Cloud Computing
Prof. Soumya Kanti Ghosh
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 32
Mobile Cloud Computing – II

Hello. We will continue our discussion on Cloud Computing; rather we will continue our discussion on Mobile Cloud Computing. So, what we are discussing about mobile cloud computing said that the mobile is a mobile devices these days are omnipresent right. There is a huge glorification of smart mobile devices which has fairly good capability of doing lot of a applications or running lot of applications. As there is a as the capability increases, the need for running more computational and data intensive applications also there is a increase in the those type of applications.

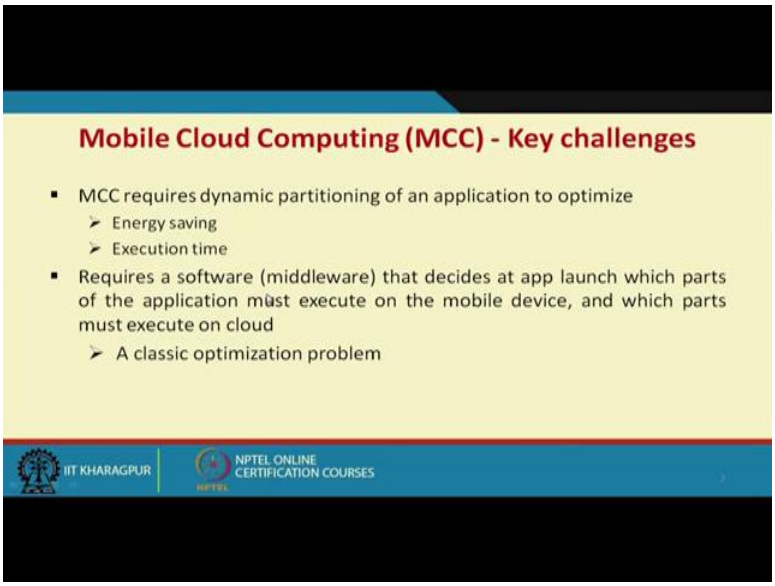
Now, in though there is a increase in this resources of the mobile devices, but what we still feel that some of the cases had it been a background, backend, some more computing services that could have been better like we execute some part here, some part offload, some part of the computation or data on the backend devices where what we have seen that cloud may play a important role. But there are several issues into the things right that any transfer of data and processes involves costing. Costing in terms of execution time specifically that if the delay is more than something, it is it becomes unviable also we see that is a question of reliability when once your application runs goes to another place to run.

So, if there is there are a lot of dependency involves like intermediate network or the wireless network, the availability of the resources at the other end or failure or fault in the other end and of course, the fault in the device at the a mobile device itself. So, whereas, when you offload all those things come into play right then we have also seen there are issues of security right. So, far the application running your own device, you have someone on the process and the data over it right, but whenever the application running in some other premises or some other devices or cloud, which is maintained and owned by some third party, then your there is a concerned about privacy and other security issues the regarding those data and a processes.

So, this becomes pretty tricky issue and nevertheless mobile cloud computing is becoming pretty popular because of several reasons. First of all you can now have think of a large application running on or rather controlled by a small devices, there can be things that the your process may be using somebody elses data. So, you run somewhere where the data is there and the process is there or the mobile device acts as a collection of data which is contributing to some central repository, some process at this device or some other device can do. So, there is there are issues of in exchanging of data interoperability issue and sharing of this sort of devices right.

Or what we see a evolution of this that internet of devices come into play where cloud plays as a central role of managing the so right. So, we will look at a some of the aspects today regarding a mobile cloud computing in continuous in our with our previous lecture of mobile cloud computing, and rather we will try to refer some of the research works which will give us a opportunity or give us a more better view of how what this what are the different challenges and how to handle them and so and so forth.

(Refer Slide Time: 04:24)



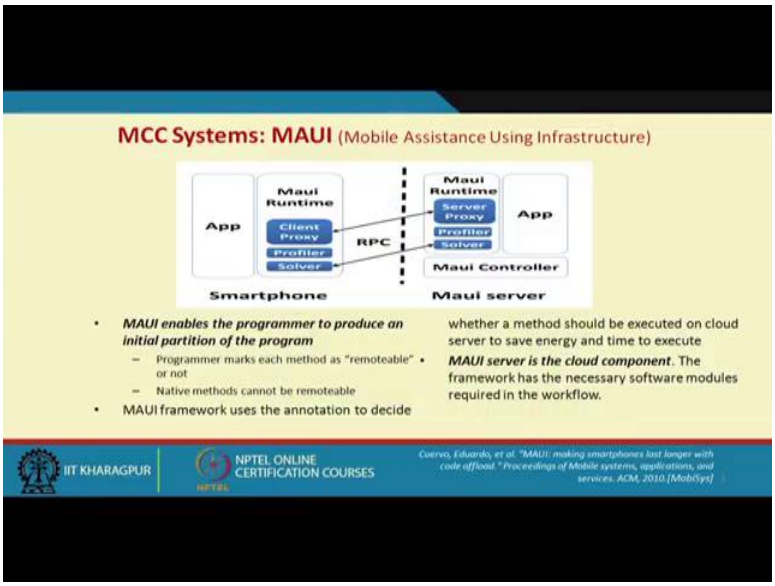
So, we will look at a mobile cloud computing, as we have seen some of the key challenges are that that requires dynamic partition of a application to optimize energy saving execution side right. So, two things as we mentioned earlier also, one is energy saving and to optimize the execution time right. If the time is beyond a limit then there may be that then that process may not be viable to run on the things also requires

software or something middleware that decides that the app launch which parts of the application will be executed locally and which part of the application need to be executed on the externally or remotely or in the cloud right.

So, it is a classic optimization problem right. So, we can do this sort of means dividing this application into different parts, some of the parts or the component of the application need to be run locally and some of the remotely, this can be done statically like I have a priori knowledge of this sort of things can be done or in some cases I may not know the a priori that which need to be locally or locally or remotely, primarily it is if it is dependent on the data and other type of things right. So, what will happen I need to have a dynamic these and all the things.

And as we understand whenever we need this type of thing synchronize a sense scheduling and the resource management come into play in a big way right. So, all those things here and when you are specifically doing the devices and say infrastructure, which are owned and maintained by different authorities then it is a more tricky. So, like mobile devices is something and cloud is maintained by other third party, intermediate you have service provider we are who are again have a different type of authority. So, all those things makes a very complicated issue need to be solved.

(Refer Slide Time: 06:36)



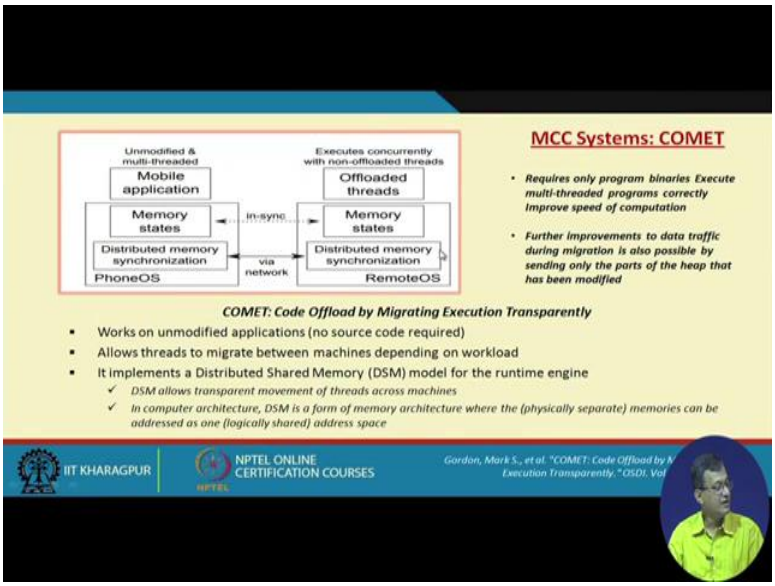
So, this is one of the work is as referred in the in this paper, there is mobile assistance using infrastructure right. So, it is one form of mcc. So, MAUI enables the programmer

to produce an initial partition of the program. So, programmer marks the each method as remoteable or not that means, whether a remote or local, native methods cannot be remoteable which are need to be done on the on the device itself MAU framework uses the annotate to decide whether the method should be executed on cloud server to save energy and time to execute the thing right.

So, there are a couple of times the time to execute on the cloud there is a time to transfer this process and data to the cloud and get back the data if as if the need demands and MAU server is a cloud component. So, there is a MAU client which is on the mobile side there is a MAU server which is a cloud component, the framework has necessary software modules required for the workflow. So, that there is a client server type of architecture where what they propose and can be able to synchronize.

So, if you see; there is the app and there is a profiler and the solver profiler basically profiling of the application and profile along with the server partitions the application these gates executed here, and with RPC calls and then it in a two and four things Lan server communicate to the profile as sync among itself when the apps run in a some part on the app and some part on the some part on the device and some or what we say smart mobile divider, SMD and some part on the cloud right.

(Refer Slide Time: 08:27)

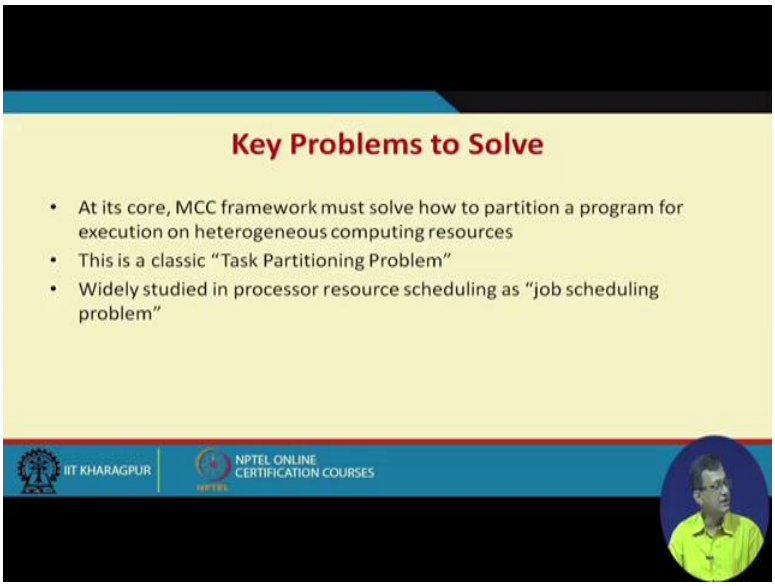


So, there is another work which is what is called comet quad, code offloading by migrating execution transparently right. So, works on unmodified application no source

code required that is on the app itself. Allows threads to migrate between the machine depending on workload, they based on the things that threads can migrate, it implement a distributed memory shared memory or DSM model to runtime engine right those who have gone through computer architecture or advanced computer architecture and various type of things, they understand that this DSM model. A DSM allows transparently movement of the threads across machine. In computer architecture DSM is a form of memory architecture where physically separate memories can address as one logically shared address space, there is the core of the or there is what we say very quick view on the what this DSM says.

So, likewise there is a phone OS and remote OS. So, it need to maintain this memory states and there is a distributed memory synchronization and the executed this processes other codes are executed or offloaded by migrating execution transparently. So, it is not that the neither the source code is required nor that has nor the a user end there is known that this things are going on at the backend. So, what it requires only program binaries right execute multi thread programs correctly. So, it does not require source code only the binaries; furthermore improvement on the data traffic during migration is also possible by sending only parts of the heap that has been modified.

(Refer Slide Time: 10:26)

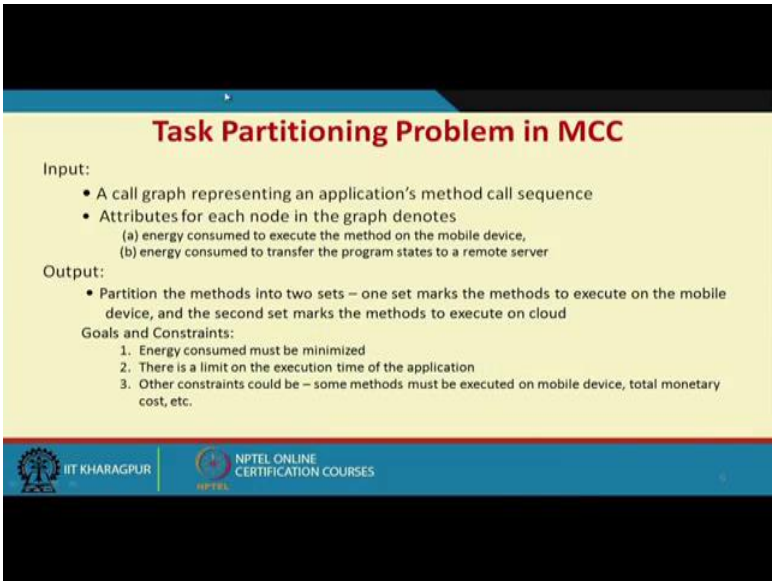


So, it is not transmitting the whole heap, but the part of the heap which has been modified. So, what we see the key problem to address or solve is a at the at its core MCC

framework, must solve how to partition a program right or execution on heterogeneous sources. Thus not only partitioning, it need to be executed on heterogeneous resources right it one on the mobile device side which may be some mobile related OS when you are running on the cloud side that is some other kind of environment. So, it is a heterogeneous environment how things need to be run need to be looked into.

So, it is a task partitioning problem as we have seen in other computer architecture and organizations, widely studied in processor scheduling as a job scheduling problem also right we do have this sort of challenges out here. So, need to be looked into from that angle. So, task partitioning problems. So, what we have typically in a task partitioning problem in mcc a call graph representing the applications method or call sequence right.

(Refer Slide Time: 11:22)



So, it is a call graph, attributes of each node in the graph in denotes a energy consumed to execute the method on the mobile device and energy consumed to transfer the program states to the remote server. So, one is the energy consumed in the mobile device and transferred the things.

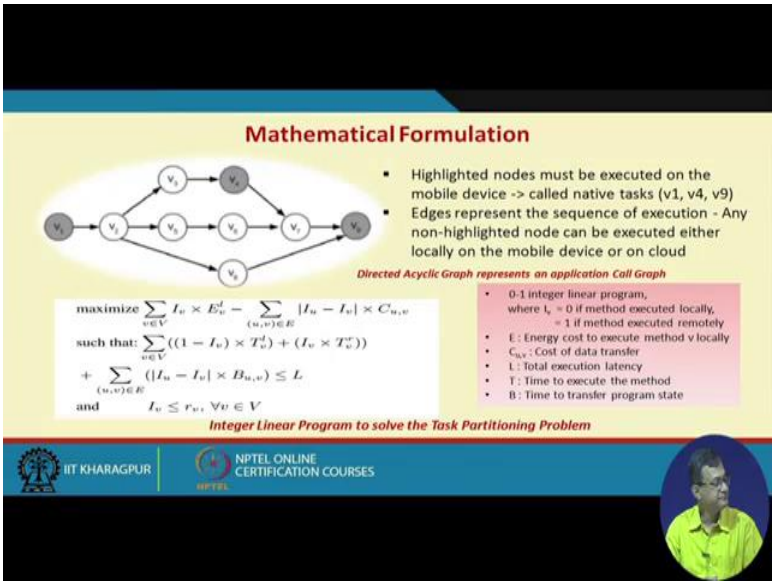
What we are considering that the other end that is the cloud end, the it is pretty resourceful and efficient. So, the energy consumes in is in executing that in the cloud end is negligible in compared to the energy consumption at this end at the device end. Rather the device does not much clear about that, but you need to clear about that executing its own locally the portion to be executed, and the transfer cost of this the same memory

state and processes and so and so forth as things required right and data memory state and so on. So, output what we get as a partition the methods into two sets one set of the matter executes at the mobile device, and the second state set it goes to the execute on the cloud.

So, go if we look at the goal and constraint, energy consume must be minimized. So, I need to minimize the energy consumption, there is a limit on the execution time of the application. So, it should within the threshold say time capital T, other constant could be some methods must executed on the mobile device total monetary cost etcetera. So, two we have couple of a constant. That means, we need to goal is to reduce the energy consumption, time should be within that particular threshold value. And there can be other constant like some of the application some part of the application may need to be executed in the device itself like I may not be able to offload it.

And I may look at the total costing of the thing the total cost should not execute some something some portion more than that, so these things are necessary.

(Refer Slide Time: 13:54)



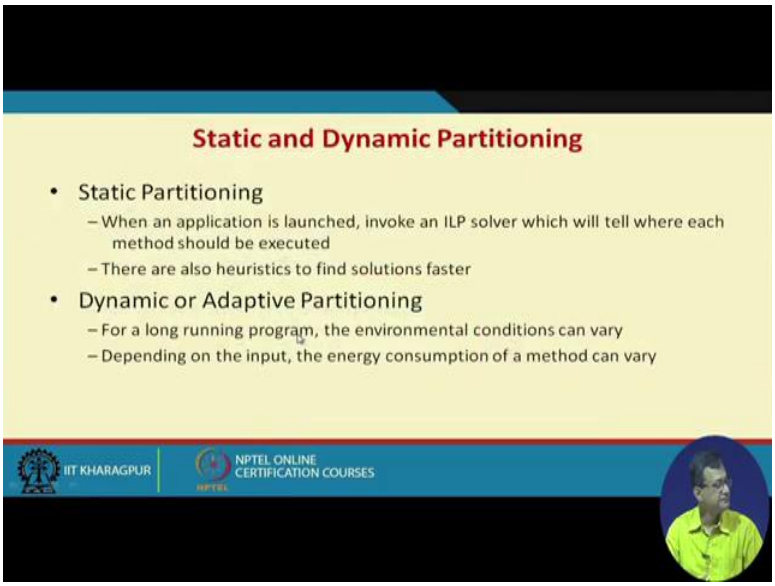
So, if we look at that little mathematically. So, we have a execution graph right out of that what we say that these are v1, v4, v9 in this case are native task; that means, needs to be executed in the device itself right I cannot offload it. So, there can be so, highlighted nodes here must be executed device, edges represent the sequence of execution that how the sequence of execution will be there. Any non highlighted node

can be executed either locally or on the mobile device right can be locally or on the mobile device.

So, if it is offloaded to the local device, this first four senses that how much energy it could have saved right or in other sense had it been worked on the local how much energy could have been taken. So, that has been saved if I offload it to the remote, but I need to also look at that where the costing of transfer this data right. So, there is a cost involved or there is a energy cost involved energy expenditure involved in transferring the data from the device to the remote device right that has to be incurred by the mobile device right. So, that are needs to be there.

So, if we need to maximize this. So, if it is if this result is a positive or the maximum of the maximum if you can maximize, then I have a substantial energy saved otherwise cannot be there the there is no benefit of doing that right and of course, it should be within a total execution latency should be less than L right we should not have this whatever we do need to be less than the particular latency L right. So, that it should be within that particular things and I need to look at that a dependency also like execution of one need to be executed, if there is a dependency in the graph that need to be executed out here.

(Refer Slide Time: 16:13)

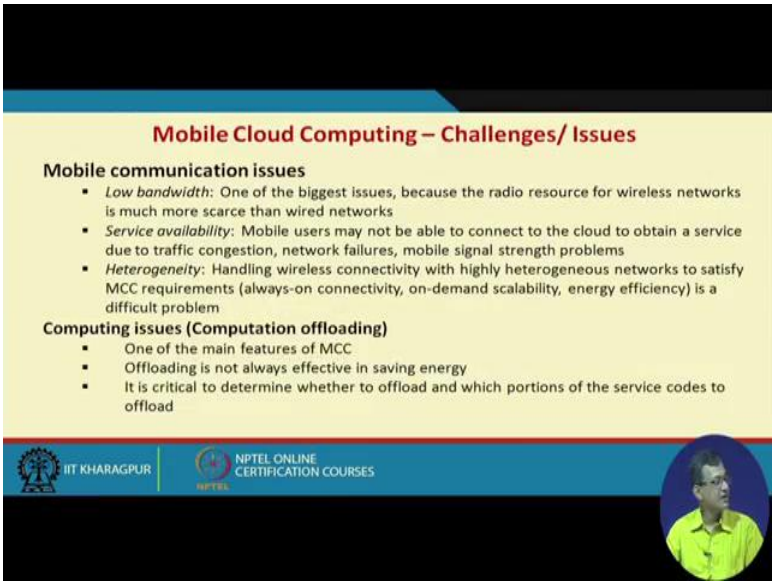


So, as we are discussing there are two type of partitioning one is static and dynamic as we have seen. Static in case of a static when the application is launched the invoke a ILP

solver and which tells where each method should be executed. So, it says that statically says that when it is launched. So, there is there are also heuristic to find a solution first term that there are we can deploy. In case of a dynamic or adaptive partitioning, what we do for long running program environmental conditions may vary right if it is a executing for a long time environmental condition may vary like a requirement of the data and even the availability of the remote resources transferring bandwidth all those things can vary.

Depending on the input the energy consumption of a method may vary so; that means, in other sense I need to take a call that which partition need to be executed locally and remotely because of that varying nature of the things there are hosting may go on changing right. So, that is another challenge out here.

(Refer Slide Time: 17:26)



So, there are as we see there are when we do this type of partitioning and transfer the job executes something locally, something remotely or everything within a time limit, a particular threshold values, threshold time and we want to maximize the energy saving right.

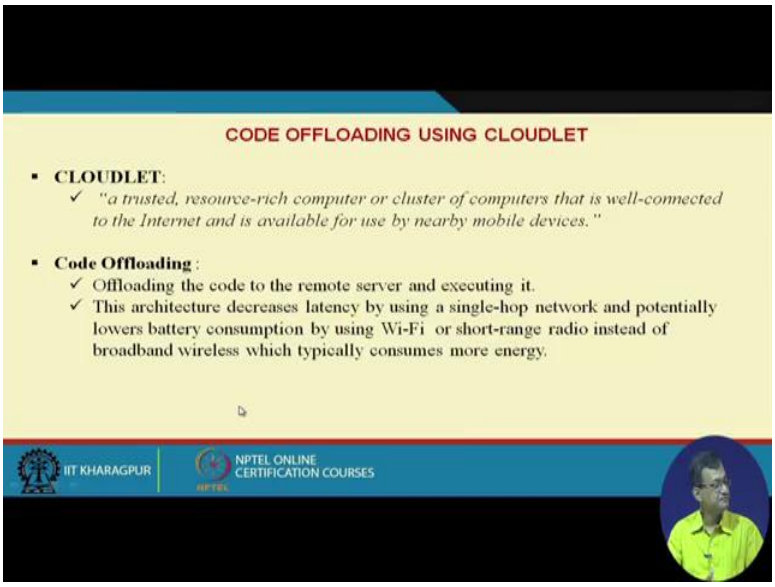
So, these are the different constraint and in order to achieve this sort of issues. So, there are several challenges and issues need to be addressed. So, mobile communication issues that is low bandwidth one of the biggest issue, because of the radio resource of the wireless network is much more scarce than the wired networks. So, the it is the availability band availability of this bandwidth mobile bandwidth or wireless bandwidth

at times is pretty less in compared to a wired type of things or. So, that is not only the less it also varies on if it is a long running execute executive means long running process then it may vary over time.

So, getting a guaranteed bandwidth is a challenge service availability mobile user may not be able to connect to the cloud to obtain a service due to traffic congestion network failure mobile signal strength and so and so forth. Issue of a heterogeneity handling wireless connectivity with highly heterogeneous network to satisfy mcc requirement always on a connectivity on demand scalability, energy efficiency is a difficult problem. So, there are heterogeneous heterogeneity involved in it and there are these are different challenges. So, these are more issues with the communication issues, there are a few computation issues or computing offloading related issues; one of the this is one of the major feature of the things like you are able to offload it.

Offloading it not always effective in saving energy as we have seen that if those parameters are not made then it may not be that energy saving, may not be what we say within that quote unquote profitable margin right. So, it may be costly stuff; it is critical to determine whether to offload and which portion of the source code to offload. So, its it is a critical challenge, especially as that there are different type of application with different type of requirement etcetera and its it is a major challenge to decide that where when to offload, how much to offload and so on and so forth.

(Refer Slide Time: 20:11)

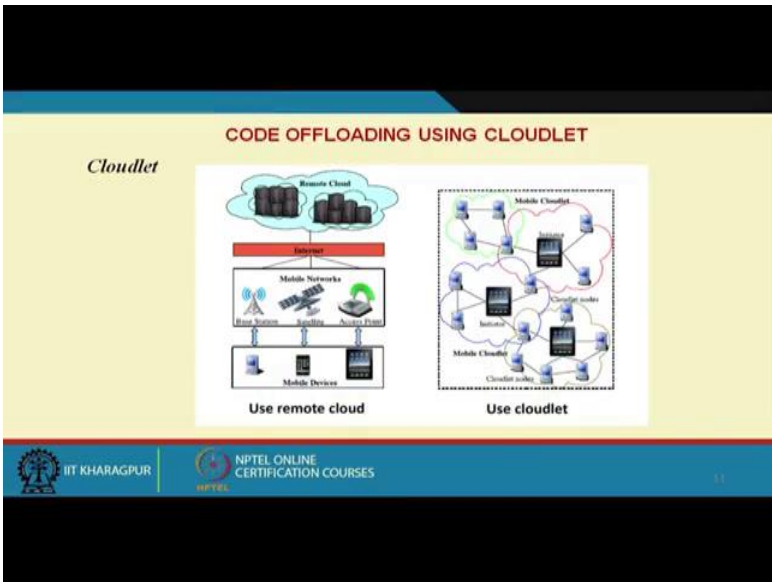


So, there are other trials or other type of a initiative one is that code offloading using cloudlet right. So, if you look at cloudlet, it is a trusted resource rich computer or cluster computers that is well connected to the internet and is available for use by nearby mobile devices. So, it is a what we say we can look at as a smaller version or of cloud, but it is resource rich, but it is more within the locality of the mobile devices or nearby regions right.

So, offloading a code to the remote server and executing it, when you look at the court code offloading part. So, this architecture decreases latency by using single hop network and potentially lowers battery consumption using Wi-Fi, short range radio broadband wireless etcetera. So, what we try to see that first of all of offloading and secondly, when we make multiple hops then the cost increases. So, what you try to see that whether I can have presence of smaller clouds, which can take care of this sort of application, this sort of a executing this portion of the application.

And if a single hop then we can have something more or what we say that we can have more the better energy saving and better response time, with having those resources much closer to the device.

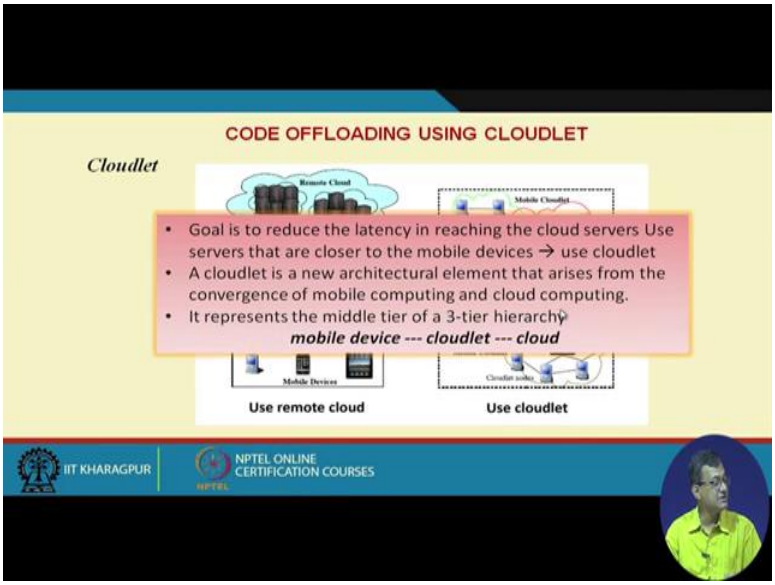
(Refer Slide Time: 21:53)



So, cloud code offloading using cloudlet. So, we have a mobile devices, mobile networks and the remote cloud and there are if you look at there are different small small different set of systems or mobile cloudlet, and cloudlet nodes which takes care of these devices

right or what we say that goal is to reduce the latency, in reaching the cloud service servers use a servers that are closer to the mobile devices.

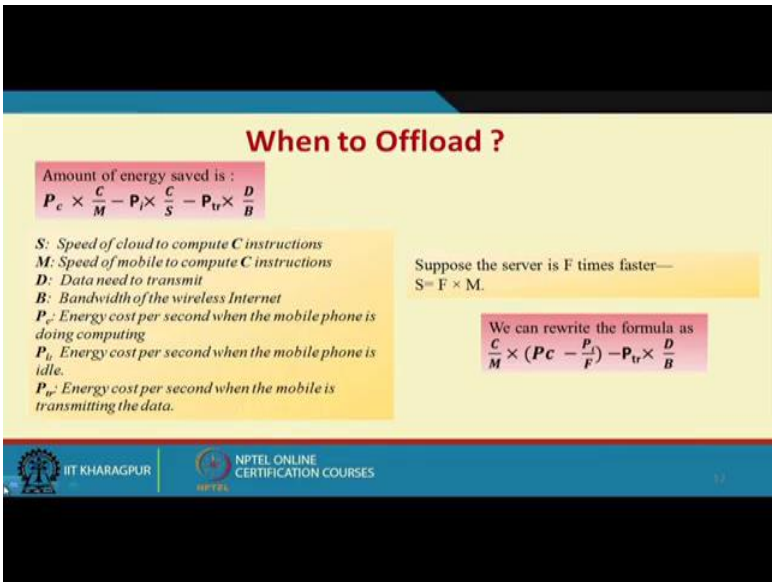
(Refer Slide Time: 22:21)



That means going for instead of the full place cloud, we can have smaller version of the clouds or cloudlet. A cloudlet is a new architectural element or that arises from the convergence of mobile computing and cloud computing right.

It represents a middle tire of a 3 tier architecture like mobile devices cloudlet and cloud. So, that in other sense what we are trying to do in order to reduce the latency, and a better energy savings possible better energy saving in one side the cloud is there, one side a devices we made a we make a intermediate a thing what we say cloudlet right; so which takes care of the things. Now when to offload is again a serious question.

(Refer Slide Time: 23:17)



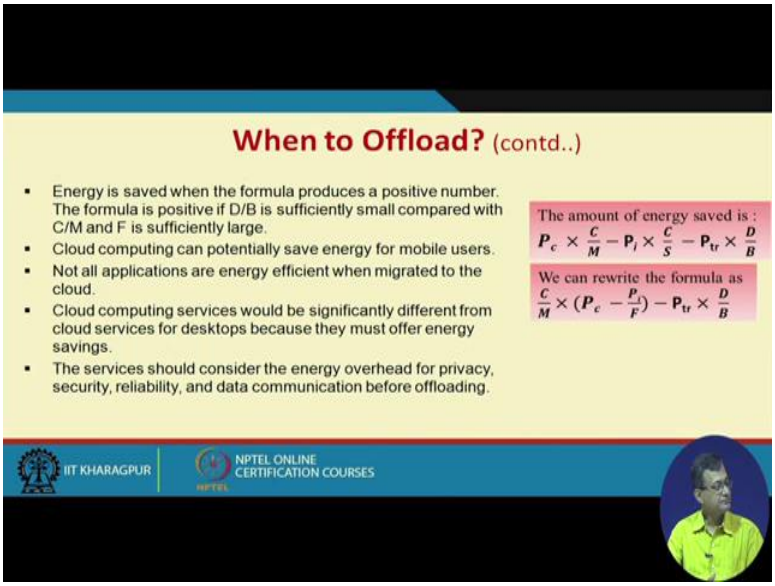
So, if we look at a very straightforward formulation of the things. So, if we look at that equation that. So, what we say that P_c is the energy cost per second when the mobile phone is doing computing; that means, mobile phone is computing P_i is that energy cause the mobile phone is idle.

P_{tr} is the per second a cost per second when the mobile phone is transmitting data. So, a it a mobile device can be either computing or idle or transmitting data. So, this three component are P_c, P_i or P_{tr} or when energy when transmitting data. Now if you consider c as the number of instruction, and m is the speed of the mobile compute the c instruction. So, what we see that if we see that $P_c \times \frac{C}{M}$ if it is offloaded $P_c \times \frac{C}{M} - P_i \frac{C}{S}$. So, what is s the speed of cloud to compute the C instruction.

So, if it is s is the p Ps s is the speed of the cloud then I we can see that idle P_i is the idle time, C there is number of instruction by the speed. So, this P by C by S and the transmitting P trd by B D means the total data to transmit and B is the bandwidth available. So, if we offload. So, this is my overall Ballpark figure right this should be a positive number. So, it says that I save energy right. So, this was my energy consumption if I offload this consumption are not there and then then this is this is save of the saving of the energy or if I say the server or the cloud server is f time faster, then I can look at $S = F \times M$ right.

So; that means, your speed of cloud to compute C instruction and M is the speed of mobile phone to compute C instruction if it is f times faster, and we can see that S equal to F into M right or in other we can rewrite the program or rewrite that particular formula or substitute this s equal to F by M into the formula to look to basically arrive this thing $\frac{C}{M}\times\left(P_c-\frac{P_t}{F}\right)-P_{tr}\times\frac{D}{B}$. This is a very simple arithmetic you just substitute and if you take appropriate common. So, we come to a figure called in this sort of scenario right $\frac{C}{M}=\left(P_c-\frac{P_i}{F}\right)-P_{tr}$.

(Refer Slide Time: 26:40)



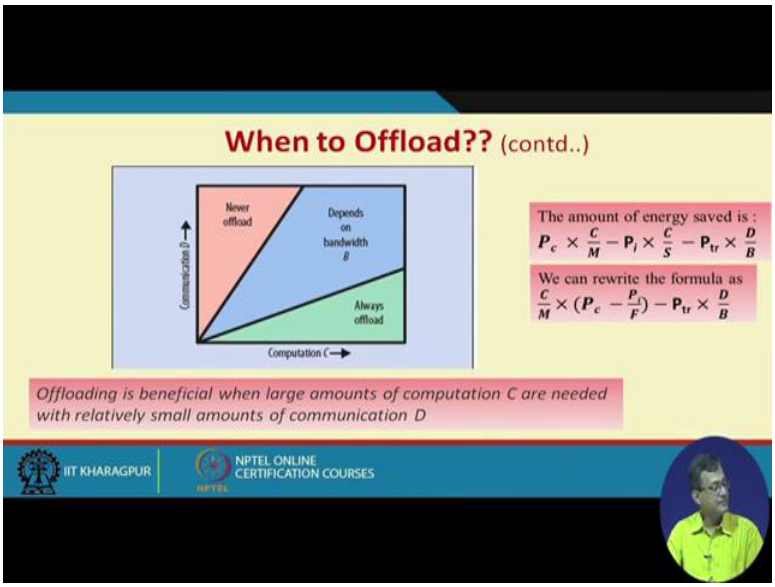
So, what we see from here that energy is saved when formula produces a positive number definitely if it is a positive then only energy is saved the formula is positive if D by A, D by B is sufficiently small compared to C by M right. So, one side is D by B should be sufficiently small compared to C by M. So, that I have a larger number on the left side. So, this is what we say that this is one of the condition cloud computing. So, what cloud computing can potentially save energy from the mobile devices, note that all application in the energy efficient when migrated to the cloud are may not be energy efficient right.

So, from this formula already say that first of all D by B should be must lesser less than C by M and F should be sufficiently higher which makes sense right. F is the how many times the cloud server is faster than the mobile device; it is likely that the cloud server

should be sufficiently higher otherwise there is no point in offloading and getting executed at externally right if it is you know. So, if that conditions are satisfied then what we see that we get a positive number, and it is it is obvious that all the applications may not be always beneficial when you offload, cloud computing service should be significantly different from the cloud services for desktop because they must offer energy savings in this case.

The services should consider the energy overhead for privacy, security, reliability, data communication etcetera which we have not directly considered in these cases.

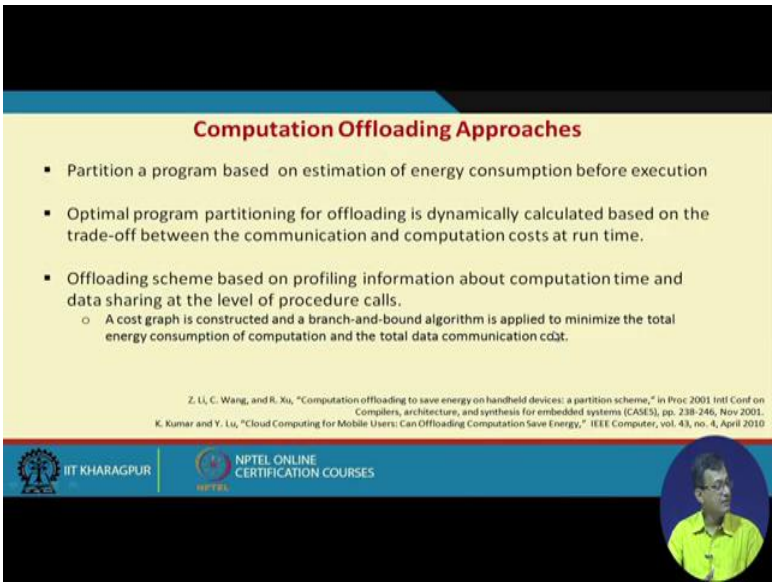
(Refer Slide Time: 28:37)



So, it is a ideally it should take care of other type of features which in while calculating the things. So, if we look into different other way of looking at us same a formulae formula is that offloading is beneficial when large amount of computation C are needed with relatively small amount of data communication right.

So, we require large amount of computation right with relatively less transmission of data right. Here also as we have seen that C should be sufficiently large with D should be sufficiently small. Keeping like F must be much faster and do we see that it is a thinks and we can see that in different scenario how things will be there right.

(Refer Slide Time: 29:24)



So, party computation overloading there are several approaches, you can see several research papers also partitioning a program based on estimation of energy consumption before execution. So, what we say that pre partition estimation of the thing, optimal program partitioning for offloading is dynamically calculated based on trade off between communication and computational cost at the at run time.

Offloading scheme based on profiling information about computation time data sharing at the level of procedure calls right. So, at the level of procedure calls we can do. So, a cost graph can be generated or constructed and branch and bound algorithm to applied to minimize the total energy consumption of computation and total data requirement. So, what we are looking at that it is not vanilla offloading some of the things to the things. So, there are a the lot of calculation involved as we come back to the our previous premise that one is that your energy savings should be higher, need to be maximized whereas, your latency should be preserved.

And we see that there are several other aspects of the components are there like security, reliability, fault in the server system etcetera and there are issues which are not directly under not and may not be a directly under the control of the cloud, neither on the mobile devices which is what we say that intermediate wireless network that also plays a things. So, that is there are issues of heterogeneity need to be addressed and need to be looked into a in a totality it is not that isolated I offload something, get some executed and type

of things need to be looked as a in totalities. So, that the user at the mobile end or the user get a seamless feeling or better resource utilization or better energy savings at that things.

Thank you.