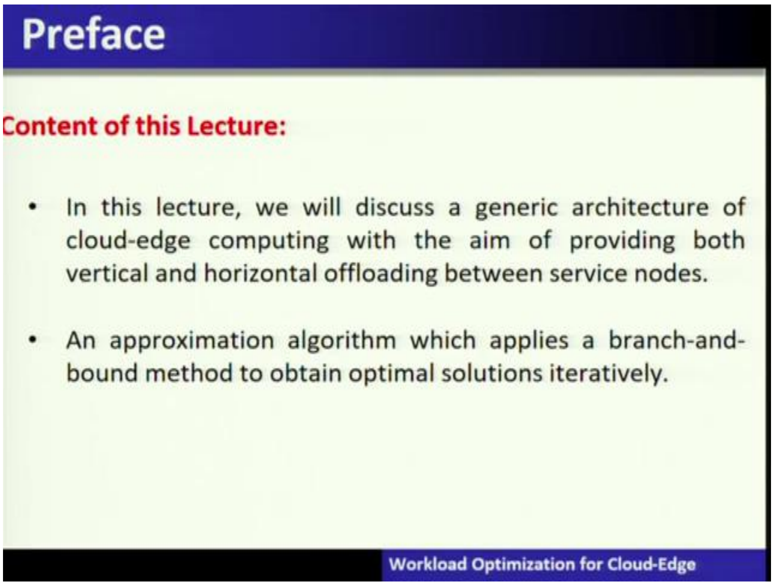
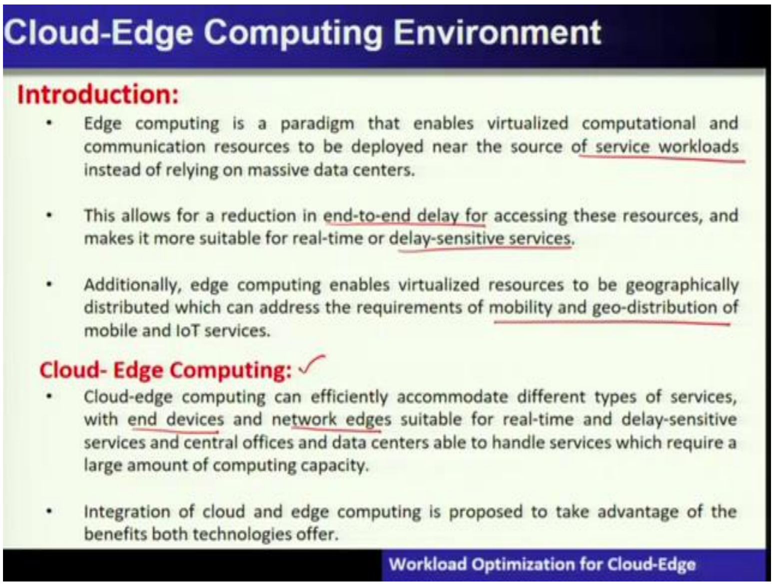
Foundation of Cloud IoT Edge ML
Professor Rajiv Misra
Department of Computer Science and Engineering
Indian Institute of Technology, Patna
Lecture 14
Vertical and Horizontal Offloading for Cloud-Edge

(Refer Slide Time: 0:21)



I am Doctor Rajiv Misra from IIT, Patna. Title of this lecture is Vertical and Horizontal Offloading for Cloud-edge. Content of this lecture in this lecture we will discuss a generic architecture of cloud edge computing with the aim of providing both vertical and horizontal offloading between the service nodes.

(Refer Slide Time: 0:33)

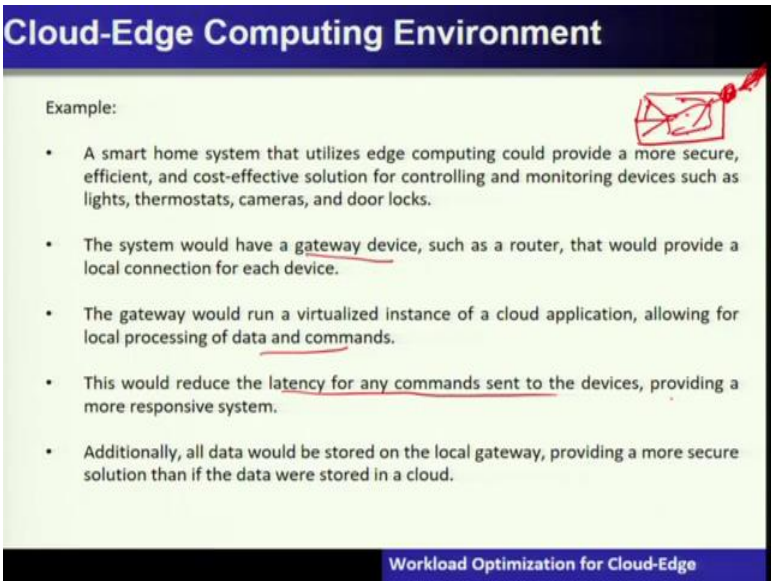


So, cloud edge computing environment let us go for the introduction. So, edge computing is a paradigm that enables virtualized computational and communication resources to be deployed near the source of the service workloads instead of relying on massive data centers. Now, this allows for the reduction in end to end delay for accessing these resources and makes it more suitable for real time and delay sensitive services.

So, additionally, the edge computing enables virtualized resources to be geographically distributed which can address the requirements of mobility and geo distribution of mobile and IoT services. Thereby, let us introduce about the cloud edge computing. So, cloud edge computing can essentially efficiently accommodate different type of services with end devices and network edge is suitable for real time and delay sensitive applications.

Now, integration of this edge computing and cloud is proposed to undertake the advantage of benefit is of both the technologies in this.

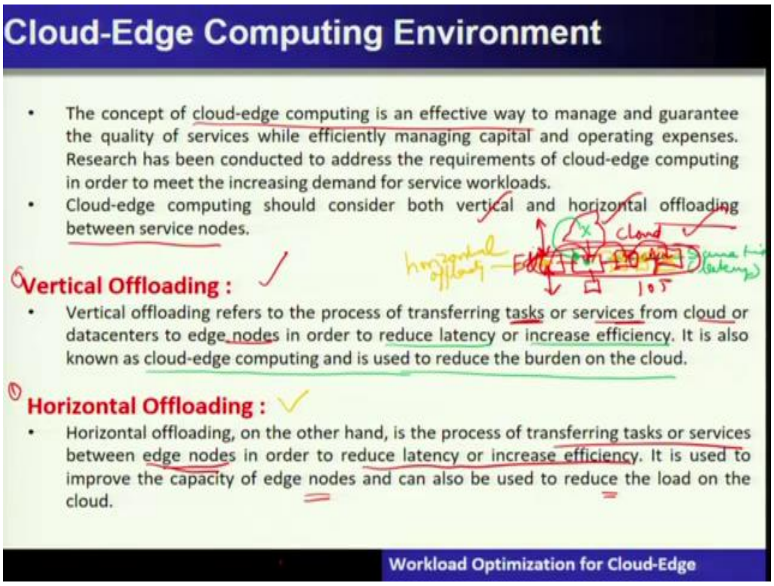
(Refer Slide Time: 1:41)



So, let us take this example where this cloud edge computing model or environment is more suited. Now, consider the smart home situation which utilizes the edge computing could provide a more secure, efficient cost effective solution for controlling monitoring devices such as lights, thermostats, camera door locks, the system would have the gateway device such as the router that would provide the local connection to each device.

So, think of that this is a smart home where all these devices are connected. And they have a kind of a gateway as a router which will connect to the internet and all the devices are internally connected. In this particular case so the gateway would run a virtualized instance of cloud applications allowing local processing of data and command this will reduce the latency of the command sent to the device and provides more responsive.

(Refer Slide Time: 2:41)



So, the concept of cloud edge computing is an effective way to manage and guarantee the quality of services and a lot of research is going on to address this requirement of cloud edge computing in order to meet the increasing our demand for service workloads. So, the cloud edge computing we will see that often depends upon the task offloading from the low or from the resource constrained devices they are called IoT devices which are deployed in most of these applications often require the connecting with the cloud earlier days.

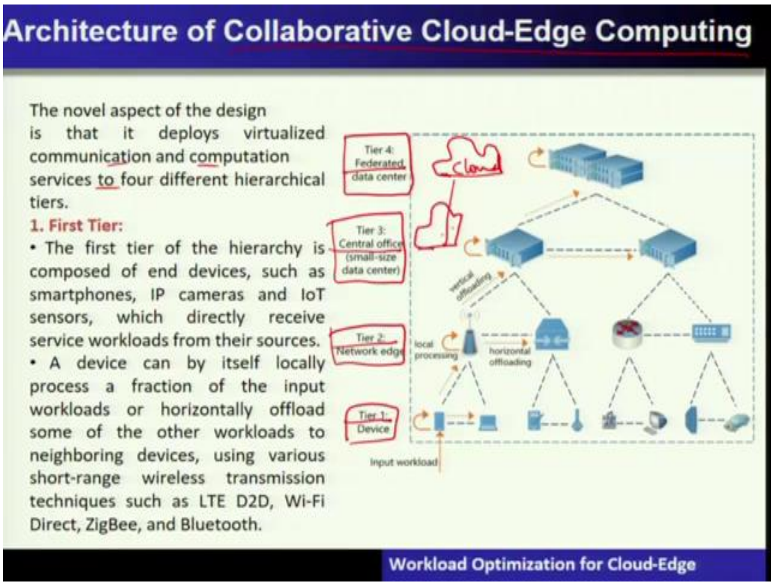
But now with the edge computing coming into the picture the edge and cloud together brings into the different concepts of offloading often called vertical and horizontal offloading between these service nodes which we will cover so let us understand about the vertical offloading. So, when you say that vertical offloading that means you have the cloud you have the edge and here you have an IoT device.

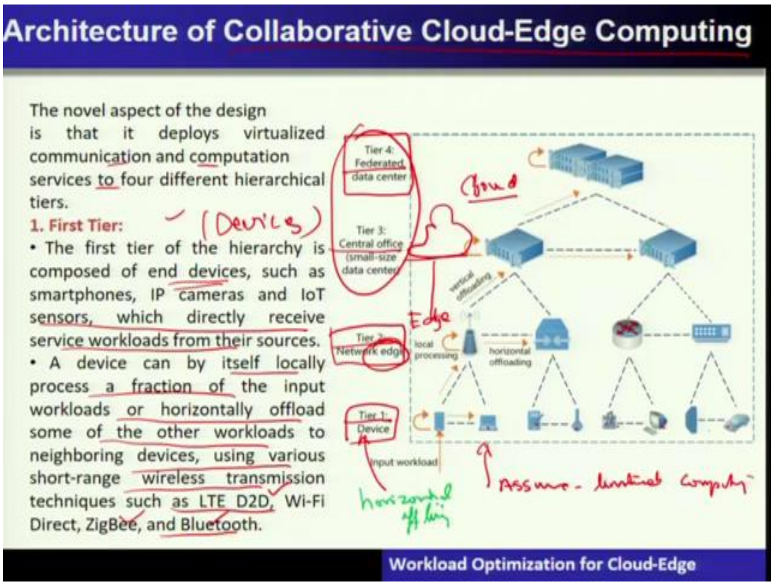
So, when you say vertical offloading so this means that the offloading from edge to the cloud so vertical offloading refers to the process of transferring the task or the services from the cloud or the data center to the edge nodes. So, this particular way of transferring the task from cloud to the edge so that means if let us say that processing if it is done on the edge not at the cloud it will going to save the time and thereby we call it as latency and thereby increasing the efficiency it is also known as cloud edge computing and is used to reduce the burden on the cloud. So, this is called vertical offloading.

Now, let us see about the horizontal offloading. Now, in horizontal offloading what you can see here is that on the other hand the process of transferring the task or services between the edge nodes so it is not having only one edge node but several edge nodes. So, the task is if it is offloaded across multiple edge nodes this is called horizontal offloading. So, horizontal off loading often considers the existence of multiple edge devices and this particular transfer that is task offloading across multiple edge devices called horizontal offloading.

So, horizontal offloading on the other hand is a process of transferring the task or the services between the edge nodes. So, in order to reduce the latency or increase the efficiency it is used to improve the capacity of the edge nodes and can be used to reduce the load on the cloud. So, therefore, these 2 types of offloading are very important in the development of edge computing and environment.

(Refer Slide Time: 5:46)



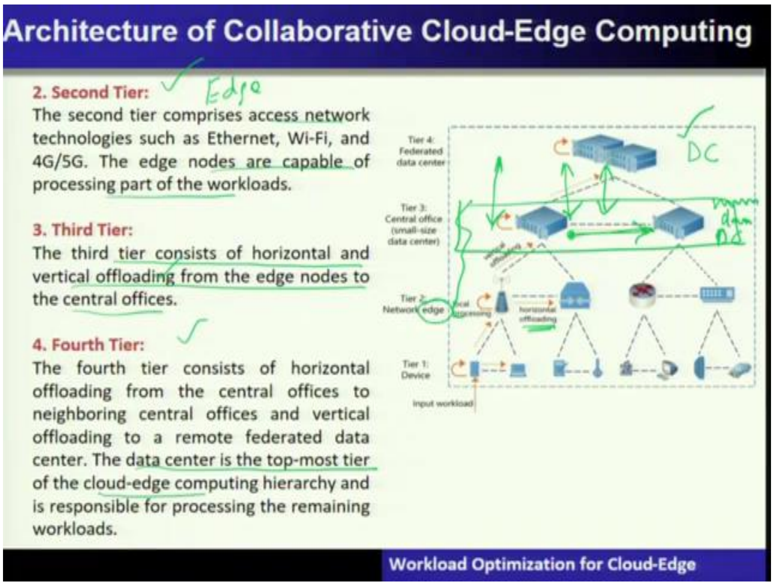


Let us understand through architecture of a collaborative edge cloud computing this particular concept. So, the novel aspect of the design is to deploy that it deploys the virtualized communication and computing services across 4 hierarchy tiers. So, they are that is tier 1 is nothing but the devices tier 2 is nothing but the network edge tier 3 is having the central office that is small sized data centers tier 4 is a federated data centers.

So, we can understand that federated data center is often a cloud and in this particular data center is a let us say that this is also a type of a cloud or you can say that it is a federated cloud and this is the cloud and this is the edge and this is an IoT device. So, this is an edge and this is these 2 are the cloud and this is so the first year definitely is about the devices. So, you can see that there are different devices and these devices are having we are assuming to have some kind of very limited computing capabilities at the device level.

So, the first tier is the hierarchy is composed of the end devices such as a smartphone IP camera IoT sensors which will directly receive the service workload from their source. So, the device can it itself locally process a fraction of the input workload or the horizontal offloading and some of the other workloads in the neighboring devices using short range communication let us say LTE device to device Wi Fi ZigBee Bluetooth. So, there is a possibility in the tier 1 devices also for horizontal offloading at tier 1.

(Refer Slide Time: 8:07)

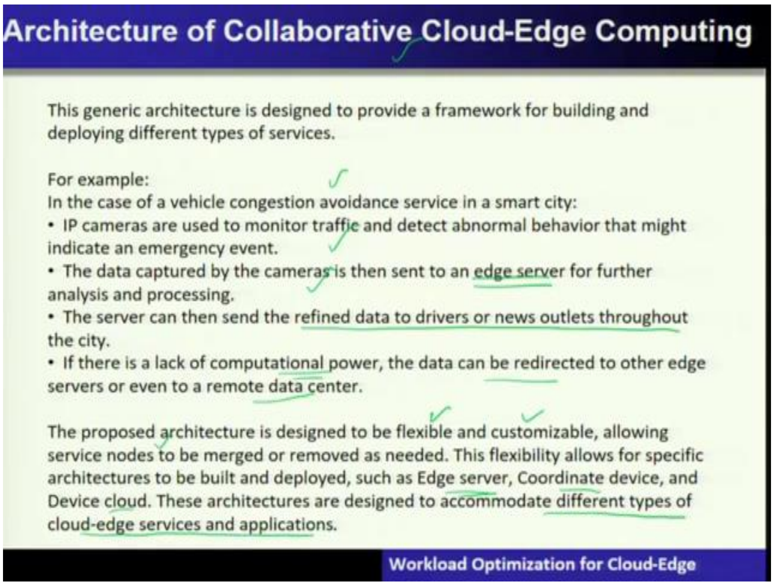


Now, then comes to the second tier which is called edge. So, the network edge or the edge nodes forms the second tier comprises the access network technologies and these edge nodes are capable of processing the part of the workloads. So, therefore horizontal offloading is also possible at the network edge or the edge nodes. So, network edge or the edge nodes. So, edge also contains little amount of compute or is a micro data centers 30 year consists of horizontal and vertical offloading from the edge node to the central offices.

So, this particular concepts of the edge nodes with a small data center called micro data centers micro DC and this is the data center of a cloud. So, this will have 2 types of this type of offloading that is between the edge node between the edge node and the cloud is called vertical offloading and between the edge nodes called horizontal offloading so it has the capabilities so these add node has both the capabilities whether to go for the horizontal offloading or to go for the vertical offloading with the cloud.

First tier consists of horizontal offloading from central offices to the neighboring central offices and vertical offloading to the remote federated so the data center is at the topmost tier data center there is a cloud a topmost tier of the cloud edge hierarchy and is responsible for processing the remaining workloads.

(Refer Slide Time: 9:40)

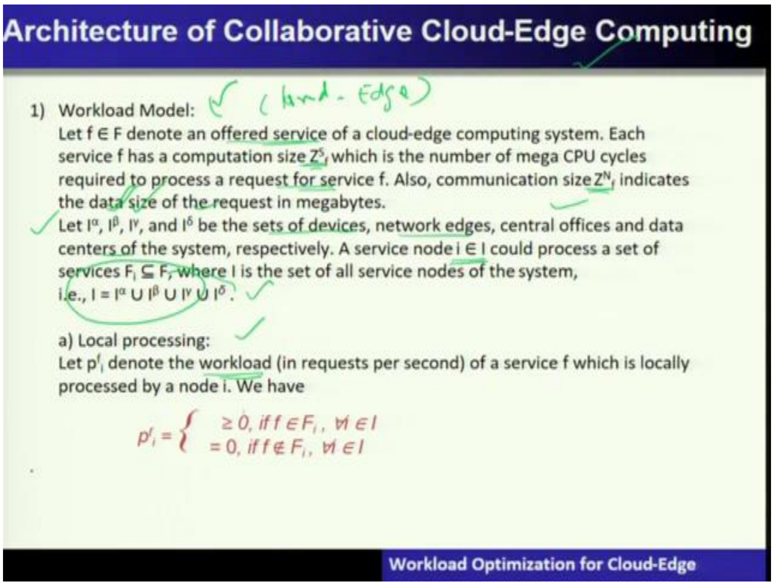


So the architecture for this collaborative cloud edge computing is shown over here. So, for example in the case of a vehicle congestion avoidance service in the smart city IP cameras are often used to monitor the traffic and detect abnormal behavior. So data captured by this camera is sent to the edge servers for further analysis and processing these servers can then send the refined data to the drivers or the news outlets and if there is a lack of computational power at the edge nodes then data can be redirected, to the other edge nodes or to the remote data centers.

Therefore, both horizontal and vertical offloading is to be done in an intelligent manner to solve the problem and of the application constraints of latency sensitivity and yet solve this particular application. So, therefore, this kind of architecture that is cloud edge and environment B is designed to be a flexible and customizable allowing the surface nodes to be mostly removed as the need arises and architectures are built and deployed as such as edge server coordinate devices and device clouds.

So, different type of cloud edge services and applications are now being made possible under this architecture of the collaborative cloud edge computing.

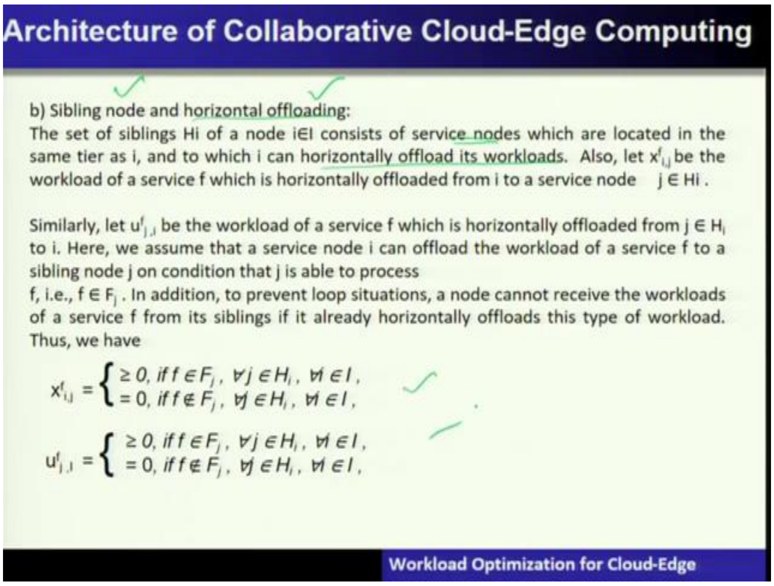
(Refer Slide Time: 10:54)



Let us see about the workload model. So, workload model let us say that f denotes the offered services of the cloud edge computing system and each service F has a computation size which is number of CPU cycles are required to process the request of the service F also the communication size and is needed the data size of the request is in the megabytes. So, let us assume that I alpha beta and gamma and delta be the set of devices network edge central offices and the data centers of the system responsible respectively.

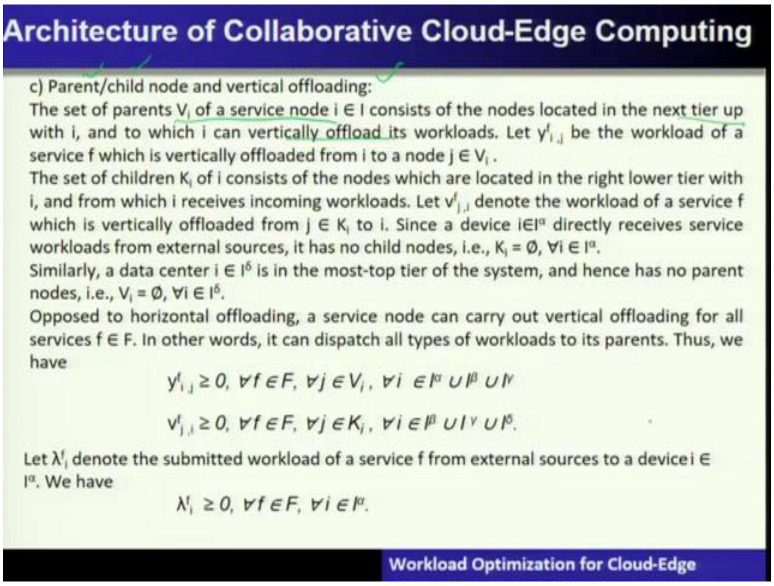
So, the service node that is I could process the set of services where is the set of all services of the nodes. So, this is all about specifying this cloud edge system using the capital I. Now that is called a workload model. So, we are assuming that collaborative cloud edge computing will give workload model. As far as the local processing is concerned define the workload in the requests of a service where F is locally processed by the node.

(Refer Slide Time: 12:05)



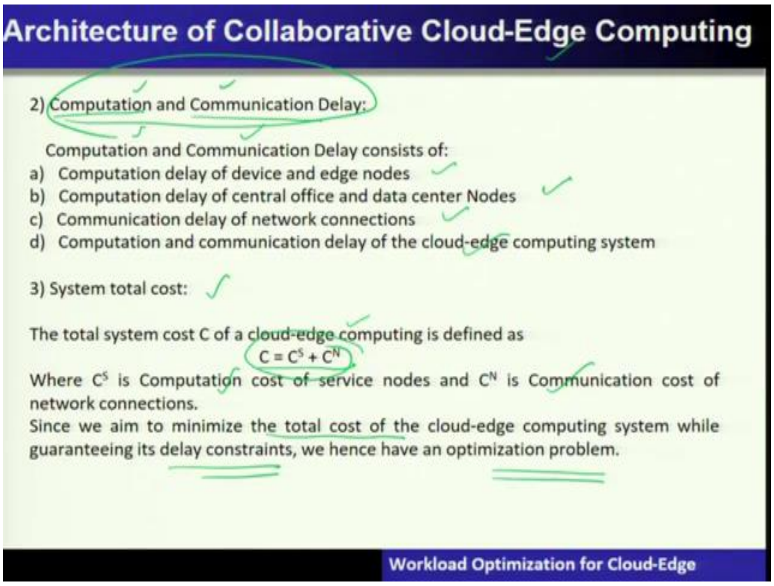
And we have this particular sibling. sibling node and the horizontal offloading. So, the setup sibling Hi of a node consists of the service nodes which is located in the same tier. So siblings and horizontal offloading whenever you are considering the horizontal offloading we are assuming that there exist a sibling nodes and these service nodes that is in the form of sibling allows the horizontal offloading which are located in the same tier can horizontally offload it is workloads and often this is represented by these equations.

(Refer Slide Time: 12:36)



Now, as far as if there is a parent child nodes are there that means the cloud and edge and cloud that is apparent it will become the cloud and edge node is a child node then there is a possibility of the vertical offloading. So, the set of parent nodes of the service consists of the nodes located to the next tier up with I and where I can vertically offload it is workloads.

(Refer Slide Time: 13:02)

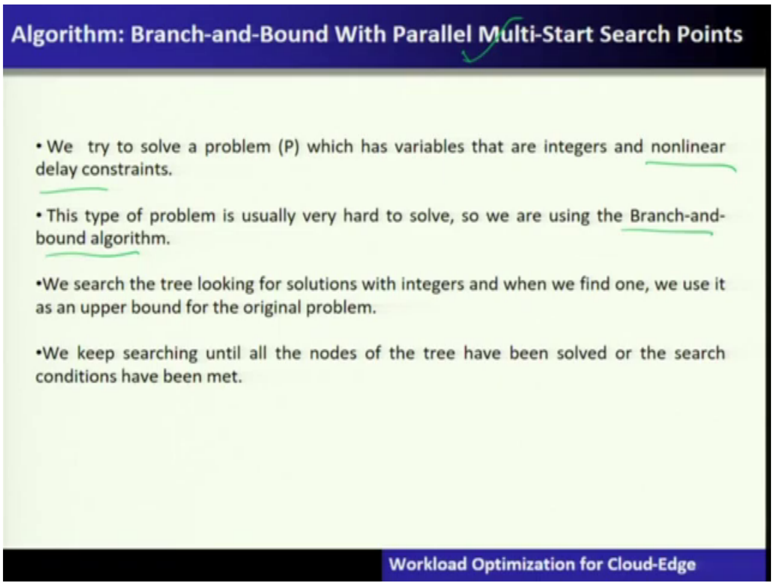


So, they are all represented here in this particular equation. Now, let us see that in this cloud edge computing model there are 2 types of delays one is called computational delay the computation delay means that whether the compute has to be done whether in tier 1, tier 2, tier 3, tier 4, so wherever it is done there is an amount of computational delay similarly to roof the oriental offloading or vertical offloading there also exist communication delay into this part.

So, computational and communicational delay consists of computational delays of the devices and the edge nodes are doing the computation computational delay of the central office and data center and computational and communication delay of the network connections and communication and computation communication delay of edge cloud system.

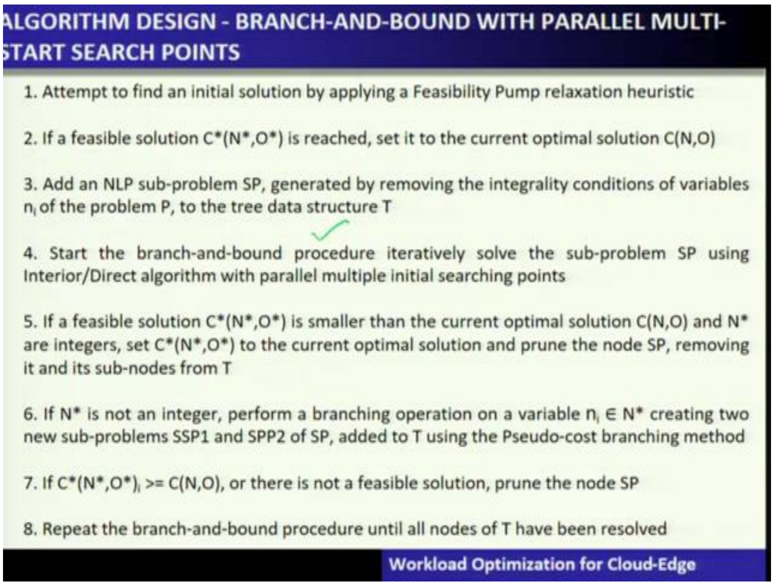
And let us see the total cost total cost of the edge cloud computing is shown over here using C S is the computation cost of the service nodes and C n is the communication cost of the network connection. Since here the aim is to minimize the total cost which is nothing but the communication and computation cost of the cloud ad system. While guaranteeing these particular delay constants, we have an optimization problem setting in this particular way.

(Refer Slide Time: 14:18)



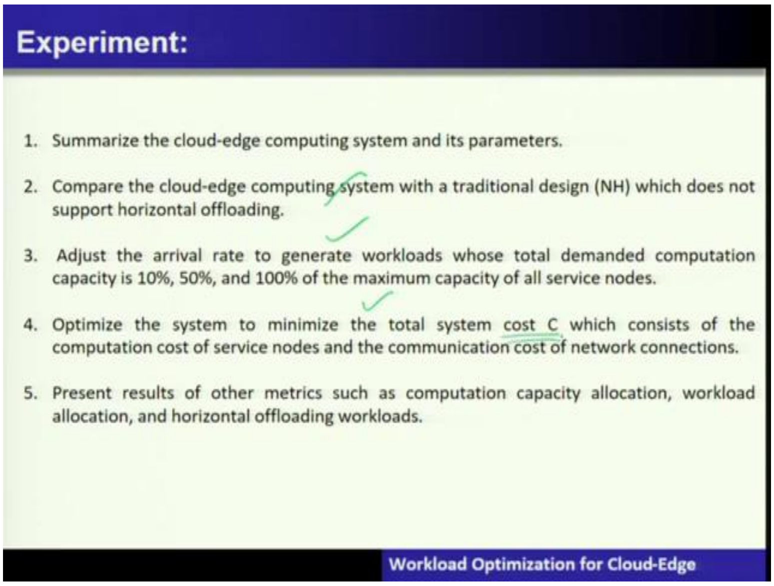
Now, there are various techniques to solve this kind of problem without the machine learning this kind of problem often formulated as the branch and bound problem and these problems are very hard in computer science in PC branch and bound problems are often hard and turns out to be the integer and nonlinear optimization problems without going in more detail about these branch and bound problems.

(Refer Slide Time: 14:44)



We will see about how these particular problems can be solved. So, here we can see that these problems are modeled or they are equal into the branch and bound procedures. And therefore a lot of algorithms are possible to be designed in this approximation algorithm domain.

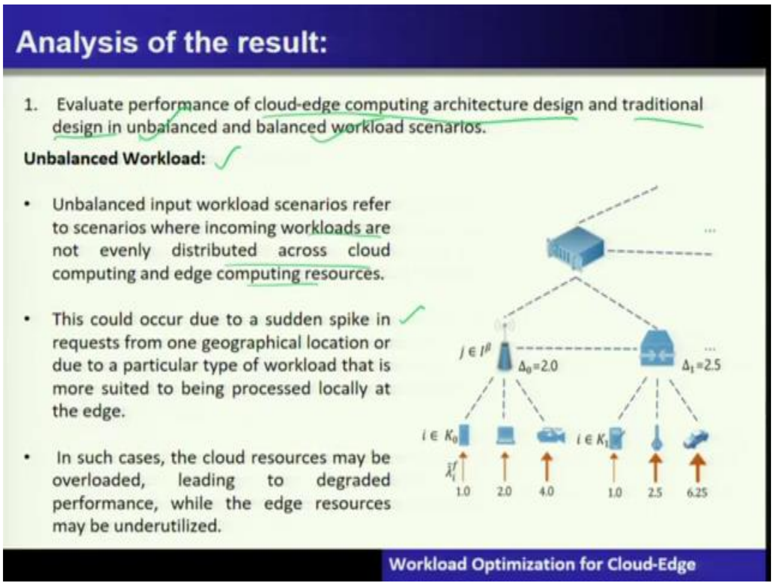
(Refer Slide Time: 15:01)



So, let us see that how we are going to conduct the experiments to find out this cost functions that is a computational and communication cost function under this cloud edge computing model. So, if you want to compare the different strategies whether oriental offloading or vertical offloading to solve horizontal offloading vertical offloading for service nodes.

So, there is a possibility of comparing different strategies and then finding out which one will be the best one. So, therefore, this requires a different level of a different type of configuring the experiments and to measure the performance using the experimental evaluation of these methods and how to for optimization to minimize the total system cost here in this particular manner.

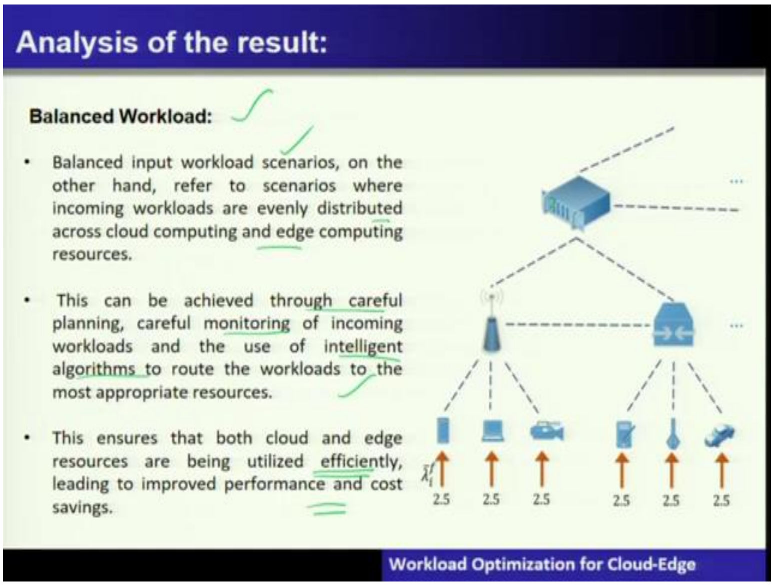
(Refer Slide Time: 15:48)



So, if you see the analysis broad way without going in formulating the problem as a branch and bound problem you will find that there are 2 important observations you will see in that cloud edge computing architecture and if you compare with a traditional design so traditional design often suffers from unbalanced workload situations. So, the traditional approach is offering leaves to the unbalanced workload situation says that some of the servers are heavily loaded whereas other servers are not that much loaded.

So, unbalanced workload scenarios often refer to the scenarios where incoming workloads are not evenly distributed across the cloud and edge computing resources. Another hand if let us say you go by a generic heuristic which is called a balanced workload scenario so this will overcome from any problems of sudden spike and all these problems which will not be able to scale the problem or without overloading the problem at the servers.

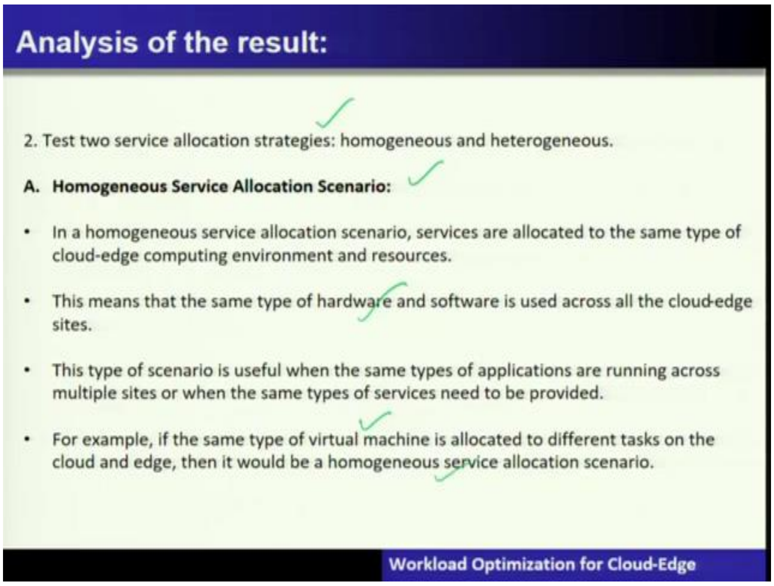
(Refer Slide Time: 16:52)



So, this particular concept will have with a balanced workload so balance input workload scenarios on the other hand refer to the scenarios where incoming loads are evenly distributed across the cloud and edge resources and this will eventually carefully planning monitoring of incoming workloads and some intel there requires intelligent algorithms to be designed.

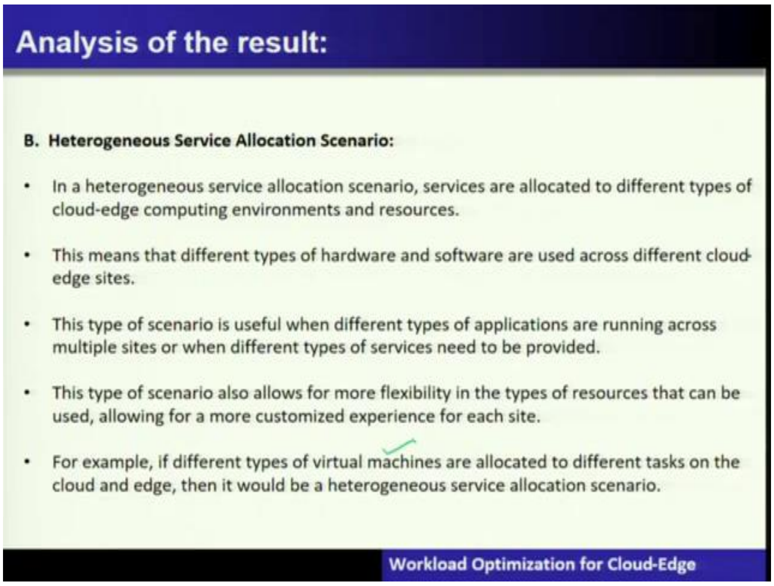
Here we are not discussing but there is a possibility that to achieve this balanced workload some machine learning or intelligence computation has to be there. So, to achieve this so this ensures that both cloud edge resources are fully utilized efficiently. if let us say you maintain the balance workload situations and the cost savings here in this case.

(Refer Slide Time: 17:33)



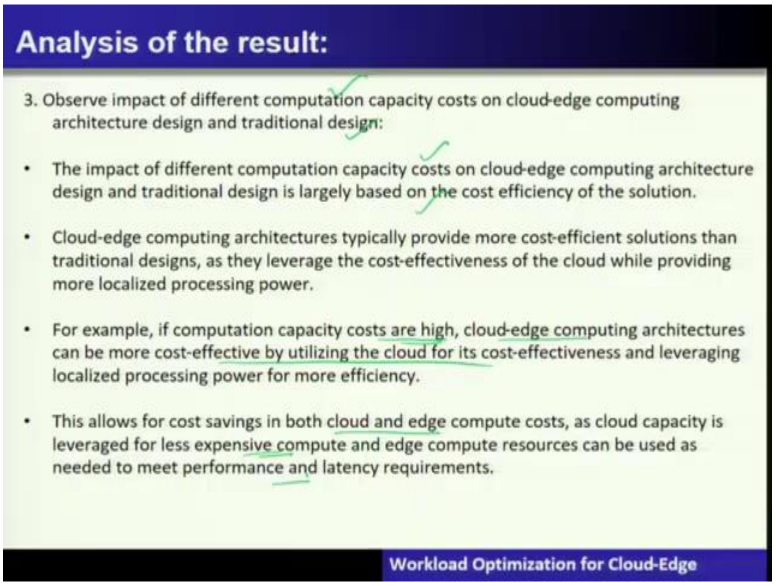
So, there are 2 service allocation strategies homogeneous and heterogeneous homogeneous service allocation strategy if you see that services are allocated to the same type of environment that is in the cloud edge computing environment are the resources this means that the type of hardware and software used almost same type is often used in running. So therefore if the same type of virtual machines are allocated to different tasks then it will be homogeneous type of services.

(Refer Slide Time: 17:58)



Whereas, heterogeneous type of services are not having this kind of similarity and therefore different virtual machines are allocated to the different task on cloud edge then it will be heterogeneous type of services.

(Refer Slide Time: 18:10)



So, we have to observe that the impact of different computational capacity cost on the cloud edge computing what says the traditional design approach is rarely studied here in this particular discussion of this lecture. So, the impact of different computational capacity cost of cloud edge computing architecture and traditional is largely based on the cost efficiency of the solutions.

For example, if the computational capacities are high the cloud edge computing architecture can be more cost effective by utilizing the cloud for it is cost effective and leveraging the localized processing power. This allows the cost saving in both the cloud and edge compute cost and cloud capacity is leveraged for the less expensive compute and edge compute resources can be used as needed to meet the performance and latency requirements. Thank you.