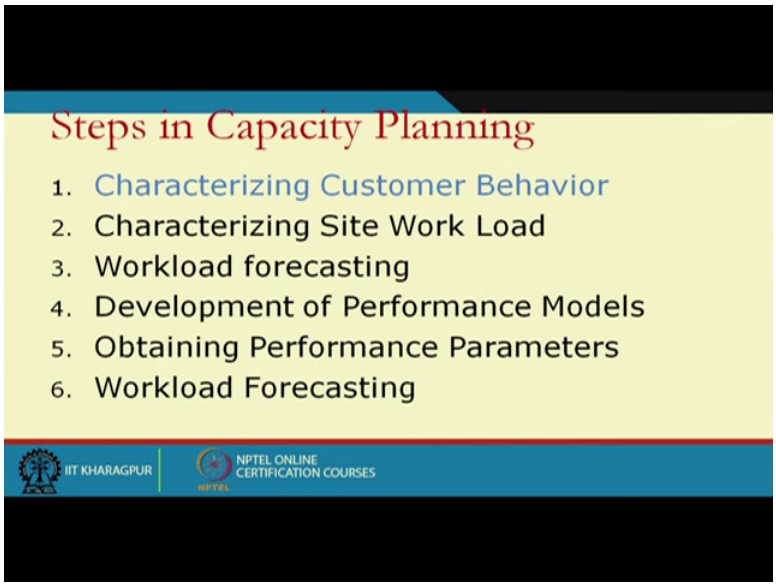
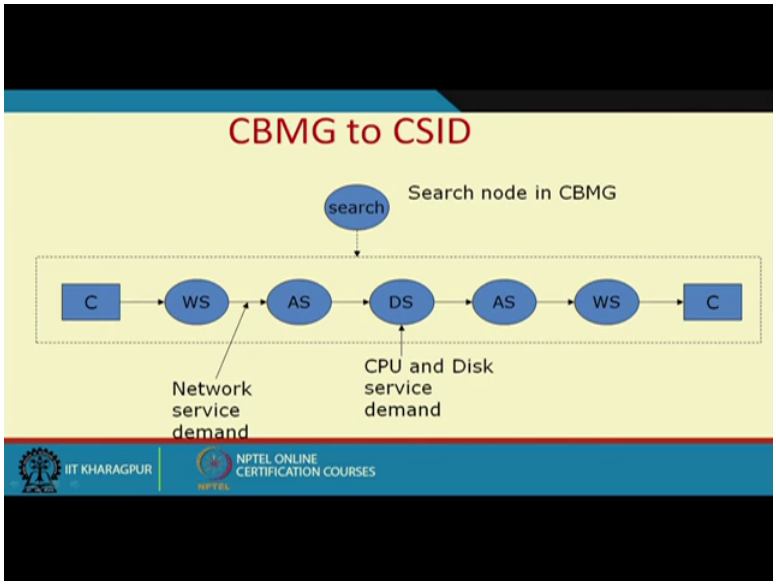
E Business
Professor Mamata Jenamani
Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Lecture-51
E-Business Capacity Planning (Contd)

(Refer Slide Time: 0:31)



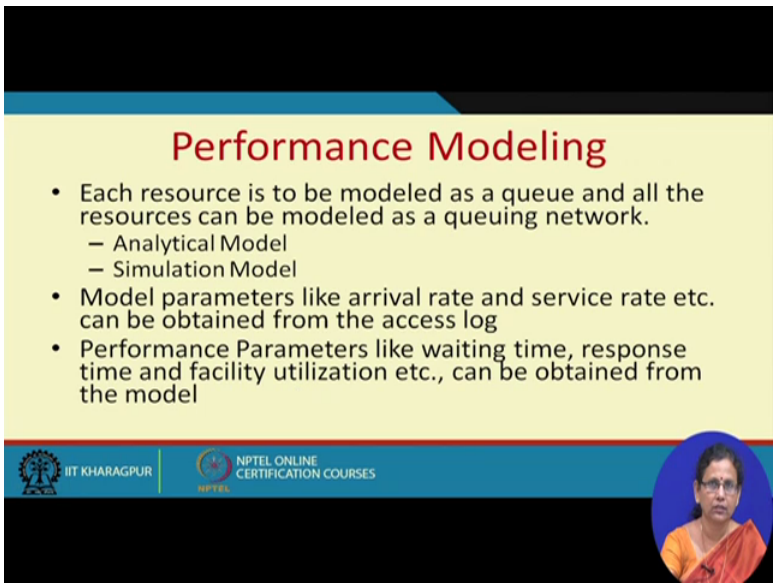
Welcome back, today we are again going to continue our discussion on E business capacity planning. In fact, we already know from the last class that the steps involved in capacity planning are character having customer behaviour which we did some 2 classes back, then characterising site workload which also we did it and we also saw one queuing example for that, then workload forecasting, development of performance model, obtaining performance parameters, et cetera and the other steps.

(Refer Slide Time: 1:02)



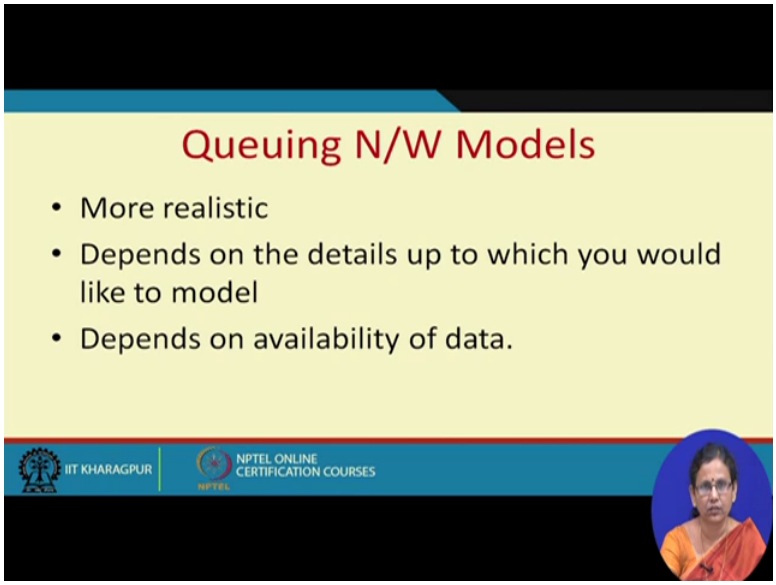
So here we continue with that discussion when we were trying to see that whenever we go for performance modelling whether our site is capable of handling the load or not, each of the functions that is offered by the site we can actually break it down into a number of sub functions in terms of how the plant and server interact so here this plant-server interaction has been represented in the form of plant-server interaction diagram, and we saw last class that how to use a queuing model to solve this to find out various performance parameters.

(Refer Slide Time: 1:48)



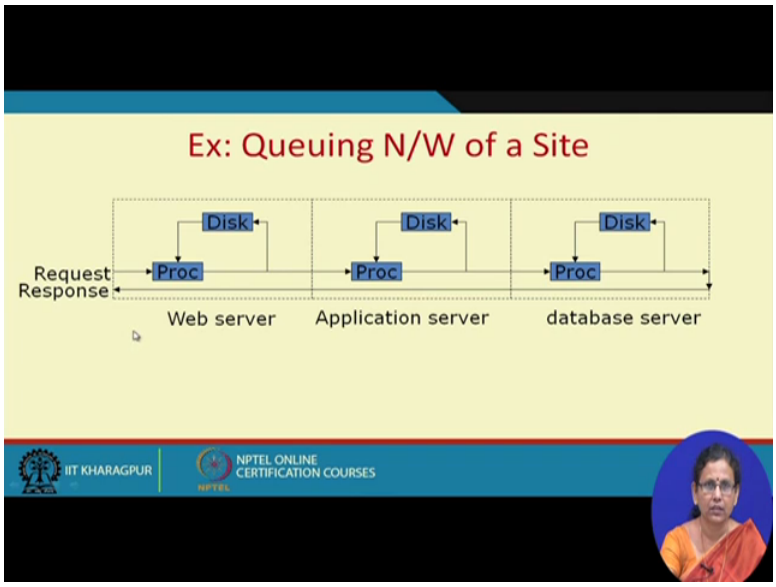
Now here during this performance modelling, each resource has is to be modelled as a queue and all the resources can be modelled as queuing network. So last class we saw one example of a queue single server queue.

(Refer Slide Time: 2:10)



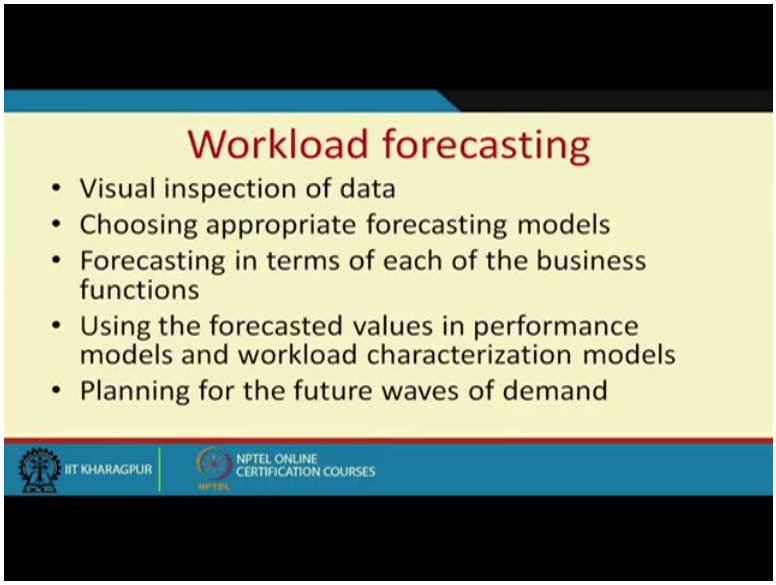
But in case you have more number of resources then it is more practical to represent it in the form of a queuing network.

(Refer Slide Time: 2:19)



So in a queuing network you can even go further detail by showing how the interaction within each of the servers takes place (())(2:28) and so on. In case this is just one example, in case you have multiple sites, multiple mirrored side then the network is going to be even more complicated. But anyway, as I have told you our focus is not on teaching you about various kinds of mathematical models involved, it is just to show the applications of those models taking few examples, if you are interested you can always take your course related to that in this case for the queuing model.

(Refer Slide Time: 3:05)

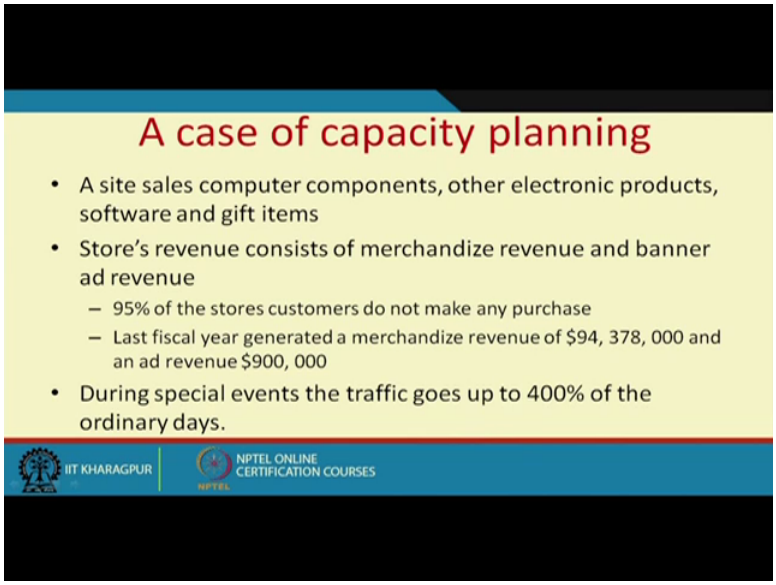


But anyway, the next task is your workload forecasting. During your workload forecasting the data again has to come from the log file, your log file contains the last request made to the site, all the requests so far made to the site so analysing this you can actually again forecast the workload. Here you will be using various forecasting models but we are not going to discuss about the forecasting model you can if you wish you can read about them. But the steps involved are first you have to visually inspect the data, you have to choose appropriate forecasting model and this forecasting model in terms of each business function you have to see like which business function is overloaded, which business function is not.

It may so happen in the specific period of particular business function is getting overloaded. So if you break this demand on each of the specific function, demand in the sense the number of request in terms of specific functions in different periods of time you can even forecast make the forecast not only at a very gross level where you consider all the functions together, but you can also make the forecast where you are making the forecast for individual functions let us say search product search, when the load on product search is very high, payment; when the load on payment function is very high and so on.

So using these forecasting values in performance models and workload characterisation models you can actually plan for the future waves of demand, in this context we are going to see one example.

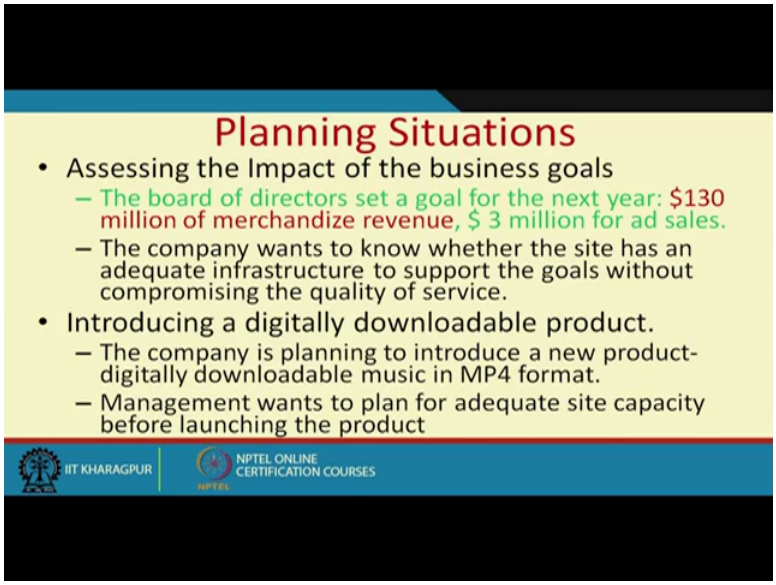
(Refer Slide Time: 5:07)



In this particular case, a site sales computer components and other electronic products, software and various gift items. Now the store's revenue consists of merchandise revenue and revenue from banner ad. By merchandising me we mean by selling they will be by selling the items they will be getting some benefit that we say it is a merchandise revenue. Similarly, by banner ad showing various banner ads in their site they also get some they also earn something so that we call as banner revenue. So the situation goes like this, 97 percent of the store's customers do not make any purchase and the data shows that the last fiscal year generated a merchandising revenue of some 94 million dollars and an ad revenue of some 900,000.

So anyway whatever is irrespective of the value, this is the merchandise revenue, this is the ad revenue. Now during special events the traffic goes up to 400 percent then that of a ordinary day like Diwali or let us say Christmas during some even the traffic even goes up in the site okay.

(Refer Slide Time: 7:06)

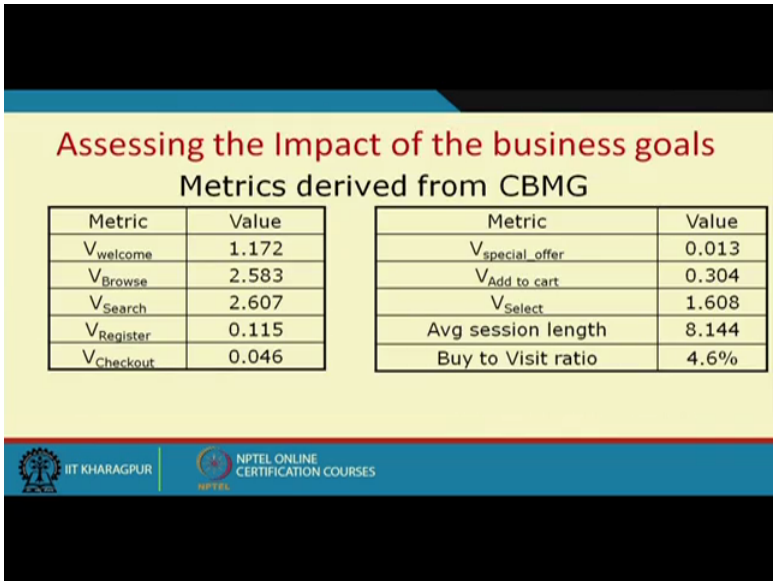


Now in this capacity planning situation it faces many problems, one such problem is one such problem is it has its board of director is setting a goal of certain merchandise revenue and certain ad revenue for the next fiscal year. Now the problem is how do you figure out whether your site will be able to take the load or not. Similarly the second situation it may so happen that the site who was otherwise selling software products is now introducing a new product which , earlier it used to make provision for downloading only the software products, now it has provision for downloading music from its site. Now depending on the expectations of the management how should you plan for your workload in the site and whether your infrastructure is ready or not, how do you evaluate it.

Anyway, this can be some 2 situations for the particular case and of this we are going to look at the first one and see that how the data generated from access log and the the management's expectations can be used to figure out whether the site is ready and whether or whether the infrastructure is ready to handle the future wave of (())(8:29). Now, for what is the management sector expectation? Management expectation means some 130 million merchandise revenue and 3 million ad revenue from the website.

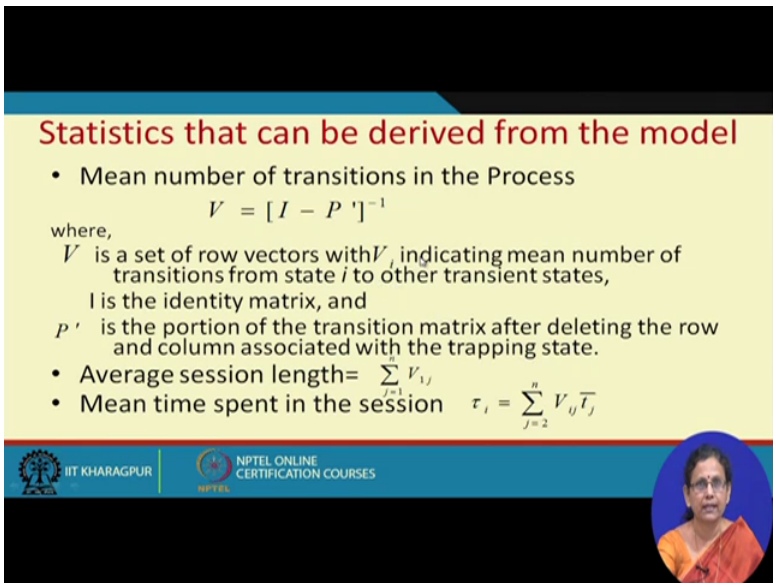
And what was the earlier? Earlier it was 94, from 94 you are going up to 130 and here it was 0.9 and here you are going up to 3 million. Now what should the company's strategy be and how with this management's expectations what should be the company's strategy be for the infrastructure, so let us try to see from various data sources how we are calculating it.

(Refer Slide Time: 9:44)



Now look suppose from the customer behaviour model class these matrix are derived, what are the matrix? The average number of visits to each phase and as a result the average session link, within each session how many pages and all together visited. So that how many pages are all together visited within that session that many number of download page requests are going to come on an average. And this is your buy to visit ratio, how many people came to the site and of them how many actually made the purchase.

(Refer Slide Time: 10:26)



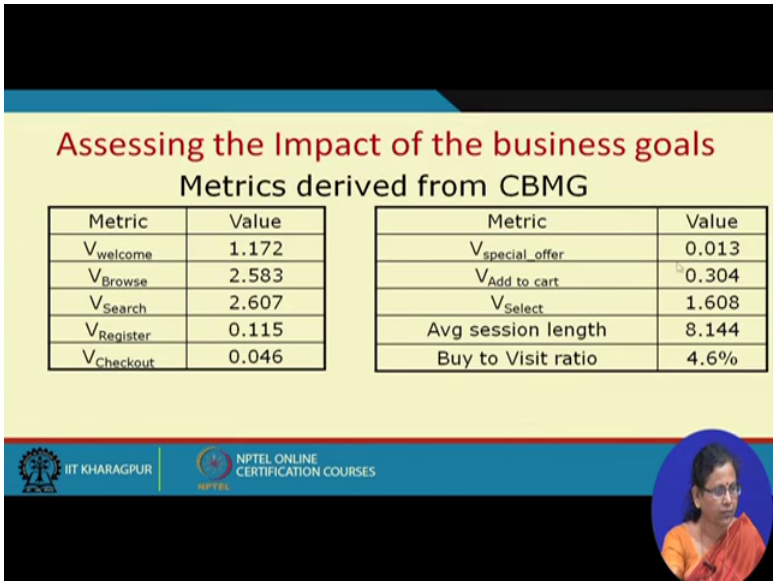
So to remind you be derived this when we discussed about the customer behaviour model, we saw that we could this slide we have already seen in one of the earlier lectures. Here, this V is the average number of transitions in the process the average number of visits to each state

from one of the earlier states, so this V is a set of Rho vectors. But anyways irrespective of the calculation we know that this V represents the from the model only we could derive the average number of visits to each page and then we are using this number for our further calculation. So you must imagine that the steps that was involved in the capacity planning how they are interrelated and how they can be used in a very co-ordinated manner to solve certain business problems?

Look at this situation, what was your target? Your target was your target was 130 million US dollar for sales revenue. Now, how many purchases are required to generate this specific revenue that you have to figure out, then with this many number of purchases how many unique sessions will be required that also you have to find out okay. Now what is the revenue throughput? Let us first find out what is the revenue throughput. Now what is the revenue throughput? Revenue throughput is the dollars earned per second. So Management's expectation was 130 million dollars per year, so per second it is 130 million divided by number of seconds per year that turns out to be this number 4.122 so this is your expected revenue throughput, next year you want this much as your revenue throughput.

Then your observed value of average sales is given to be dollar 225, so your revenue throughput that is earning per second is actually equal to sessions per second, how many sessions have been executed per second into buy to visit ratio into average sales. Now buy to visit ratio is multiplied here, now out of all the sessions only few sessions involved the selling of the product, not everybody who is coming to the site is actually buying is coming to the payment state and making the purchase, only few of them will be making the purchase, now what is that number? How many percent?

(Refer Slide Time: 14:18)



So that is given here that buy to visit ratio is 4.6 percent, so with this 4.6 percent buy to visit ratio and this is your average session length. So out of on an average and everybody will be visiting this many number of pages per session and out of that this many percent will be actually making the purchase, so number of sessions per second into buy to visit ratio into average sales. So revenue throughput is this one that we have calculated that is dollars earned per second and this is will also can be calculated from this problem that is sessions per second into buy to visit ratio into average sales. So revenue throughput be have already calculated 4.122 now this, so therefore putting 4.122 here your number of sessions per second is this 4.122 divided by buy to visit ratio into average sales.

This average sales is again given from is calculated from the past sales data okay. Now with this it is understood that the sessions per second required to fulfil this desire of earning 130 million dollar turns out to be 0.398 sessions per second okay. So thus the server if it has to fulfil the management's expectation has to actually deliver 0.398 sessions per second, now with the current infrastructure whether it is possible or not that is we need to figure out okay. Now with these many sessions per second, so how many requests now will be made to the server? What should be your system throughput? So expected system throughput is the number of sessions into average session length, now what is the average session length?

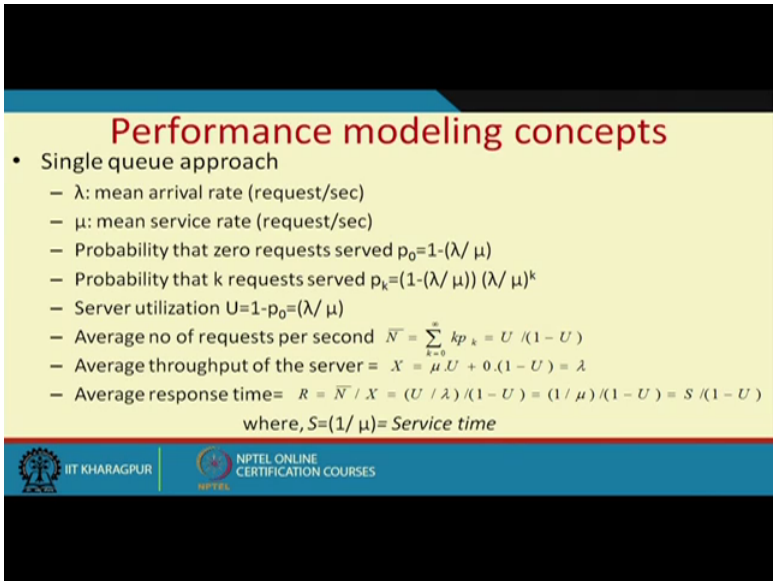
Average session land we have understood from here, average session length is 8.144, so 8.144 into the number of sessions per second. So which means there has to be 3.241 transactions per second okay. This is in general, but sometimes the traffic will be too high,

from past observation it is understood that the observed traffic burst whenever there is some special selling season also, it will be some 20 times from the normal load. So if it has to be 20 times from the normal load, the site has to serve these many requests per second on a very high traffic day. So now this many requests per sec and if the your site has to deliver, is your site ready? That is the question.

Now what we have to figure out is, we have from the expectations of the management somehow we are able to calculate the value that how many requests per second the site has to deliver. Now considering again your performance model where we consider the arrival rate to be lambda and service rate to be Mu, your service rate remains constant because if you are going with your own infrastructure, your service rate will be your constant and your arrival rate is 64.82. So if you and how do you get your service rate? Your service rate you will be obtaining from your past data which is available from the access log, on an average how much time the server is taking to deliver a page from that data you will be calculating and the total if you consider the total number of requests served during that period, you can calculate the average time the server takes to serve.

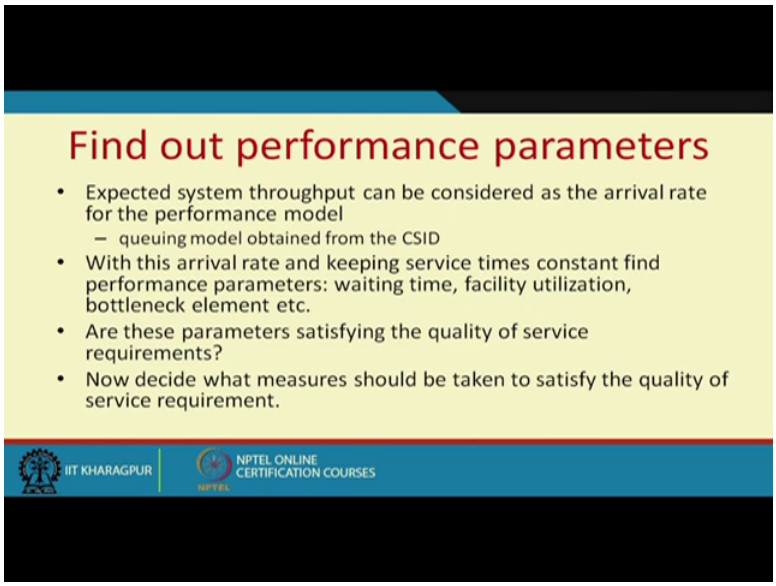
In fact, one example in this connection we have already seen in the last class. We are not going to discuss but in this similar setting you have to first find out what is the what is your Mu value, what is your service rate by your current infrastructure and given this Mu value that is 64.82 requests per second now you need to figure out what is your whether your site will be capable of taking adequate load or not. Here we are not going to further details because similar situation we have already encountered in the past the last class so we are not going to show this detail, but using that queuing model in fact we can show I can show you that queuing model which I am talking about, this is the one.

(Refer Slide Time: 20:53)



This is the one but this is again the example right now we are considering is different but this formula will remain same, here the lambda will be the value 64 point something that we calculated, Mu will be the service rate which has to come from the access log data then one has to find out the server utilisation. And as I have told you the server utilisation indicates how loaded the server is and if you learn the theory of queuing model, a server load if it approaches 1 which means the server is going to break down so it should be almost I mean if it is some 0.5, 0.6, etc it will be really good. So therefore continuing with your current example, if we use those kinds of formulas and figure out whether your infrastructure is ready for realising your business goal or not okay.

(Refer Slide Time: 21:59)



Now this is what I was telling you, the expected throughput can be considered as the arrival rate for the performance model that the queuing model that we are discussing. Then with this arrival rate keeping the service time constant which is again derived from your access log data, one has to find out the performance parameters like your waiting time, facility utilisation that is your server utilisation, which is the bottleneck element and so on. And with these values you should figure out the quality of service expectation that your management kept whether you are able to satisfy it or not, if not then you have to take adequate steps. Now what kind of adequate steps you can take?

If you feel that your server is going to be overloaded that utilising server utilisation value is somewhat let us say above 0.7 or so, then what do you do? You will be either scaling up because this is your expectation, this is not a short-term goal that like we were discussing last class, this is not a short-term goal only they selling season I am just hosting a offer period of let us say some 3 days, within these 3 I will be having some selling festival, so therefore during this period I can hire certain facilities, certain other servers and I can somehow manage for this period but this is a different situation altogether.

In this situation, we are actually planning for the next year, so the whole year we are going to see this kind of situation so therefore either we can scale up or scale out. By scaling we mean we increase the capacity of the server, why scaling out we mean we can have more number of servers to deal with the situation may be cluster of servers so that is again a management decision okay. So if you find out these parameters then you can see whether the quality of service requirement is satisfied or not, so now you have to decide the measures that should be taken to satisfy the quality of service requirement, so with this we finish this lecture and from next class onwards we are again going to start some interesting topic, thank you very much.