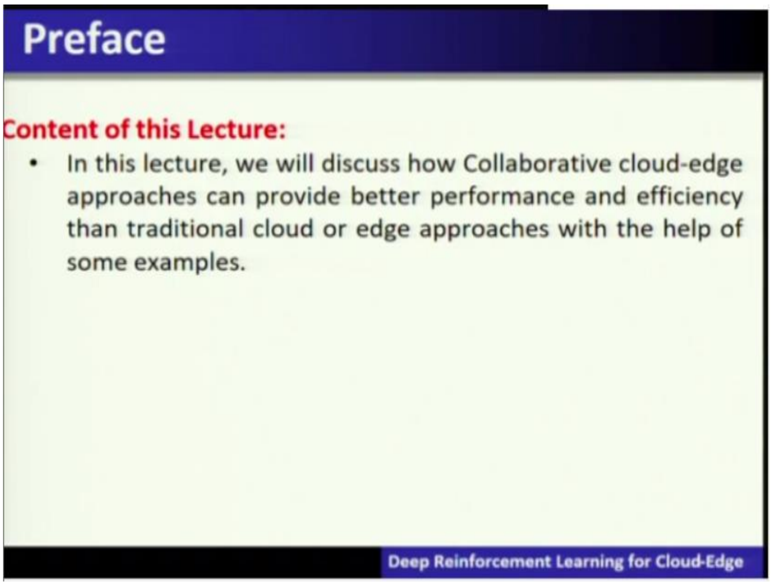
Foundation of Cloud IoT Edge ML
Professor Rajiv Misra
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Lecture 10
Deep Reinforcement Learning for Cloud-Edge: Example

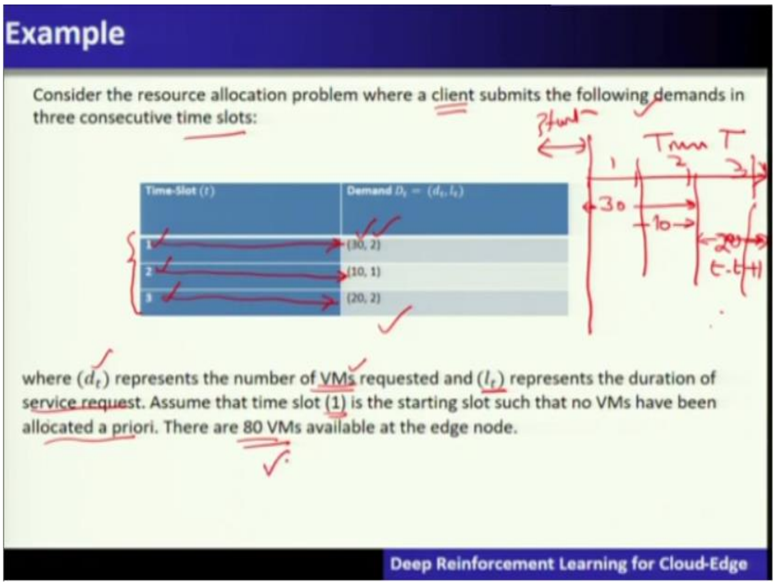
Deep reinforcement learning for Cloud edge an example, I am Doctor Rajiv Misra from Indian Institute of Technology, Patna. The title of this lecture is Deep Reinforcement Learning for Cloud Edge an Example.

(Refer Slide Time: 0:29)



In this lecture, we will discuss how collaborative cloud-edge approaches can bring better performance, efficiencies compared to the traditional cloud and approaches using this deep reinforcement learning for which we will show with the help of examples, this particular approach of using machine learning or a deep reinforcement learning in this part.

(Refer Slide Time: 1:00)



So, us take the example. For that, we will explain you this example and then we can move forward considering the resource allocation. So, consider the resource allocation problem where the client submits the following demand in the three time slots. So, we are considering the three time slots, 1, 2, and 3. So, that is the time is divided into 1, 2, and 3 different slots.

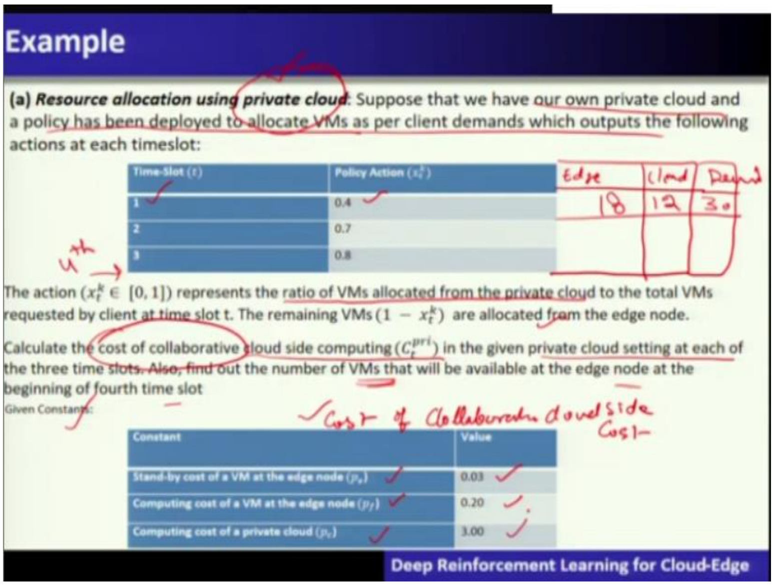
So, the time is divided into 3 different slots and each slot, the demand; capital Dt is expressed with a pair that is dt and it. So, dt, we have explained that dt means the demand, user demand at time t. Second thing, second component is it. So, it is the user's demand, needs the resources for the time period mentioned as it.

So, let us understand this. So, that means for a time period 1, the demand is about 30 units of resources for, let us say the time duration of 2 units. Similarly, for time slot 2, that is for this time slot. Here, let us write down. It is a 30 unit is required here for 2 units of time. So, this, so 30 unit for 2 times it needed. So, the first time slot T1 requires 30 units, and this will be for 2 time slots. Similarly, for time slot 2. So, that means for times slot 2, it requires 10 units and this is for 1 times slot.

For 4 times slot 3, for times slot 3, it requires 20 unit and for 2 times slot. So, that means fourth time slot or the next time slot is also to be considered or the next time slot. So, 20 means t plus 1. t and t plus 1, both are needed. So, this is shown here that, it is explained that small dt represents a number of virtual machines which are requested. So, let us say that the resources are represented by virtual machines and lt represents the duration of that request of that resources. That is the computation required for this duration that is represented by lt.

So, I assume that the time slot 1 is starting times slot such as no virtual machines have been allocated a priory. So, that means it is the start point. And we also assume that there are 80 virtual machines which are available. So, available virtual machines are 80, and if you see the demand is only 60 demand. So, that means the available virtual machines, which are available at the edge node is good enough to accommodate that.

(Refer Slide Time: 5:23)



Now let us see the strategy. So, resource allocation using the private cloud if you do, suppose we have our own private cloud and the policy has been deployed to allocate virtual machines as per the client's demand, which outputs the following actions. So, in a time slot 1, the policy action says that allocate 0.4. So, 0.4 represents the ratio of virtual machines, which are allocated from the private cloud to the total virtual machines which are represented.

So, that means 0.5 of the demand of the resources. That means 0.4 will be allocated from the private cloud out of the total virtual machines, which are requested by the client, and the remaining are allocated from the edge nodes.

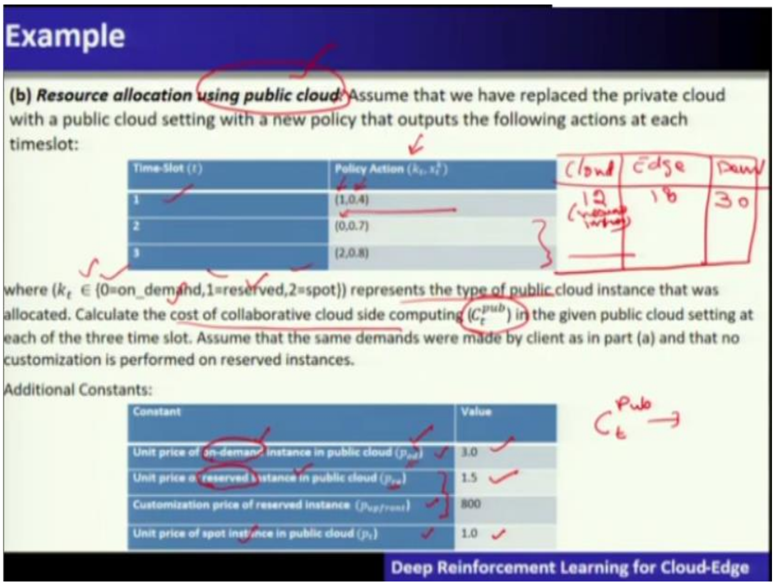
So, that means this will, if you consider here the resources coming from the edge, and if you see that the resources which will come from the cloud using this policy, you can see that if, let us say that this particular demand, let us say this demand in the previous slide was shown. If the demand is, let us say 30, then let us see the distribution. 12 will come from here then and 18 will come from edge.

So, you can see that the large share will come from the edge and less share will go from, will come from the cloud. Similarly, 0.7, if it is the requirement of the time unit 2 that also can be met and shown were here. So, let us see that the cost, the calculate the cost of collaborative cloud side computing, total cost in the given private cloud setting at each time slots, and also find out the number of virtual machines that will be available at the edge at the fourth. So, that is the third and at the fourth slots.

So, let us see the given constant to calculate this cost. To calculate the cost, what are the parameters to calculate the cost of collaborative cloud side that is in the model, we have already considered the cloud as a private cloud. So, what are the constants which are required to calculate the cost, which is shown over here is that is standby cost of a virtual machine at the edge node is p is let us say having a value of 0.03.

Now, computing cost of the virtual machine at the edge node which is having, let us say 0.2 and the computing cost of the private cloud is, let us say 3. So, these are the constants which are need to calculate the total cost.

(Refer Slide Time: 8:46)



Similarly, if you consider the public cloud, then in that case we assume that we have replaced this private cloud with a public cloud setting and with a new policy that outputs the public action. Now the, when you say a public cloud, then the demand will be parameterized. Let us see the parameterized policy action. So, at a time slot 1, the policy action says the k, k is the parameter. Let us say 1, 1, let us assume that 1 means the reserve instance. So, for reserve instance, 0.4 is the VM or the resource allocated.

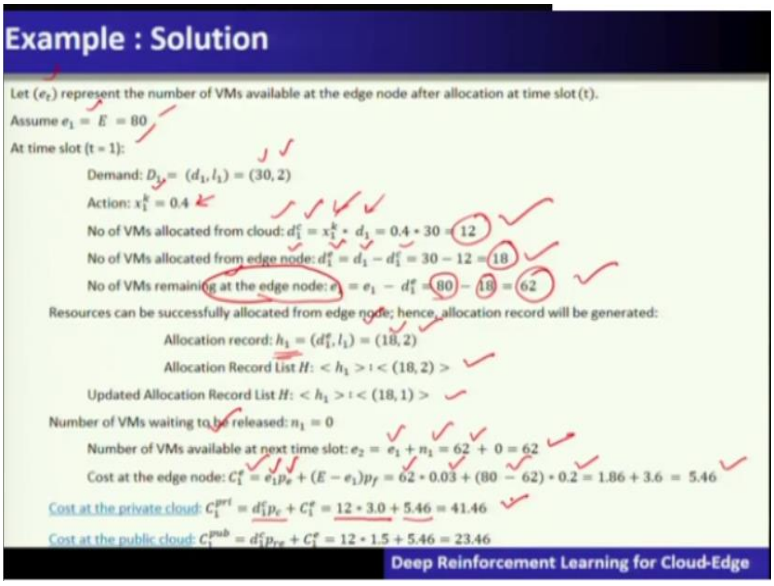
So, in that case, if you see over here the, so you have the cloud and you have the edge, and here you have the demand. So, the demand is 30, then 12 will be and 18 will be. So, 12 will be of which type of cloud; that is of reserved instance. 12 will be off reserved instance. Similarly, you can calculate for the others. So, that is what is simply explained by parameter kt of 3 different parameters, 0, 1 and 2. So, 0 means on demand.

So, the second example is of 0. So, 0 means on demand instance of the public cloud will be used here in this policy action. So, this type is called parameterized in the public cloud instance that was allocated. So, to calculate the cost of collaborative cloud side computing, which is expressed by C of public cloud, at a time t in a given public cloud setting, at each of these time slots, you require some more constants.

So, what are those constants? To calculate the cost function. So, it requires the unit price of on demand instance in the public cloud. That is P of od. Here the value is 3, unit of reserved instance in a public cloud. So, re means reserved instance od means on-demand instance that is return over here on-demand instance.

And the reserved instance is 1.5. Customization price of reserved instance is 800. And then unit price of spot instance, that is 1 unit. So, you can see that the reserved instance and on-demand instance and on spot instance has different constants or the unit prices shown over here.

(Refer Slide Time: 12:04)



Now let us see the solution of this particular example. Let us understand that et represent the number of virtual machines at the edge and we assume that the total virtual machines available at initially in the beginning is 80. So, after at the time slot 1. So, let us see that the demand for the slot 1 was d1, l1 is 30 and 2, and the action was 0.4. That means 0.4 fraction will be from the cloud. So, the number of virtual machines which are allocated from the cloud for this particular demand of a slot 1 d c is nothing but this particular fraction multiplied by the demand that 12, which we have already shown in the previous slide at this particular calculation.

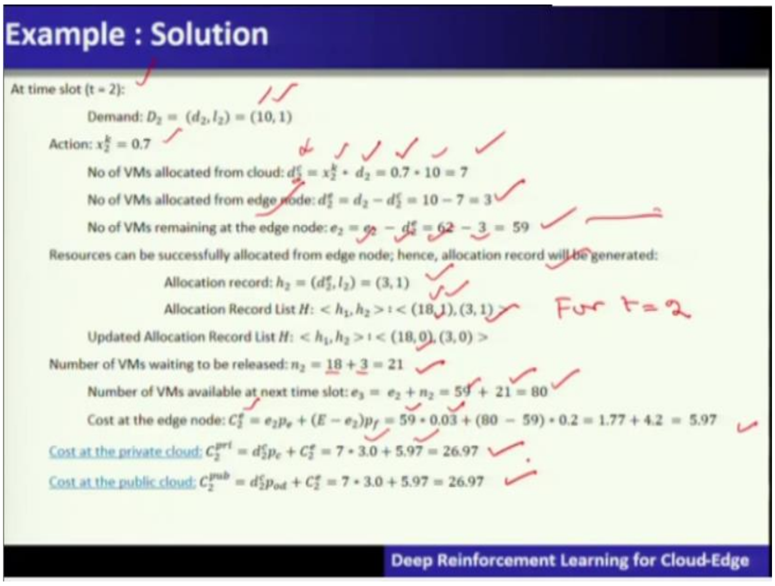
Similarly, the number of virtual machines which are calculated from that edge, that is d1, that d e at time, 1 is nothing but d1 minus d c. So, that is 30 minus 12 is 18. That also we have calculated in the previous slide and shown you along with the demand. Number of virtual machines, which are remaining at the edge node e1 that is 1 minus d1 that is 80 is a total virtual machine that is 1. And the demand, which is coming from the edge node is 18. So, the total available virtual machines at the edge after meeting this particular demand is 62.

So, the resources can be successfully allocated from the edge nodes from, hence the allocation record will be generated in this manner. So, allocation record is structure will be d e1, l1 that is 18 for 2 time slots will be needed. Allocation record, therefore has only this particular record and updated allocation record now will be for this particular time period for first time slot. So, the number of virtual machines waiting to be released is 0 in the beginning because it is the first time slot.

So, the number of virtual machines available in the next time slots that is e2 will be e1 plus n1 that is 62. So, 62 will be available for the next time slots. So, the cost at the edge node will be the e1 and pe. So, e1 is, let us say that is alone, is available 62 multiplied by that pe. That is 0.03 plus this is allocation from this is the virtual machines which are allocated that is 80 minus 62. that will be having that is 0.2, that is from the public cloud.

So, that comes out to be so therefore, the total cost at the public cloud, if you calculate, this comes out to be 12 multiplied by 3 and plus 5.46. That is 41.46. And similarly, for the other cloud is also calculated in the similar manner.

(Refer Slide Time: 15:31)



At time t2, we can calculate in the same manner that the demand for time slot 2 is 10, 1; that is 10 virtual machines for 1 time duration. And the action policy action is 0.7. So, that is the number of virtual machines which are allocated from the cloud at instance 2 at the time slot 2 from the cloud is calculated in this manner. x2 of k that is 0.7 multiplied by d2 is 10. That comes out to be 7. So, number of virtual machines remaining out of the demand 7 is 3. 3 will come from the edge.

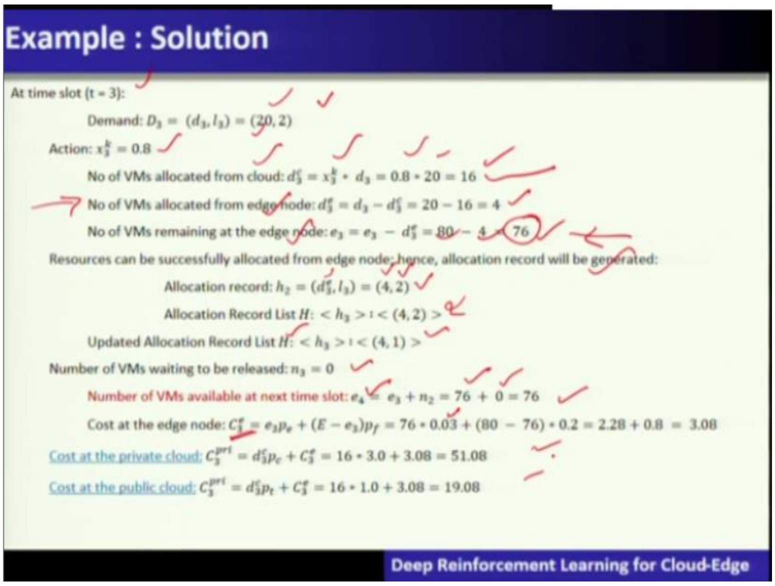
So, the number of virtual machines remaining at the edge node will be calculated here in this case e2 minus d2. So, e2, you know that only the number of virtual machines. The edge is 62 and 3 will be reserved for 3 will be allocated at times slot 2. So, 59 will be available virtual machine, the edge node. So, resources can successfully be allocated from edge node, hence the allocation record will be generated here in this case 3, 1. So, allocation record.

Now an updated allocation record will be here in this case for the second time slot for t equal to 2. So, the first demand requires that resources for 2 time slots. So, the remaining is 1 and this new 1 that is 3, 1. So, this will be the allocation record and updated allocation record will be that 18, 0 and for the for and 3, 0 for the next time slot. So, the number of virtual machines which are waiting to be released is 18 plus 3. That is 21. So, the number of virtual machines available for the next time slots will be the 21 is out of this, and 59 is from the remaining. That is 20 total 80.

So, the cost, the edge, if you calculate for the time slot 2 is nothing but e2 pe. So, you can see that pe is 0.03 and e2 is 59. So, that we have already calculated here. Plus, the virtual machines, which are allocated from the cloud that is 80 minus 59 multiplied by 0.2 is 5.97.

So, if you calculate the total cost, that is this cost from the edge plus cost from the cloud, it is 7 into 3, 7 into 3. That comes out to be the same amount, which is shown were here in this particular slide.

(Refer Slide Time: 18:28)



Similarly, for times slot 3, lets explain in more detail the demand d3 is that 22. So, 20 2 times slots is needed. Now action is 0.8. So, the action is 0.8 means that the number of virtual machines which are being allocated from the cloud is 0.8 and total demand is 20. So, the number of virtual machines which are allocated from the cloud is only 16. Now, the remaining of that particular demand that is the 4, will be allocated from the edge node. So, the number of virtual machines, which are still available out of the full capacity will be 80 minus 4. That is 76.

So, the resources can successfully allocate it from edge node, hence the allocation record will be generated. So, let us see that the allocation record is 42. So, that means 4 is the number of virtual machines from the edge, and they are to be allocated for 2 time slots. So, the allocation record will have this entry for this time slot 2, updated allocation record in capital H if you see, will have 4 and 1 and 1 will be for the next time slot. So, number of virtual machines which are waiting to be released here and 3 is 0.

The number of waiting virtual machines will be available in the next time slot. That is e4 is nothing but whatever is available and 0 that is 76. So, the cost of the edge at the time slot 3, if you calculate that, comes out to be e3. So, e3 is 76 and it is it is unit cost that is 0.03 plus this is the unit cost coming for the allocation. So, this comes out to be this one. So, with this example, we, we conclude this particular lecture.

So, we have covered how the user demand is being formalized with the help of capital D, which is nothing but a small d and le. That is the time for which this demand is required to execute the application with the resources. And then we have seen that the policy action or the action requires this allocation from the edge and how much fraction is required from the cloud and when in the cloud, whether it is the public cloud or it is a private cloud, all these things.

And then we have also shown how to calculate the cost and the cost function required to be minimized. That is what is the objective, and that is being used with the help of deep reinforcement learning for a large, discrete and continuous state space. Thank you.