Cloud Computing

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

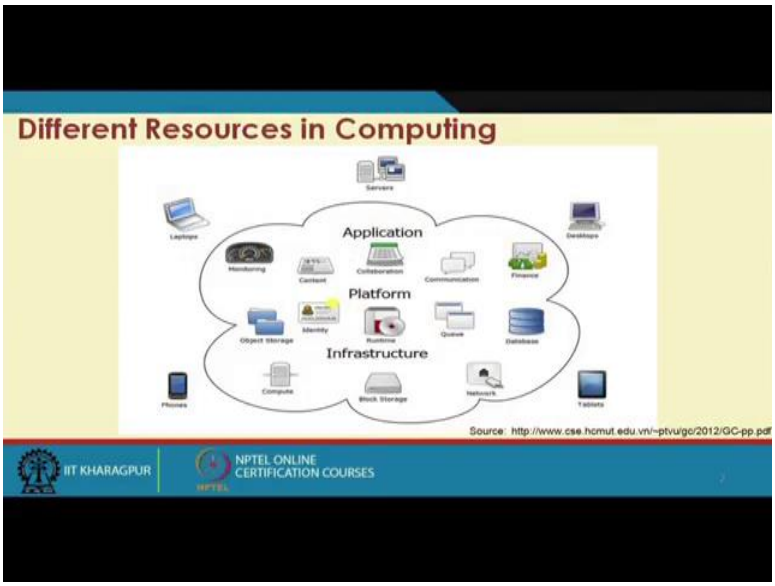
Lecture - 25
Resource Management – II

Hello. So, we will continue our discussion on Resource Management in Cloud. So, as we discussed last lecture or last 2 lectures on the resource management; what we have tried to look at that it plays a important role in overall cloud service, right. So, it is important not only from the service provider point of view; it is also important for the service consumer point of view, right.

So, provider want to have maximize its utilization of his resources with minimal energy cost and maximizing it profit right if you look at the from the consumer point of view. It wants to have a guarantee or a particular quality of service or support for its SLA, right. So, that the SLA is not valid. So, nevertheless this whole resource management he plays the important role for this; what we say quote unquote success of this cloud computing paradigm, right.

So, taught today; what we will try to look at some of the aspects of these resource management right we will try to look at a particular for a review paper and take up some of the aspects of resource management. So, I do not want to claim that release the all the aspects, but these are some of the important aspects where what a particular cloud computing environment or cloud computing platform. So, look at right.

(Refer Slide Time: 01:56)

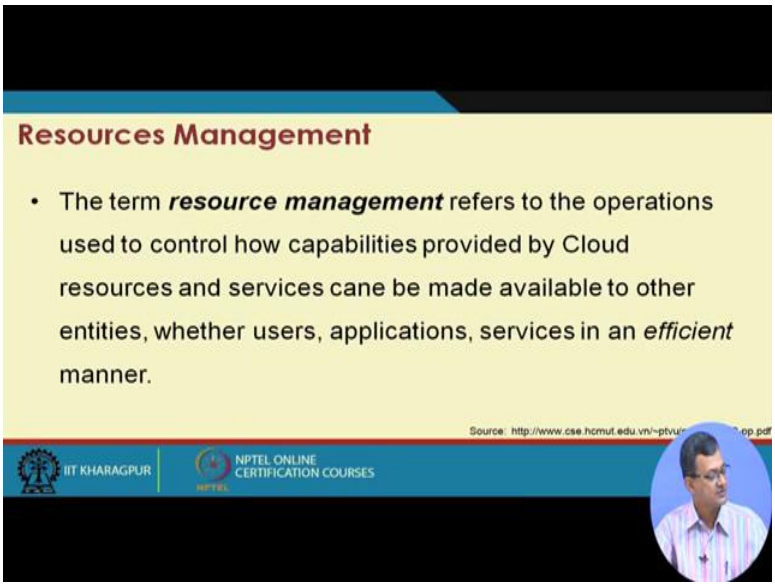


So, couple of slides maybe imputation from the other. So, what we made by resources just to recap quickly. So, I have at the core infrastructure platform and application or IaaS, PaaS, SaaS and there are different kind of user across the means; user for this clouds means they can be either human user or it can be some process or machine which are indirectly consuming cloud service to the for other services. So, what we want to look at that how optimize these resources can be managed at the core.

(Refer Slide Time: 02:41)

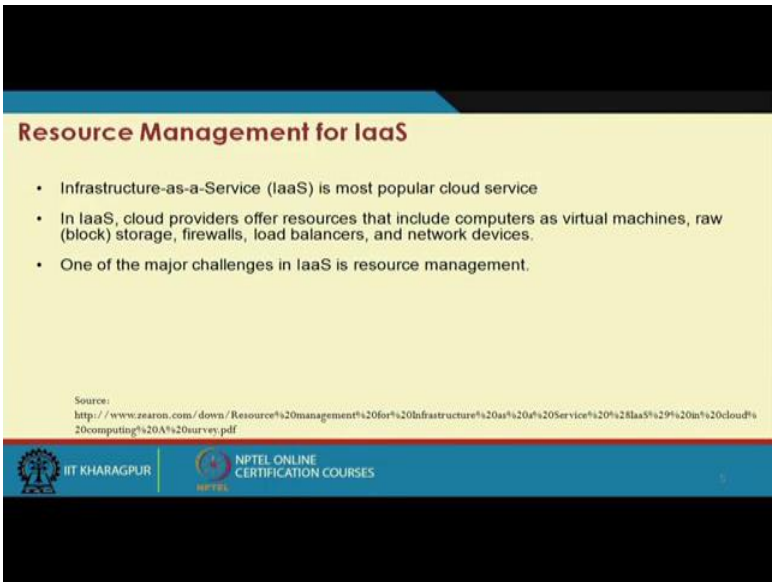
Now as we have seen, there are 2 categories of user; physical 2 category of resources; one is physical resource, another is a logical resource. So, the physically; what is there and logically like applications monitoring and type of things; so, both plays a important role in the resource management.

(Refer Slide Time: 03:01)



And also we seen these we have gone through these particular underlining definition of resource management it refers to operation used to control over capabilities are provided by the cloud resources and services and can be made available to other entities users applications in an efficient manner.

(Refer Slide Time: 03:25)



Now if we look at this resource management mechanism or resource management approaches the maximum or the maximum or the major thrust is on the management of the IaaS type of resources or infrastructural resources, right other resources like platform or SaaS; though they are also management is necessary, but those are mostly dictated by the amount of underlining backbone hard resources you are having, right.

So, some of these type of techniques are applicable across the different type of services whereas, some of the things are more good to the IaaS. So, what we look at today is more about that; what are the different approaches for IaaS type of resource management, right. So, infrastructure as a service is the most considered to be most popular or seen to be most popular cloud service among these different type of services. So, in IaaS, cloud providers offers resources that include computer as virtual machines raw storage firewalls load balancer network devices and so and so forth.

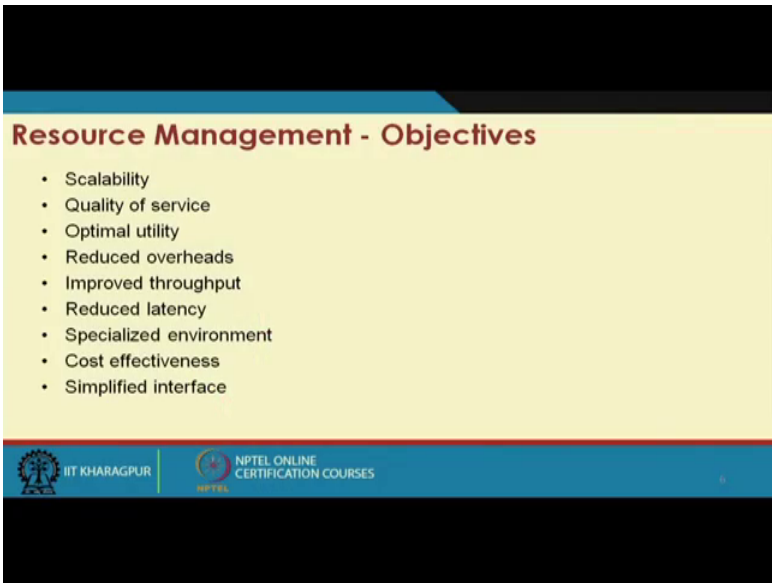
So, these are the different category of things which we consider as when we talk about infrastructure as the resource and one of the major challenges in IaaS is the resource management; how to optimally manage the resource and as we have seen as energy plays a important role for overall functioning of the cloud. So, it one of the aspect is with minimal or the energy consumption how I can give service at a particular level, right.

So, that is that is very important when we talk about IaaS; when we talk about resource management. So, I want to maximize profit from the provider point of view maximize

utilization of the resources and minimal requirement of energy right and of course, on the other hand, we have; we need to satisfy these quality of services and SLA right rather there are several metrics which we will see when we discuss today, but these are the aspects we need to look at when we look at the resource management aspects.

So, we will be following or we will be taking inputs mostly from a survey paper which where the link is provided you are free to download and look at the things and there you will get lot of other corresponding paper who are interested in further research or further study on this type of resource management are welcome to look at it.

(Refer Slide Time: 06:20)

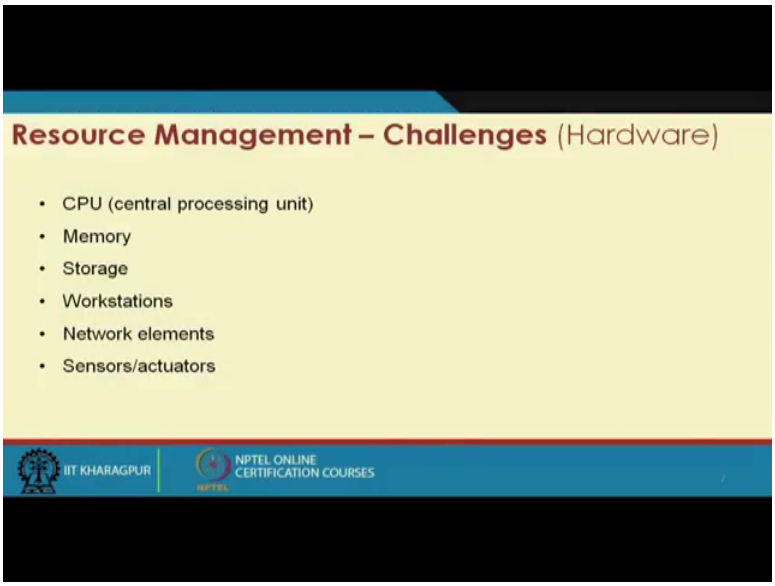


So, if we look at the resource management, if I broad objectives or the broad goal. So, to satisfy things that my scalability of that for which that cloud is one of the properties that scalability should be preserved that I can scale up scale down and ideally infinite scalability quality of services should be preserved optimal utility reduced over its like in 1 or 2; do a resource management protocol or algorithms if the overhead is high then I loose on performance, right. So, the it should be a optimal overhead rather reduced over it and improved throughput. So, that overall throughput should be improved reduce latency. So, that it should not increase the time latency of the overall system specialized environments like whether I want to have a specialized environment in order to have say as last day we discussed as a specialized environment for rack management power and so and so forth.

Cost effectiveness that overall; it should be cost effective, it should be financially beneficial to the both provider and consumer the provider should not spend more and the consumers should not have to pay more subscription for that and it should be a simplified interface the interface should be again simple, it should not be very cumbersome interface. So, that it is or I can say that we should have a ease of use will be there. So, it is easy to use that type of environment.

So, these are our broader goal or I can say broad an objective of cloud service provisioning without compromising this whether I can have a better resource management that is the objective of the things.

(Refer Slide Time: 08:23)

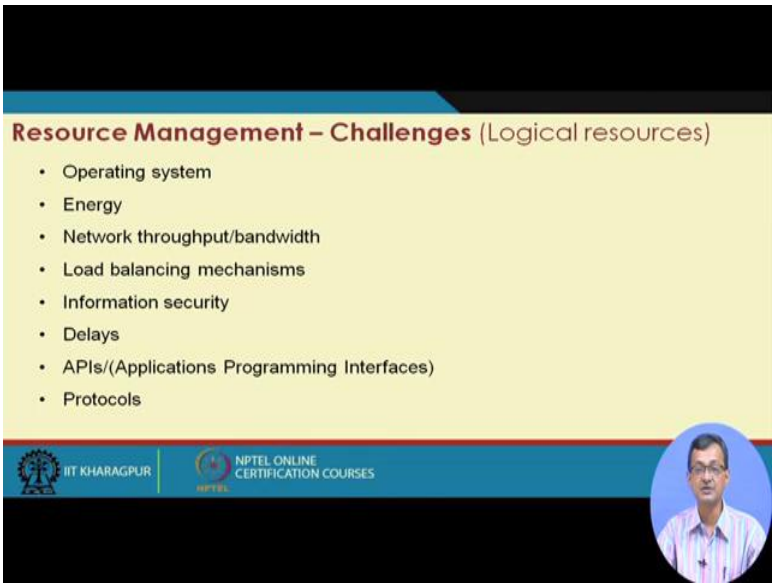


Now, there are several challenges like if you look at the hardware or the bare metal or the backbone of the things the one is that CPU management, memory management, storage management, workstation, network element, sensor actuators and so and so forth. So, these are the different components which are there which need to be properly managed and it is it; these are not isolated things right like CPU memory storage these are not isolated components. So, they have a Intel linking while operations on the thing. So, you cannot have a very high power CPU in low memory and type of things then the performance will not be there.

So, if the coordination between this bare metal or the backbone resources are important. So, when we manage resources we need to take care that those are preserved, right, I

cannot make a optimal management of storage without ignoring the other component of the like network component and other things a make a faster storage my network accessibility is still slow then purpose are not solved. So, that needs to be looked into.

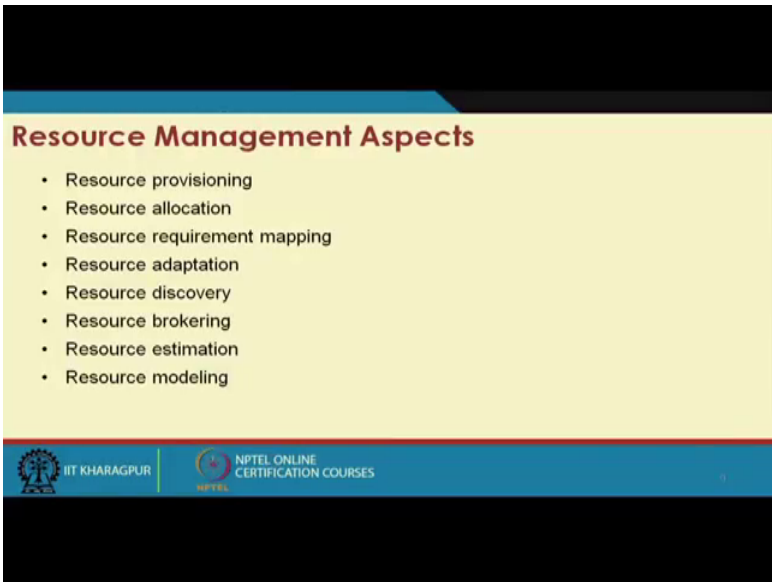
(Refer Slide Time: 09:35)



So, there are other logical resources. So, those are what we say physical hard resources there are logical resources like operating system energy management network throughput or bandwidth load balancing mechanisms information security which is coming up in a big way or which are looked into in a big way when you are leveraging lot of things on the cloud and specially your sensitive or semi sensitive information on the cloud or privacy preserving things delays that how much delay or time delays are their application programming interface or API. So, the API is whether need to be redone on new type of API has to be there and there are various protocols.

So, these are all what we say soft resources or logical resources which plays a important role and these hard resources and soft resources are not separate to each other they are intermingled rather need to be looked into in a integrated way. So, these are the different challenges or broad objectives what you look at.

(Refer Slide Time: 10:48)



Now, there are different type of approaches researchers followed and or to have resource management aspects or we say the different resource management aspects one is resource provisioning right we will see what are the resource allocation right. So, I need to provision then allocate the resources resource requirement mapping. So, whether I can map the resource requirements right or sometimes whether I can do a priority mapping of the resource requirement, right, like I am forcing that this sort of resource requirement are happening is there, then I provision it accordingly right like resource requirement in different parts of the different time scale of the day are different right different time period of the day are different-different time period of the year are different maybe over the years things are different so on and so forth.

Then we have resource adaptation right how resource can be adapted resource discovery there should be a mechanism that how can I discover resource faithfully like where do I find those resources and type of things and where how do I as a user can look for the resources. So, again I repeat that use and may not be always human user it can be another process another set of processes all together.

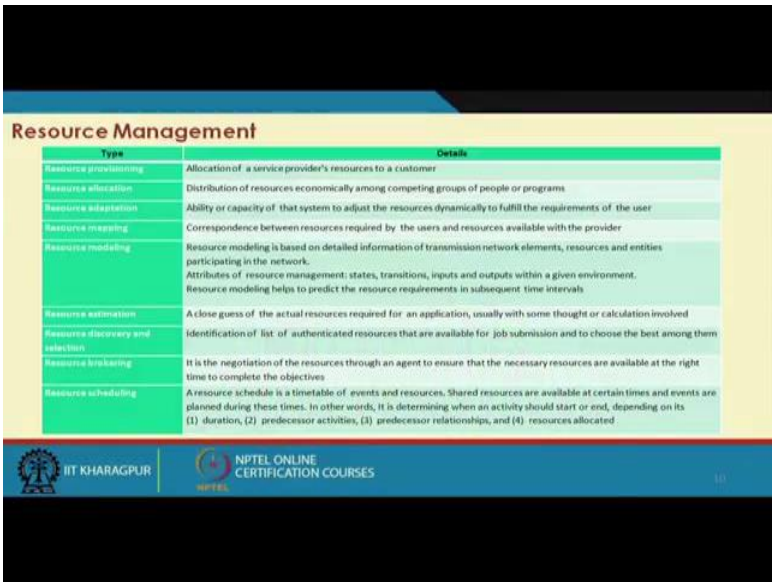
So, for optimal and that may be a part of a larger applications. So, in order to do that; I need to discover resources and type of things. So, there should be some provisioning some cataloging registry type of things whether resources can be there; there are resource brokering, right. So, some sort of a brokerage or agent based things where those are can

be where which acts as agent to have provide me a optimal resource; rather I should say that when a user request for a some resource then that initially it hits a agent or a broker which tries to look out that which are the which are the resource available which are less loaded nodes how VM can be allocated and so and so forth.

So, this is a important aspect that booker brokerage type of things resource estimation that is estimating that what sort of resource requirement will be there these are sometimes important when we do higher level of things like I am looking at a SaaS or PaaS label need to estimate that what sort of back backbone resources are required and there is a thing called resource modeling that how I can model the thing to the resources for considering estimation considering my present load and type of things.

So, these are different aspects and you can see that these are not all are independent aspects they have Intel linking between them also, right. So, these are different aspects and the emphasis may vary based on type of application or type of requirement you are there, right. So, in some cases the some of the aspects may be higher priority and so and so forth nevertheless these are not isolated components again they have a inter linking between the things.

(Refer Slide Time: 14:09)



So, this slide; what we are trying to look at that what are the different type of what are the different aspects and what are the different what they mean, right. So, what one

aspect is resource provisioning a location of service provided resources to a constant customers right the customer can be a user or a process, right.

So, resource allocation stands for distribution of resource economically among the computing groups of people or programs or processes resource adaptation ability or capacity of that system to adjust the resource dynamically to fulfill the requirement of the user. So, that based on the user requirement the overall; the system adjusts itself, how this resource; the available resource can be optimally used among the users, again I should say that without compromising the SLA and other quality of services and type of things.

Then we have resource mapping which says that the correspondence between the resource required by the users and the resource available with the providers. So, based on the resource available and they require requirement how we map resource modeling resource modeling is based on detailed information transmission network element resources entities participating in the network, right.

So; that means, attributes of resource managements if we look at they are different states different transitions different outputs with a given environment; so, every resource management. So, if I look at the resource management as a as a entity or a frame work. So, it has it goes in 2 different state it has different transition from one state to another and every state has the output type of things. So, I can have realized some sort of a state chart diagram or type of things and based on that I need to model that based on this; how this transition will go on.

So, resource estimation; so, how I can closely guess the actual resource required by application usually with some thought or calculation involved I can do some a priori I may have some a priori knowledge about the application or I can have some meta information at the application like this application may require so much memory. So, much displace. So, much threads and type of things and based on that I relocate the resource discovery and selection.

So, as we are discussing identification of list of authenticated resource that are available for the job submission and choose the best among them. So, it is always possible that you have multiple resources or multiple providers with resources are available. So, that discovering that which are the resources and which is the suitable thing and allocating

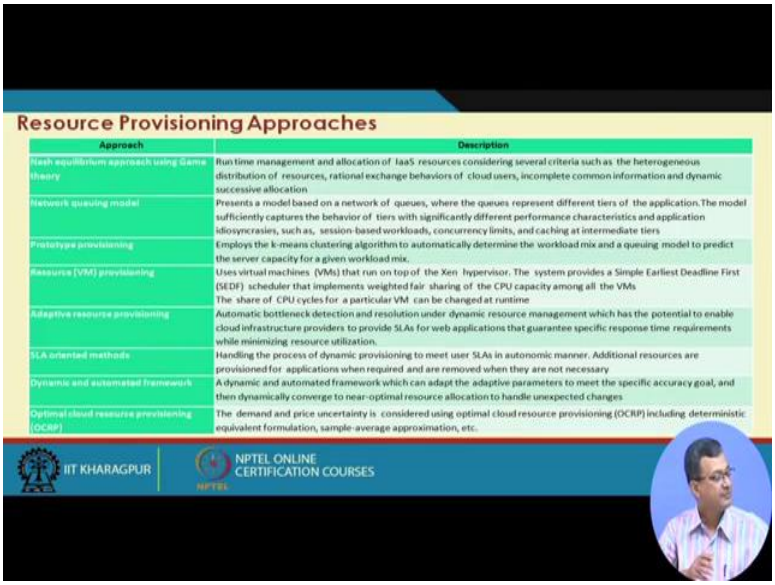
the most optimal and based about thing and resource brokering. So, negotiation of the resources through an agent ensuring; necessary ensuring that the necessary resources are available at the right time to complete the objectives.

So, I broke up because I have a requirement as a user I have a requirement as a user process, right and then I want to broker I want to negotiate with the agent that which are the things available, right, how things will be available. So, that I my objective is fulfilled and the objective may be resource wise objective the objective sometimes can be on the pricing objective also, this much cost and this much things I have to choose.

So, there is see there is a need for optimization of the whole thing, right. So, I require a brokering service for that and finally, resource scheduling, right. So, is this a scheduling is a timetable of events and resources, right. So, say our resources are available at certain times and events are planned during those times, right. So, it can be; so, I have resources I have my operation procedure. So, time I require some sort of timetable of sibling the resources that exactly a sibling problem per se.

So, I may have lot of this may have lot of components like duration ha some predecessor activities some predecessor relationship resource allocated and so on and so forth. So, there can be different component for to determine that start end and type of things.

(Refer Slide Time: 18:58)



So, if we look at some of the approaches or some of the different type of aspects like as we discussed previously, resource provisioning resource allocation and try to see that what sort of what sort of approaches people are following or researchers are following into things like first one let us see at the resource provisioning approaches like. So, what we have Nash equilibrium approach for game theory.

So, using some sort of a game theoretic approach to find out that optimal uses of the resource, right; so, what it does the runtime management and allocation of the IaaS resources considering several criteria like heterogeneous distribution of resources, rational exchange behavior of the cloud users in complete common information and dynamics successive allocation and so on and so forth.

So; that means, I based on this different components like heterogeneous duration of resources or users pattern of the cloud users and type of things I want to have a game theoretic approach to look at this. So, it is; we can look at as a game where one side is that the consumer which are a; who are hungry for the resources or looking for the resources other side there is provider who are provisioning the resources and I want to find out a optimal way of allocating the resources. So, that is Nash equilibrium based using game approach.

So, there are there are research and or there are methods and approaches people are following there we have network queuing model. So, which a model based on network queue like those who have gone through network queuing model or network queue models in data networks, etcetera you can understand, it is more of a again resource provisioning mechanism queues where queues represent different tires of application.

So, the model significantly or sufficiently captures the behavior of the tires with significantly different performance characteristics and application like session based workload concurrency limits and caching the intermediate tires and like that. So, it is try to you; what we do what we are what they are doing here to try to exploit network queuing model there are approaches for prototype provisioning employs the k means clustering algorithm to automatically determine the workload mix and queuing model to predict the server capacity for a given workload mix.

So, what it is trying to do. So, it tries to cluster using say came in cluster k-means cluster to automatically determine these; what is this workload means of the different user and then to predict that what are the what how can be provisioned.

There are other resource provisioning like VM provisioning, things like user virtual machines that runs top of the Xen hypervisor. So, the system provides say some sort of a scheduler like in some work they propose a simply simple earliest deadline first scheduler that implements the weighted fair sharing of the CPU capacity among the VMs, right.

So, what it is doing it is taking the VM which run over the hypervisor and scheduling it based on the type of type of the load it is it is getting that share CPU cycles a particular VM can be changed on the runtime and so on and so forth. So, if I have requirement from more resource requirement, then I migrate from one VM to the other VMs and type of things can be done.

There are other methods and approaches like adaptive resource provisioning which tries to automatically detect the bottlenecks and residues and resolve that using dynamic resource management there are things called sea SLA oriented resource methods, handling process and dynamic provisioning to meet user SLAs in a automatic manner dynamic an automatic framework which adapt the adaptive parameters which adapt the parameters to meet the specific users or accuracy goals. So, it goes on provisioning the recruit resources based on the quality of services or the type of SLA, it has to support and also there is optimal cloud provisioning mechanisms what which tries to look at the demand and price uncertainty considering those try to optimize.

So, we see there are several approaches which can be used for this sort of resource provisioning mechanisms.

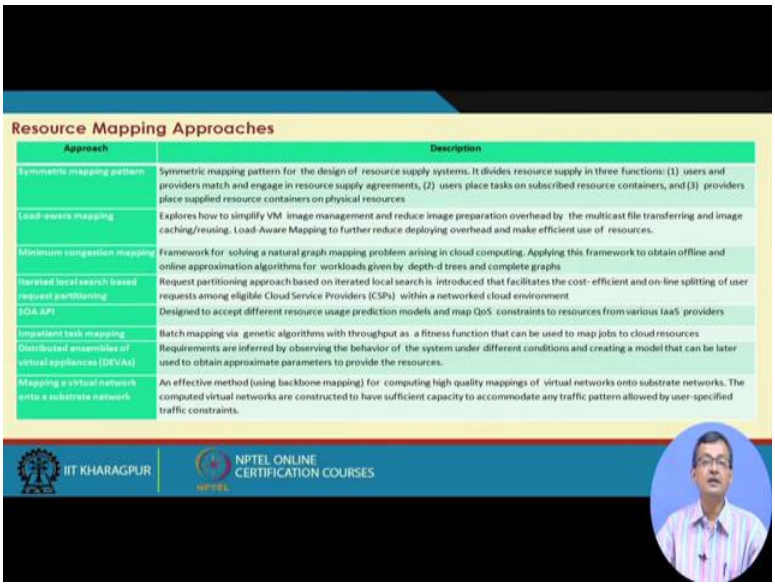
(Refer Slide Time: 24:11)

Similarly, if we look at the resource allocation there are again several approaches few are listed here like market oriented resource allocation. So, which are driven by the market requirement market demand on the things. So, we try to do model predictive model predictive control to find its solution of the of that particular resource allocation there are intelligent multi agent model primarily looking for resource view virtualization to automatically allocate services resource available specifically for devices which are mobile, right.

So, it is I can have a intelligent multi agent model to allocate optimal resources energy aware resource allocation. So, this allocation is energy aware. So, that I can do a optimal energy provisioning measurement based analysis on performance. So, it; allocation again based on different metrics or measurement parameters dynamic resource allocation methods real time resource allocation mechanisms like if there is a real time demand on the things how resource can be allocated. So, designed for helping small medium size IaaS cloud providers to better utilize their hardware resource with minimal operational cost by a well designed underlining hardware infrastructure, right.

So, in order to help for this especially small and medium sized IaaS cloud provider; so how it can be allocated in a real time and dynamic scheduling and consolidation mechanisms over and above I can have a dynamic scheduling and consolidation mechanisms of available resources.

(Refer Slide Time: 26:20)



There are again several approaches for resource mapping like symmetric mapping pattern that is for designing resource supply systems it divides the resource into three major functions users and providers match and engage resource supply management agreements.

So, that is users and the providers match that where the requirement are matching and do that type of matching before that users place tasks on the subscribe resource containers right. So, that it subscribe resource container place the tasks and the mapping is done or provider place supplied resource container on physical resources and type of things.

So, these are driven by container based services which is another type of; another technology which is coming up in a bigger ways that this container classes and container things. So, user can subscribe resource continent place their tasks or providers can place supplied resource container on the physical resources.

It can be a mapping of the load aware mapping. So, explore how simply VM image management and reduce image preparation over it by multicast file transferring and image caching and using. So, it is based on the load, it does a load aware mapping to reduce deploying over a rate and make efficient use of the resources. So, based on the available load it does the load availability there can be technique for iterated local search based request partitioning.

So, whether I can partition request partitioning approach follow a based on iterated local search to facilitate a cost efficient and online splitting of is your request among eligible cloud service provider. So, user may be requesting on a user requests, I can basically partition into smaller part if there is a there is a way of partitioning it, I can do a intelligent partitioning algorithm and then allocate the things into different CSPs like different cloud service providers and it that is; that means, that a large requests can be partitioned into smaller and look at it.

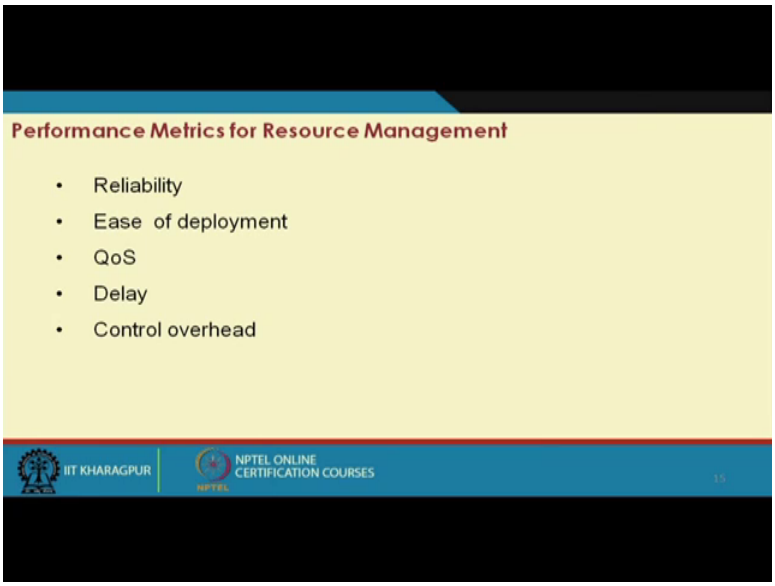
So, there are other approaches like distributed in ensemble of virtual applications like I have virtual applications ensemble name or mapping a virtual network of a substrate network. So, I have a underlining network and then I map a substrate network like why map a virtual network which is main for the user to this substrate network, right.

So, again this is a resource mapping from the network side of view there is a requirement from the user from the network and map it on the things.

(Refer Slide Time: 29:27)

And there are several adaptation approaches like reinforce learning guided control policy. So, that is a learning mechanism to look at the adaptation there are web service based prototypes, right. So, which can be used for resource adaptation there are several others like looking at virtual networks DNAs based load balancing and of course, we can have hybrid approaches for having this sort of load resource adaptation.

(Refer Slide Time: 30:01)



So, if I have this several type of techniques like as we have seen here is like resource provisioning, allocation, we discussed few of them; resource requirement, mapping adaptation and so on and so forth; how to judge that they are performance finally, what we are looking for is some matrix, right.

So, all those approaches need to be judged based on some metric like reliability ease of deployment like I have a mechanisms and it takes lot of lot of overhead to deployment quality of services should not be compromised; there should not be delay or much delay or the delays should be within the limit and control over it in order to manage these resource management things in order to control these resource management mechanisms or processes what are my control over it, right.

So, whenever we are looking for any resource management a play resource management tools or techniques we need to look at all those different aspects. So, otherwise the overall resource management may kill the basic purpose of this cloud computing paradigm, right. So, that is of uses scalability and in finite resources and type of things we may suffer.

So, we need to look at this different matrix and if you can you can see that these are the matrix may differ from different type of requirements, right. So, different user may have different or different user processes the different requirements. And where somewhere the reliability may be pretty high somewhere the quality of service somewhere some of

the applications may be delay concerned, some of the application may be accuracy concerned. And we need to take up the actual take up the resource management process resource management tools and technique based on those parameters.

Thank you.