Cloud Computing

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

Lecture - 24
Resource Management - I

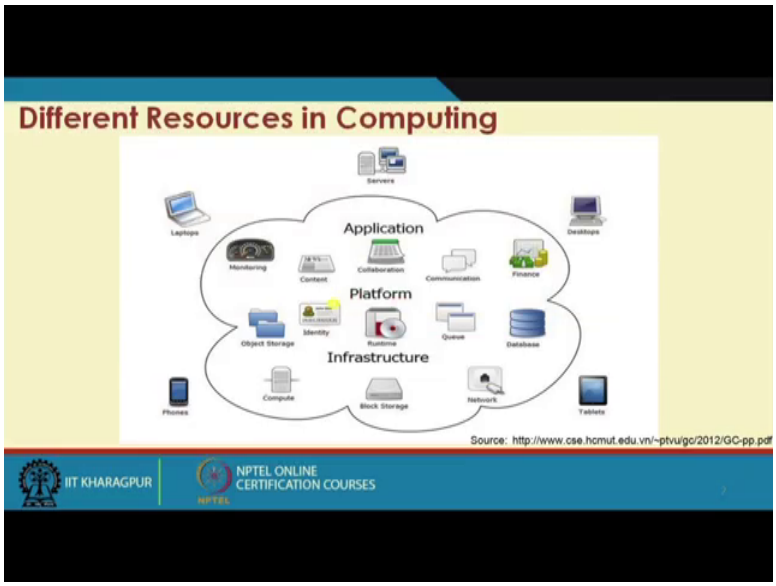
Hello. Today we will discuss one of the important aspects of a Cloud Computing or any type of this sort of service oriented mechanisms like cloud or distributed system or agreed or any type of systems that is the Resource Management. So, what we are thinking of a cloud? We are thinking that cloud is a infinite resource pool; along with that we are basically leveraging these resources to multiple users.

So, from the customer point of view or the user point of view or the consumer point of view; that it is a something a scalable service and a consumer can get as many as resources based on something of the concept of pay as you go or metered services. So, from the provider yield; it is important to properly utilize the resource in both terms that it can serve the consumer in a better way. And in other sense also, which is limited or what we say finite amount of resource; it can serve as many as consumer or the customer.

The thing is that from the providers there is a business angle also; he need to optimize the or he need to maximize his profit without compromising on the quality of services or the SLA violations. As we have seen, if there are violation of SLA's; it has a long implication, there has to give penalty and so and so forth. There is another aspect of the whole thing; even if a particular provider has say money or can afford number of resources, but there is a limitation or there is a overall environmental aspects of it; how much of your carbon footprint, how much energy resource you are having.

So, there is a restriction or there is a obligation on that mind of. So, keeping all these thing into mind, this resource management plays a important role in the thing. So, what we will do in a couple of lectures; we will look at that what are the different aspects of a resource management in the context of cloud computing.

(Refer Slide Time: 03:02)



So, if we look at different type of resources in a typical cloud computing things. So, what we have as we have mentioned; we have infrastructure, so basic infrastructure where is a compute resource, storage resource, networking resource or a related resource on the things. There is a platform which is above the infrastructure or here if I say that infrastructure as a service is the major thing, then we have a platform as a service.

And then at the end we have the application as the resource. So, the resource can manifest in a different way; it can be the infrastructure. Though primarily when we talk about resource or resource management, we usually fall back to infrastructure as a service, but considering the overall cloud or overall operation of the cloud; so, basically at least we are having infrastructure as a service, platform as a service or software as a service, also we have infrastructure as a resource management thing or platform as a resource management thing and application but though prominently a infrastructure plays a bigger role.

So, these are the different type of resources and the customer basically hook into the cloud with different type of heterogeneous systems. It can be servers, it can be laptops; maybe a smart phones, tablets, desktop and so and so forth. Number of cases, a cloud; we can say that a cloud or a service can take the cloud service or service providers thing to leverage on other things. I provide service taking some service from a service provider and so and so forth.

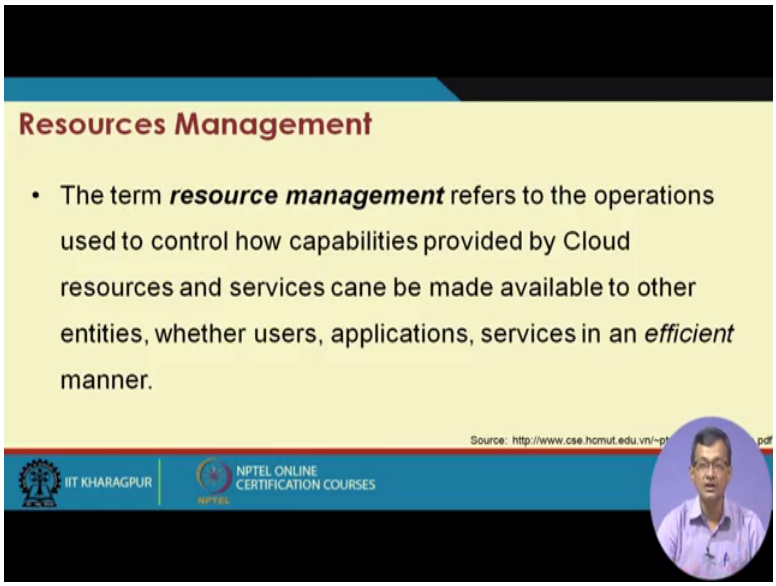
So, overall this management and optimization of these overall operations is a very tricky one. So, what we try to look at today is that; what are the different aspects of the things at all; whether with the same type of a hardware another things available, is it possible to manage the resource in a efficient way. That is our basic call for this particular resource management.

(Refer Slide Time: 05:27)

So, as we have seen resource type can be physical resources like as we see that computer, disk, memory databases, network, scientific instrument. So, database is again it is a soft resource so what we say that integrated this; means that hardware systems meant for the databases and also. There are other logical resources like what we say; there are applications like communication applications and other applications; executions of some processes which execute on the CPU's.

Monitoring and management of the things, so these are monitoring tools, management tools etcetera; these are different other resources. Some resources are directly utilized; like if I say that CPU and a hard disk and other things; some are basically meant to manage those resources.

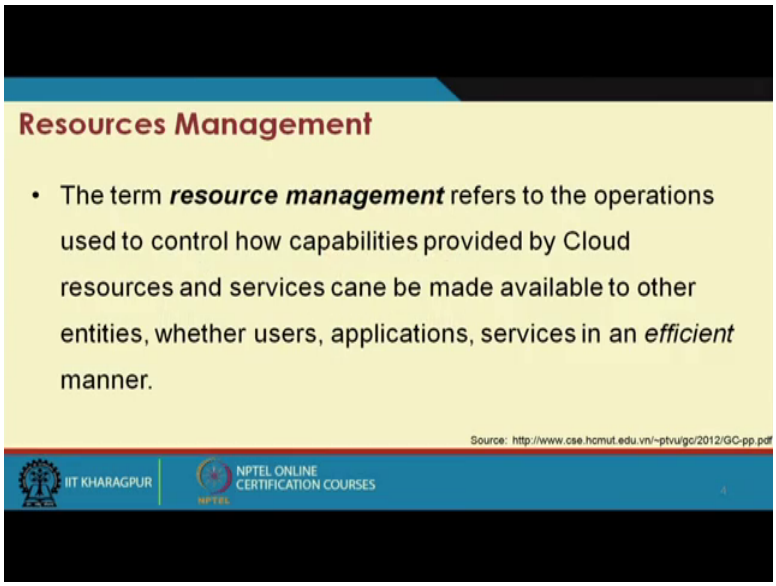
(Refer Slide Time: 06:26)



So, if you look at in a broad term like; what we are trying or what if we say the resource management; what we want to do with a resource management, the term resource management refers to the operations used to control how capabilities provided by the Cloud resources and services; there is a typo can be made available to other entities, whether users, applications, services in an efficient way.

So, what we want to do that whatever the resources and services; that means, hard and soft resources are available with the cloud; how it can be made available to the external entities like; it may be a user, it may be a application, it may be a services; in a efficient way. So, there is quote unquote the term efficient is tricky, but efficient means it can be efficient in maximizing the profit of the ISP, it can be efficient in energy optimization or in efficient in giving or a respecting SLA's and providing best quality of services; nevertheless the resource management plays a role in all those things.

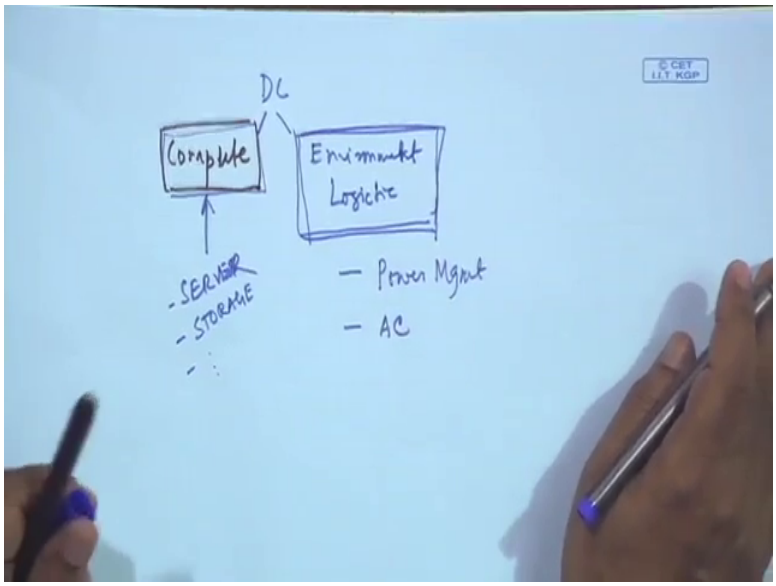
(Refer Slide Time: 07:44)



Now, if we look at data center a power consumption. So, what we are referring at? That is a ISP or a internet is or a CSP or a cloud service provider. So, they are having data centers; like major providers like IBM, Google, yahoo or; so, there are several things like Microsoft, Amazon and so and so forth. They have a huge data centers across the world as we have seen.

So, based on the things that these resources are outsourced from this data sensor to be more specific; so, what is the power thing? So, there is some report; it may not be the very recent report; it is much more higher than the whatever we are talking about. So, it is estimated that servers consumed around 0.5 percent of the world's total electricity usage; so, it is a huge (Refer Time: 08:47). Now, if it is not only the server; if you look at a data center.

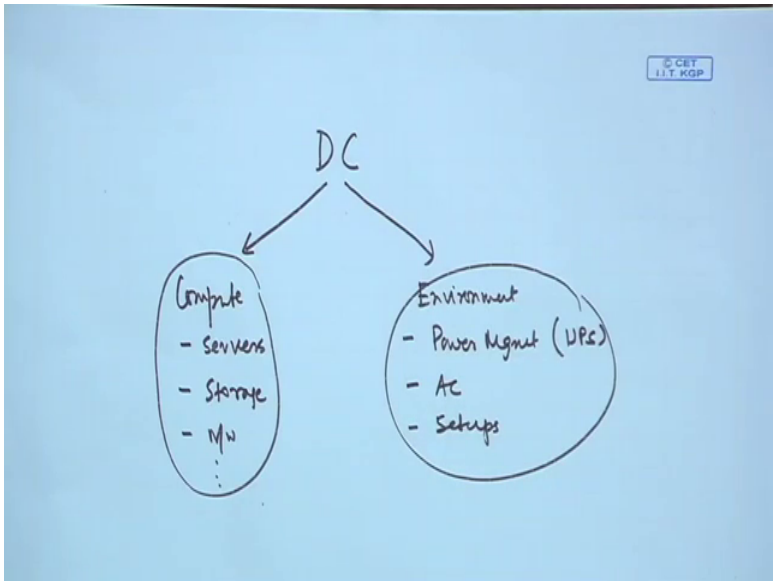
(Refer Slide Time: 08:56)



So, data center is; if you look at it, there are two major component; one is that compute infrastructure, another is what we say overall environmental infrastructure or data center other logistics; what we say; Environment or Logistics of the data center.

So, what we want to do? These basically gives you say servers or say storage and maybe adjoining thing; like network etcetera. These are providing primary looking for your power management or power supply or air conditioning and other related stuff which is related to these data center things.

(Refer Slide Time: 10:26)

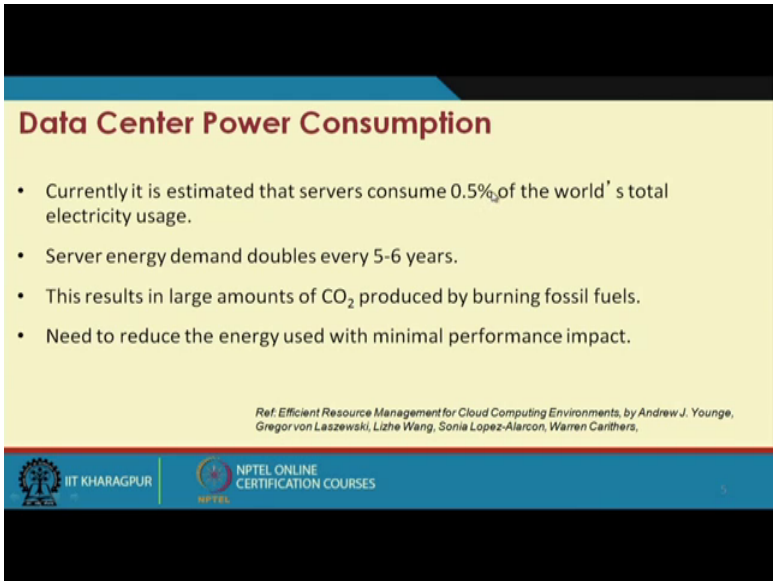


So, if I have a DC; so, it has two component; incidentally this two have; if not equally power hungry things. So, if you look at the amount of power; so, if we make it. So, we have one; if we have DC; one component is towards compute, one component is towards maintaining the overall environment of the data center.

So, here what we having a servers, storage network and other other, whereas, here primarily your power management; maybe UPS; sort of power, then AC and other overall logistics arrangement or other means; different type of setups and it is said that to have the house the data centers. Now, there is a power requirement out here; there is a overall power requirement out here. So, if you look at; and not only that this is not only power, there is a power cum space, power cum space.

So, if you look at the overall things are; if I say the overall consumption here is x. So, it is also somewhat towards x. So, if it is not only the servers, but the consumption by the other type of; other logistic units are equally high. So, when we talk about data centers where it is 0.5 percent of the world total uses.

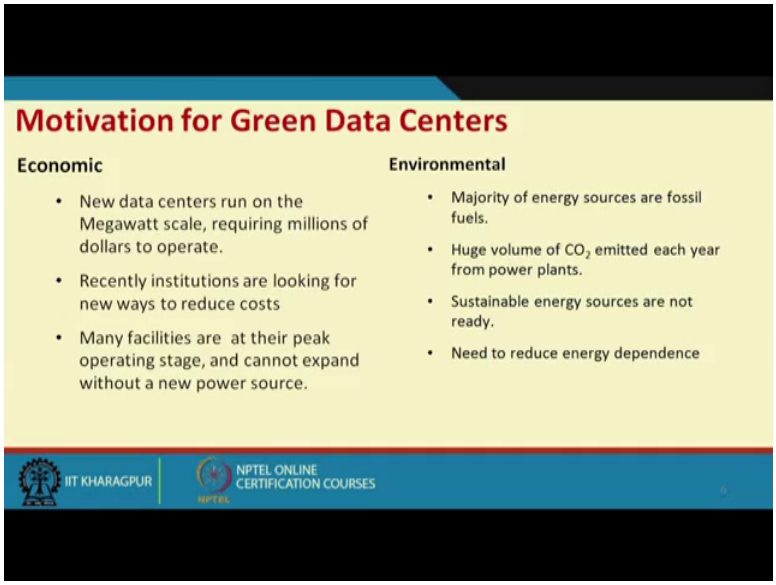
(Refer Slide Time: 12:00)



So, if you look at; if you basically make provision for other things, it may go above 1 percent of the things over power. And this is some data which is reported in some places, but it may be much higher.

And it is also saying these servers; demands when some report says that energy demand for the servers are increasing every 5 to 6 hour; a 6 years, maybe becoming double. There is a large volume of carbon dioxide footprint, which because of this fuel; burning of this fossil fuels. Need to reduce energy use for the minimal a performance, so what is our goal that we need to reduce this energy utilization with minimal effect or the compromising this performance. So, that should not be practically or that should not be very minimal impact on the performance; so, that is one of the goal.

(Refer Slide Time: 13:05)



So, as you are discussing; so, the overall motivation if we say that one to make quote, unquote green data centers; which will take very less energy; perform at the highest level. So, there are different aspects, so one is this economic aspect; the new data centers run on megawatt scale; require millions of dollars or millions of money to operate.

Recently institutional looking for new ways to reduce different type of new ways to reduce cost, many facilities are there. Peak operating stage cannot explained without a new power source in some cases; like if I have; even that is why I was talking that even if I am having a fund available as a provider, but I do not have any resource to do that.

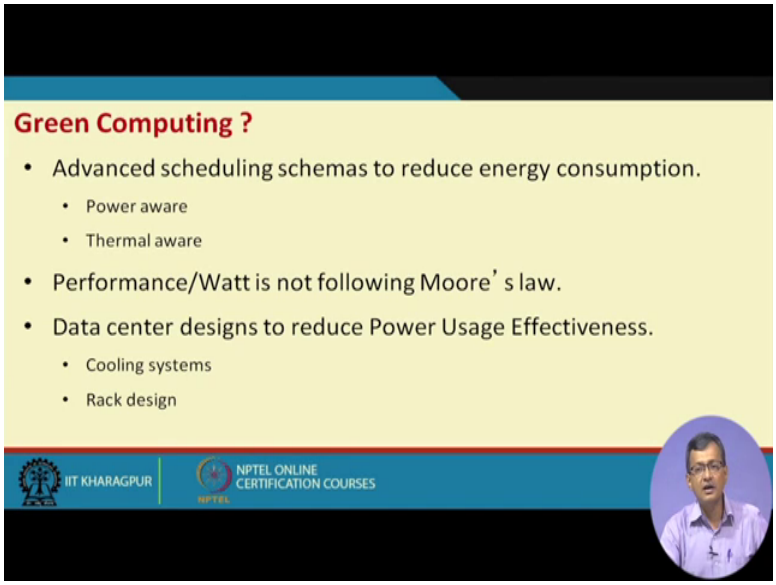
Suppose we have a data centers at IIT Kharagpur and we are taking power from the state electricity board. And even tomorrow, we say that we want to scale nothing to double and we are ready to pay, but the state electricity board may not have that type of surplus

power to supply. So, it is not only that whether I am having money or having more space etcetera, but whether that can be supplied.

There is a definitely environmental aspects and there are different type of legal obligations towards environment comes into play. Majority of energy sources are fossil fuels still to date. Huge volume of carbon dioxide emitted each year from power plants; which creates a environmental hazards.

Sustainable energy sources are not; till not ready to handle such huge requirement of the data centers. Need to reduce energy dependency; so, we need to have some mechanism to reduce the energy dependency. So, this is the two very very broad aspects of the theme.

(Refer Slide Time: 14:57)



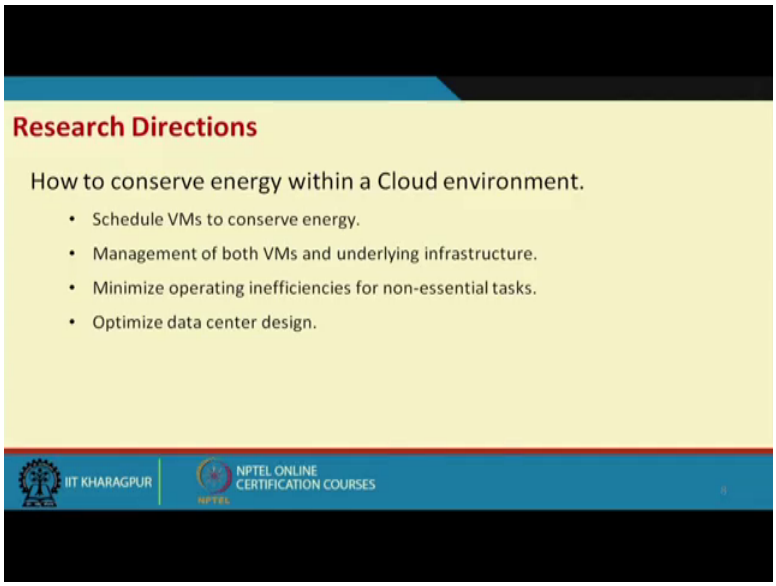
So, whether we are talking about green computing in this respect definitely. So, what we can do? What cloud providers tries to look at; is that whether this advanced scheduling schemes to reduce energy consumption is possible so, that; which can be power aware or an ender or thermal aware; both power aware and thermal aware, this sort of things.

Performance per watt is not follows the Moore's law; like our famous Moore's law. But it is unlikely that performance per watt is not following the Moore's law. So, data center design so that whether we can redesign or basically design the data center to reduce power uses effectiveness. So, that we can have effectively use the power like it can be

cooling system or the rack design or the placement of the racks and type of things that is more of the data center design; as we are talking just a couple of minutes back. That one is that having the server etcetera compute part of it, another is that; the data center design aspects of it that how efficiently you can design.

And if those who are visited some sort of data center; you know that there are different typical arrangement of the cooling, it is not conventional cooling of the may not be the whole server room or the space; it is a separate enclosure made on the data on the major server racks and the server racks are cooled specifically on a very confined area so that the energy required for this cooling purpose can be minimized.

(Refer Slide Time: 16:44)



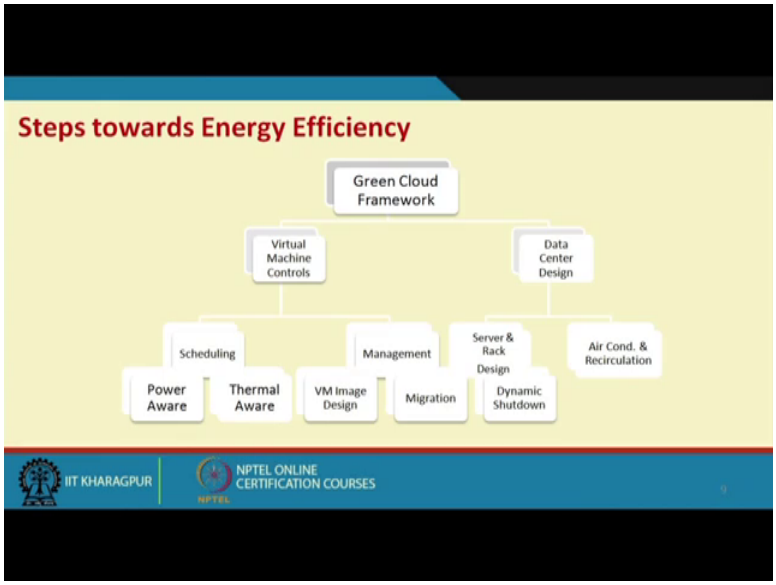
So, there can be different type of research direction and some of you; who are interested maybe doing some research on the cloud computing. So, there is a important this resource management looking at different aspects is a extremely what we say hot topic to do research, at different parts of the world people are working on this; that how to optimize this resource, how to have a optimal resource utilization for; without compromising services or without compromising the performance of the cloud.

So, how to conserve energy within a cloud environment? So one may be scheduled VMs to conserve energy, whether we can have proper VM scheduling on the conserve energy. So, management of both VMs and underlining infrastructure, minimize operating inefficiencies for non-essential tasks; so, that whether we can minimize that operational

efficiencies or non essential task. Like if you take a VM with a particular flavor and lot of resource; a lot of other packages etcetera running, but practically you do not require all those things.

Whether we can have a tailor made VM or tailor made OS; which can be run on the thing. Optimize data center design; the other aspects. Now these are; so, one is aspect of scheduling the VM; one aspect of a ease management; another aspect is design, design of the data centers and so (Refer Time: 18:24).

(Refer Slide Time: 18:32)



So, these are three broad aspects; we will try to have a quick look on these aspect so that we can have a feel of it. So, if you look at this overall green cloud framework; so one side is towards VM controls, how these VM scheduling management, other side is data center design.

So, this is more on the compute side of it; this is more of the data center infrastructure side of it. So, here the infrastructure is the hard data center infrastructure; the other come. So, if you look at the VM controls; so, one part is scheduling the view. So, which can be power aware or thermal aware; looking at the things. Other is that VM management, which is VM image designing; whatever the VM image having, whether we can have a efficient design of this VM image with basically only those packages or those services which are required are uploaded there.

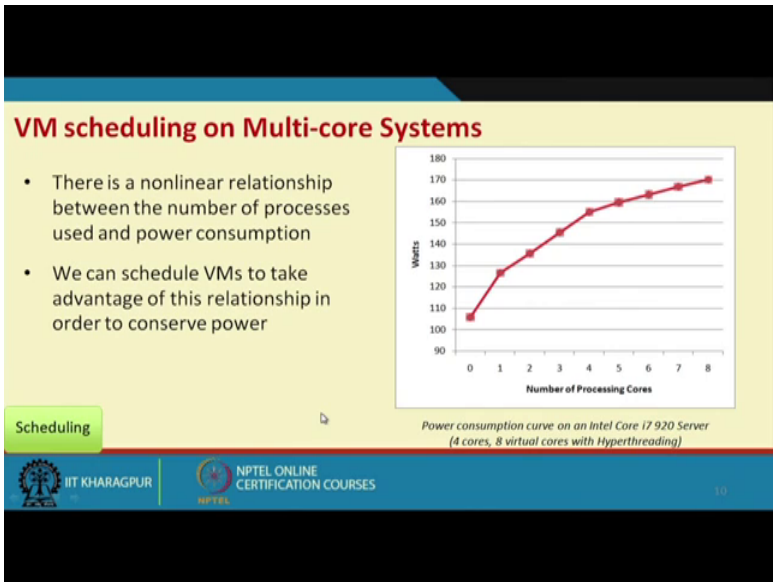
So, other is that whether you can look for VM migrations; so, what we are having actually if you look at it. So, what we are having a underlining hardware and there VMs are there. Suppose there are typically say 16 blade server; so, every blade server is running say 8 VMs. So, 16 into 8 is the total number of VMs are there, so now out of the 16 into 8 or say 10 VM.

So, 16 into 10 VMs are running or say 10 maybe on the higher side, so 16 into 5 say 80 VMs are running. Now 80 VMs; if there are at any point of sign; say 10 or 20 VMs are active, whether it is efficient to distribute over the whole VM servers available or I can concentrate the VMs into couple of servers.

So, because one is that; whether the server, VM is running on a server or not; it will go on consuming a base energy. Whenever I run a VM; it may consume some incremental energy. So based on that; it requires some sort of a simple mathematics to work out that whether it is efficient; so, one the data center design side, one is that server rack design is a important aspects and there is separately if you look at do not the cloud computing part; this is another research on the building of data center or infrastructure type of things. We are conditioning recirculation is a type of another aspects of the things.

So, there are if you look at that management side; there is another aspect of dynamics shutdown if the things are not utilized; so, whether I can dynamically shutdown there. So, there are a lot of aspects and if you see these are all have lot of research potential especially this part of the things on the compute side.

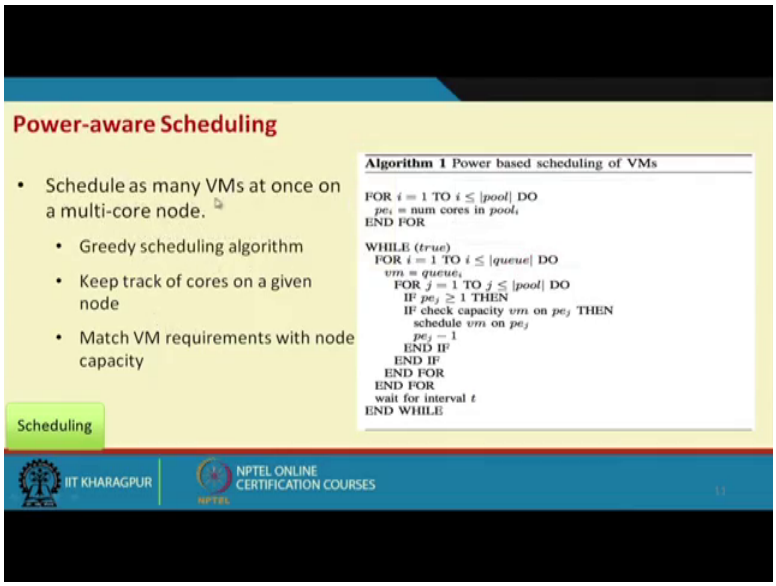
(Refer Slide Time: 21:40)



So, this is a typical example scenario if you see there is a non-linear relationship between the number of processor used and the power consumption. So, it is not like that that the number of processor used in the power consumption is a linear thing. So, if the number of processor goes on high, the power consumption is basically somewhat going towards some saturation; if not saturation the increment is not linear.

So, whether we can exploit this one; whether we can schedule VMs to take advantage of this relationship in order to conserve power. So that means, as I was talking that I want to concentrate my VMs into the underutilized server so that the VMs running on the number of servers are reduced. So, that effectively I can put this idle servers into sub hibernate or mode or low power consumption mode so that the overall power consumption of my data center reduces.

(Refer Slide Time: 22:50)

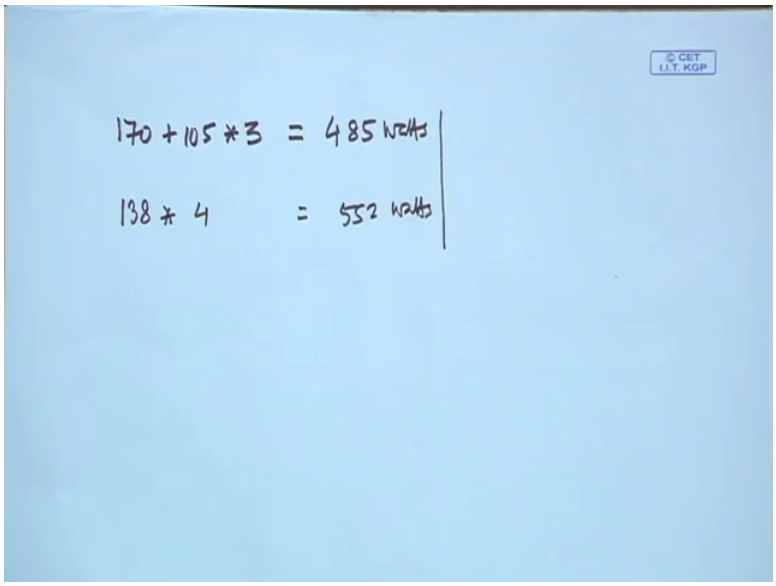


So, that schedule as many VMs once on a multi core node; greedy scheduling algorithms. We can do; there is a snap shot of a greedy scheduling algorithm, you can basically try out and see that how things work. Keep track of the core on a given node. So, we have to see that at a way for it typically in a node that how this codes are being busy or how much loading is there. Match VM requirements with the node capacity; if it is there then whether we can migrate the thing.

(Refer Slide Time: 23:26)

A simple very vanilla type example, so if I have 4 nodes; so, that it is in a idle condition or when a no load condition they consume 105 watt and when it is fully loaded with 8 VMs, it consumes say 170 watt. So, if I have some 8 VMs; then if we run on a 1 node, the overall consumption is 170 plus; 105 star 3.

(Refer Slide Time: 23:57)

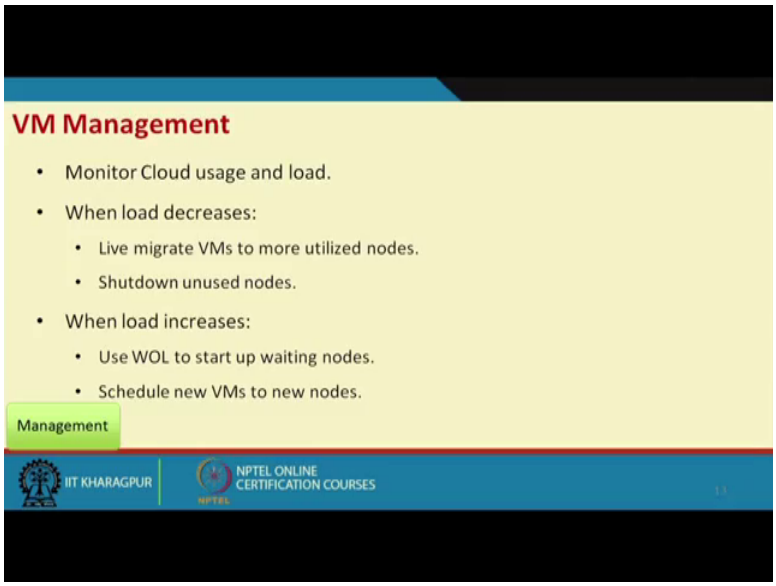


So, that is the overall consumption 4 watts in this case, this typically looking at the things. If I had it been it is distributed in thing; so, with this sort of 2 VM; let us say that is 138. So, I have 4; so, in this case 552 watts; so, that exactly what want to see that in doing so, we can basically reduce the things.

But there is a little catch in it; now in order to achieve from this stage to this stage, I need to migrate this; migration also have some cost. Migration has some cost, not only that here we are taking all VMs to be the same category. So, VM can be of different categories; so, it may not be able to put all the VMs into a 1 node or 2 nodes etcetera. So, we need to check and monitor that how much loading is there, how much free VMs are there so that we can basically migrate the VMs.

So, it is not like straightforward of multiplying that to what age; it is also this migration has some cost on doing that.

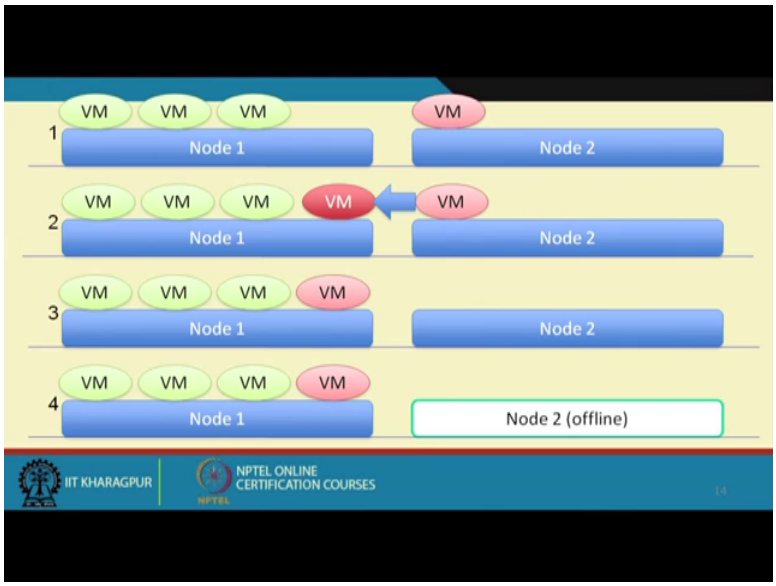
(Refer Slide Time: 25:18)



So, what we need to do? Monitor cloud usage and load. When load decreases; live migrate VMs to more utilized nodes; unused node. Shutdown unused node so, that I can basically what we are doing? We are compacting those things into the server things, so that is the live migration. Now, number of these your hypervisor or say number of VMM do support this type of live migration, though it maybe commercially costly but you can do a live migration and it has lot of implication other way also.

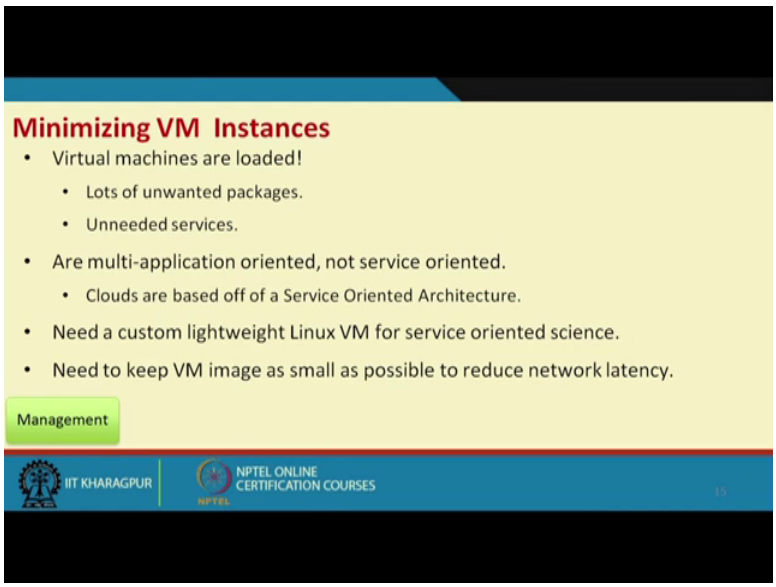
Like if a server is having problem, then you like migrate to the other servers and so on, but that is those are things are possible. So, when load increases; so, we can have some sort of wakeup call; wake up on lan sort of things to start the waiting node. So, in the load increases; I basically give a wakeup call to the nodes. So, which are sleeping to wake up, schedule new VMs to the new nodes. So, there are technologies available and if we can be efficiently use; effectively we can have a energy efficient thing.

(Refer Slide Time: 26:31)



So, that it is node 2, it is migrated there; the VM is put into service, then the node 2 become idle, then made into hibernate or offline mode; so, effectively I am running on a node 1. So, it is this; though there is a cost of this sort of a migration, but overall if I am achieving efficiency or in terms of power consumption, then that maybe a good option to look at.

(Refer Slide Time: 27:03)

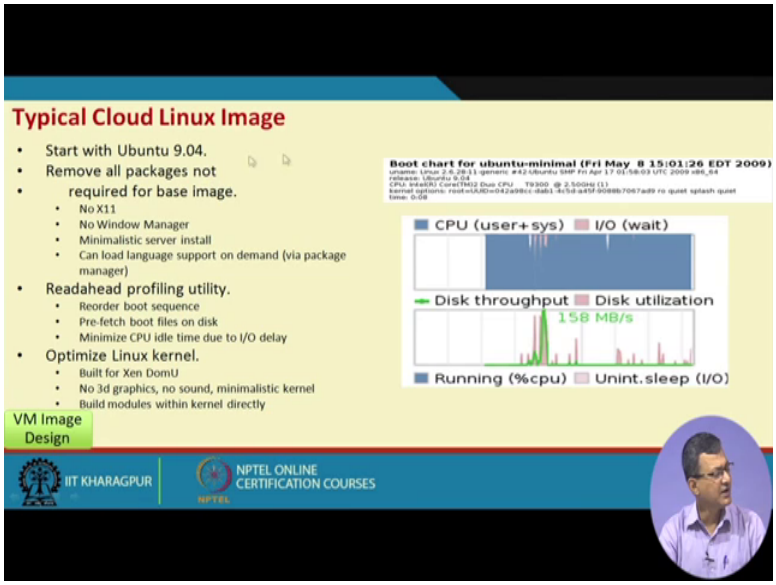


So, another aspects what we have seen that management; why is that? Whether I can minimize VM instances? VM machines are sometimes too loaded, so lots of unwanted

packages, unneeded services are there. So, you take some while VM it comes with a basic configuration; there which may not be utilized by the consumer. So, our multi application behaves on multi application oriented, not service oriented in number of cases, clouds are based on of a service oriented architecture.

So, we should have service orientation; need to customize lightweight Linux VM for service oriented science or services, so that we can need to have a customized Linux VM or the customized VM per say. Need to keep VM image as much as possible to reduce a network latency. So, if we have the VM image less then the; means while carrying of the network, the latency will be minimized. So, these are different aspects; so these are typical example scenario which starts with a Ubuntu; typically 9.04.

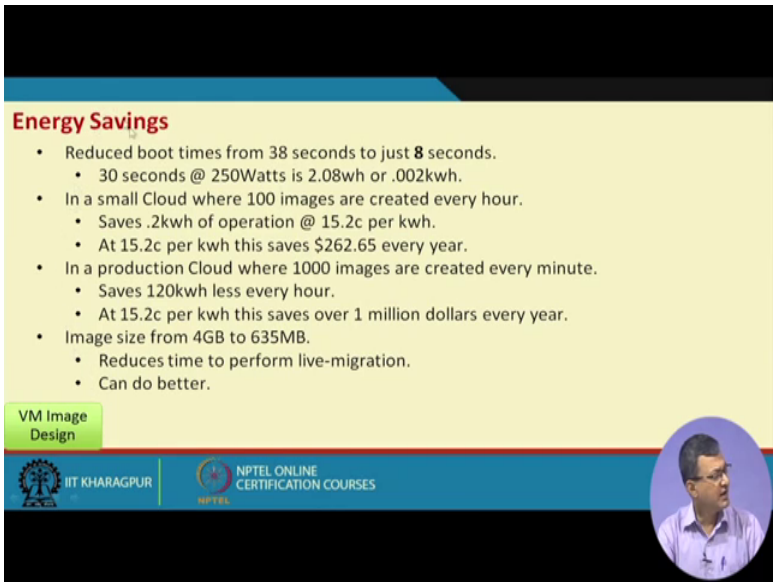
(Refer Slide Time: 28:10)



Remove all packages which are not required; like if I do not require X 11 or windows manager etcetera etcetera. Read ahead profiling utility, reorder the boot sequence so that I can have a prefetch boot files on the dicks minimize CPU idle time due to IO delays etcetera.

So, I can have a read ahead things like if I have that the steps to be followed, I can do a; a priori thing. Optimize Linux kernel; build on say in this typical case DomU. So, there is no 3D graphics, no sound so that I can have a customized Linux kernel which is primarily used for the things. So, based on the type of customer requirement or based on the services, we can basically optimize this type of stuff.

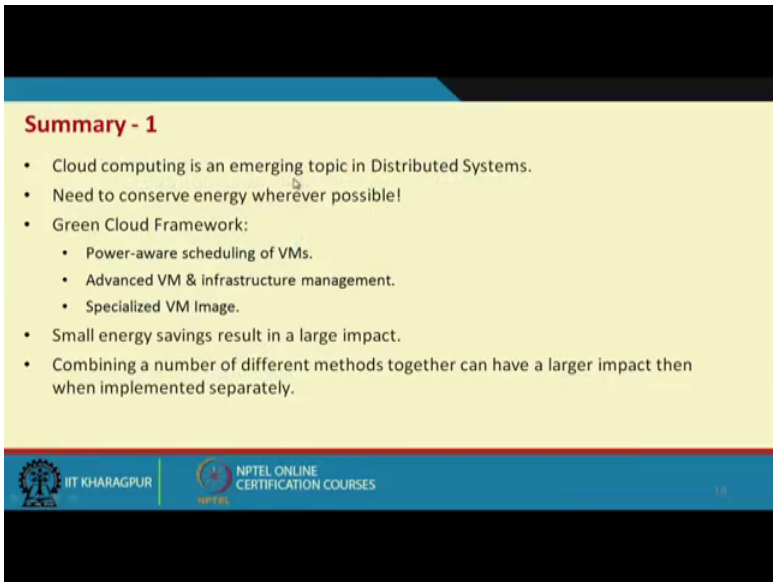
(Refer Slide Time: 29:05)



So, there are different energy saving some snaps parameters like we can; if it is reduced the boot time from 38 to 8 seconds; so, effectively we can have energy savings; even if a small cloud where 100 images per hour, so much energy savings can be there.

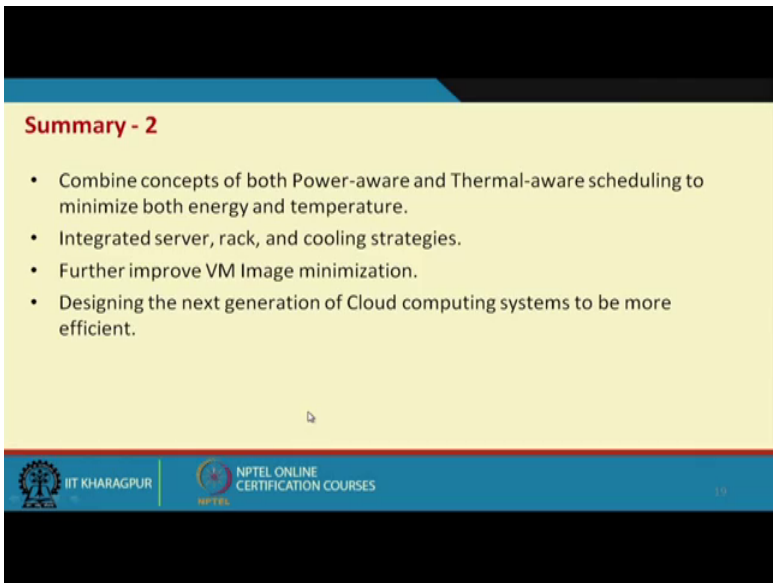
In a production cloud where 1000 images are created every time; saves a lot of energy in other way. A image size from 4GB to something 600 plus MB; so, reduces life performance; as we are telling that the life migration also comes with a cost and we can do better. So, what it is trying to say that all those things makes its energy efficient or have a proper resource manager without compromising the actual performance. So, if we try to summarize, so it is a emerging topic definitely.

(Refer Slide Time: 30:05)



So, need to conserve energy wherever possible; so, one is that green cloud framework, power aware scheduling of VMs. Advanced VM and infrastructure management, specialized VM images or customized VMs images. So, small energy savings result in a large impact, so even it is a apparently VM wise of node wise small, but considering the whole thing is a large impact. Combining a number of different methods together can have a larger impact than when implemented separately. So, it is not only piecewise having a collective or cooperative way of looking at it.

(Refer Slide Time: 30:46)



And there are lot of research interests and things combined concepts of Power-aware and thermal aware scheduling to minimize both energy and temperature. Integrated server, rack, cooling strategies; these days as our server racks are coming with integrated cooling power and so and so forth.

Further improved VM image minimizes and looking at that customer needs and type of things, customer profiling. Designing next generation cloud computing system to be more efficient; both in terms of energy, thermal and type of service provided.

So, with this let us conclude today and as we understand, there is a lot of scope of doing research and for the studies in this particular way of resource management.

Thank you.