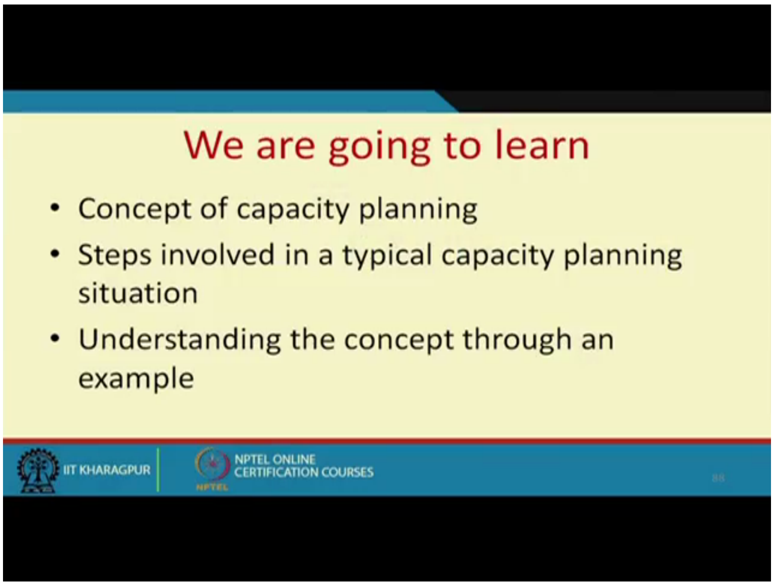
E-Business.
Professor Mamata Jenamani.
Department of Industrial and Systems Engineering.
Indian Institute of Technology, Kharagpur.
Lecture-50.
E-Business Capacity Planning.

Welcome back, we continue our discussion on how to use the access log for getting business insights and taking some decisions there off and this is a part of the series of lectures we are going to have on providing decision support in online environment to the managers. So the next topic that we are going to look plan how to, what is the capacity, how do you plan for the capacity of your e-business infrastructure, whether to, whether your capacity is adequate or not. So in this lecture, we are going to see the concept of capacity planning and going through few examples which will tell you, which will give you insights that how this access log data and the data from few other sources can be used to decide on this thing.

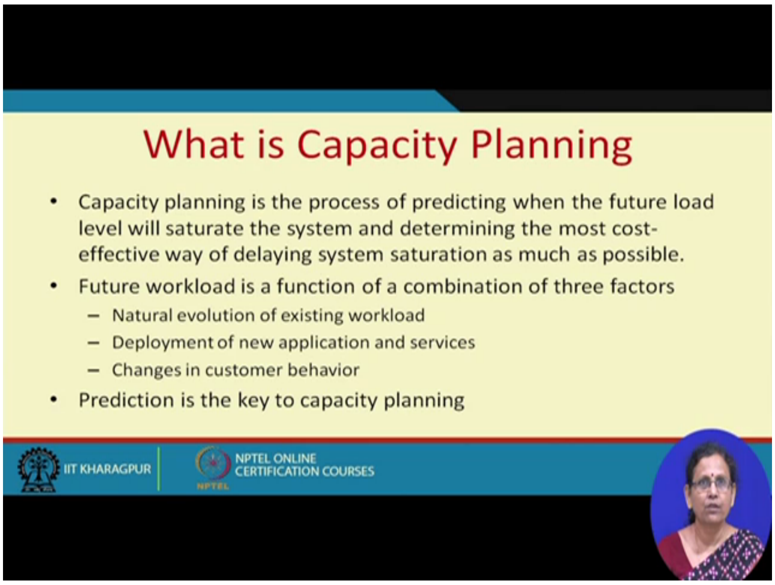
(Refer Slide Time: 1:31)



We are going to learn

- Concept of capacity planning
- Steps involved in a typical capacity planning situation
- Understanding the concept through an example

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES



What is Capacity Planning

- Capacity planning is the process of predicting when the future load level will saturate the system and determining the most cost-effective way of delaying system saturation as much as possible.
- Future workload is a function of a combination of three factors
  - Natural evolution of existing workload
  - Deployment of new application and services
  - Changes in customer behavior
- Prediction is the key to capacity planning

Again here we are going to adopt, we are going to see various modelling procedures can be used for this capacity planning. So start with, capacity planning is the process of predicting when the future workload level will saturate the system and determining the most cost-effective way of delaying the system saturation and as much as possible. What does it mean, it is the, you will be predicting what should be your future workload based on your past data. Then when you realise that your system is going to be saturated, it will be no, what is the meaning of saturation here, it is not possible to handle the load that is coming to your website.

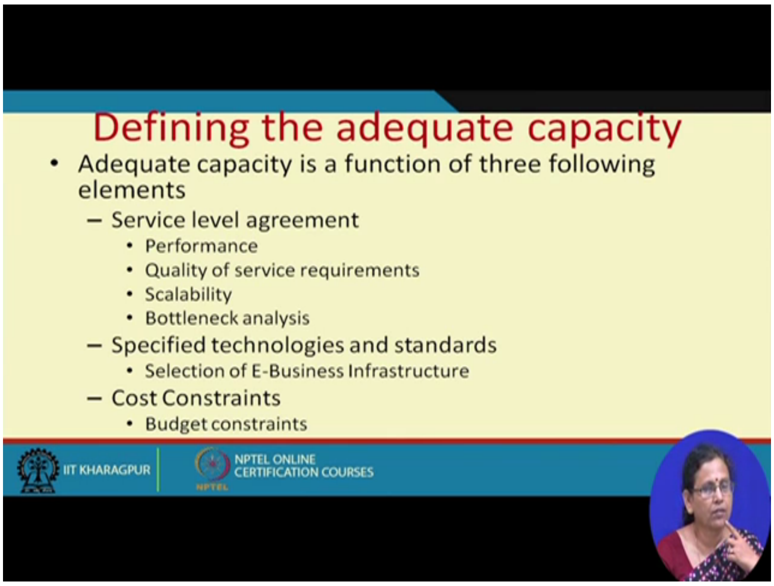
So therefore what should you do? What should be your strategy so that, because investing in the infrastructure itself is a, is a decision-making situation which, I mean, the more you avoid, the more you delay the saturation level, you can delay your investment process as well. So how do you, how do you plan for delaying the saturation? Now look here, your future workload, whatever will be coming to your site is actually a function of 3 things. 1st it is a natural evolution of the existing workload. What do you mean by natural evolution of existing workload?

As time progresses, your website becomes well-known, more and more people will be coming to your site, so it is by natural evolution, more and more load will be coming to your site. Deployment of new applications and services you attract more customers. Changes in customer's browsing behaviour because, you see, we have discussed that customer behaviour is a random process, it is a stochastic process. So therefore it may change over the time and you may not be the reason, you may not know the reason. Maybe, and sometimes you may also know the reason, for example let us say Diwali time, more sales are happening, so more customers will be coming to your site.

So if you expect such kind of situation and you are, you are planned for this, then possibly you can avoid crashing of your server at the peak time of your transaction. If you remember, just few years back possibly, some e-commerce site announced for some annual sale, people were coming in the large numbers and the website was crashing, even I remember putting few things, I mean the large number of things in the shopping cart, the site crashed and we lost all our information and then we were, that time we were free, you are trying to buy and the next hour we become busy and we did not, we decide not to purchase.

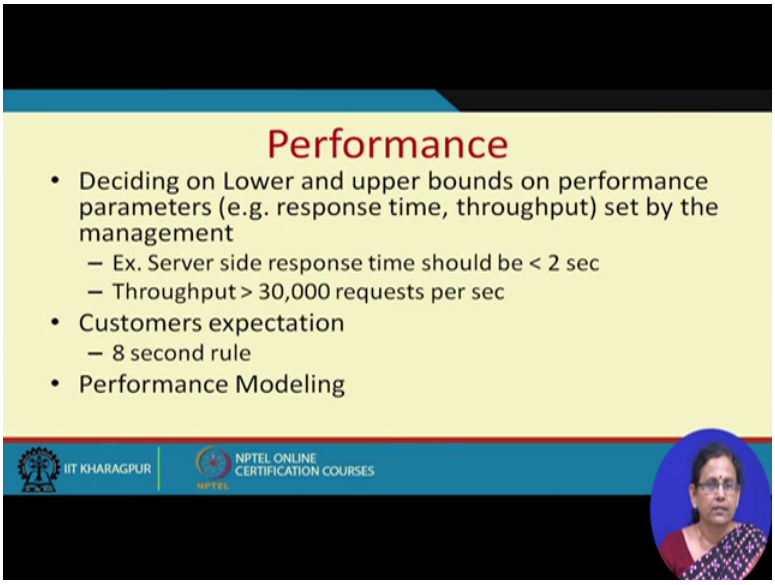
So some, like me, like us there will be many such buyers, so therefore, therefore having adequate capacity to cater to your customers especially during the peak time is a decision-making situation and you require a lot of, a lot of thought process should go into it, it is a more of, it is a combination of more strategic level activity along with some tactical level activities. Okay, let us try defining adequate capacity.

(Refer Slide Time: 6:15)



Defining the adequate capacity

- Adequate capacity is a function of three following elements
  - Service level agreement
    - Performance
    - Quality of service requirements
    - Scalability
    - Bottleneck analysis
  - Specified technologies and standards
    - Selection of E-Business Infrastructure
  - Cost Constraints
    - Budget constraints



Performance

- Deciding on Lower and upper bounds on performance parameters (e.g. response time, throughput) set by the management
  - Ex. Server side response time should be < 2 sec
  - Throughput > 30,000 requests per sec
- Customers expectation
  - 8 second rule
- Performance Modeling

So adequate capacity is a function of 3 things, 1st the service level agreement, service level agreement which is basically your performance, quality of service requirement and scalability and bottleneck analysis. In fact, some of the discussions we had in the beginning when we started talking about the e-business infrastructure. The 1st one, that is service level agreement is about deciding what should be your performance. This performance is about deciding the lower and upper bound of various performance parameters like that of response time, throughput, etc.

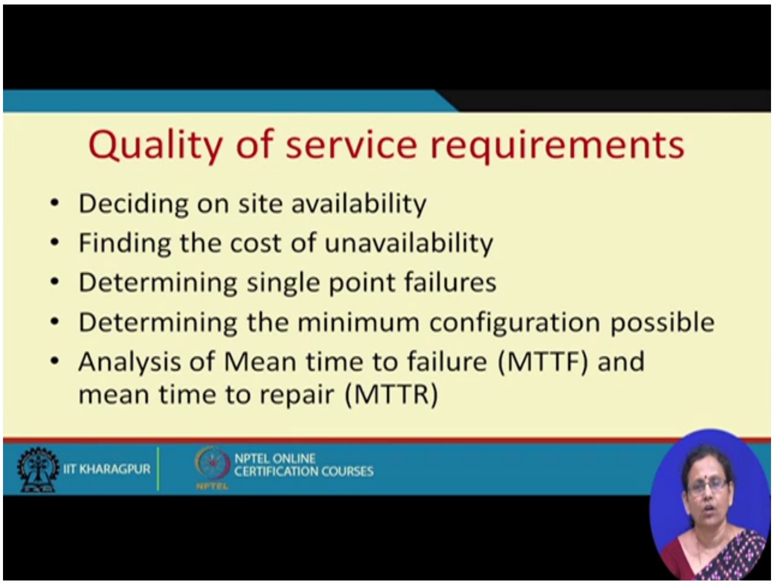
So it is again your, this is about the management to set depending on the kind of business that you are in whether you are an online e-commerce, you are an e-commerce site for whom the response time is very important, otherwise the customers will go. Suppose you are an online publisher, publishing research papers, research papers the users of that research paper, the reader of that research paper is going to wait even if your response time is little low because it is the requirement. But if you are an online shopping store, then there are many more shopping stores like you.

Unlike your digital library where the unique product, that is the research paper is the unique product which is not available anywhere, here if you are online shopping site, there will be many more shopping sites dealing with the same product or at least a substitute. So people will go, so depending on the kind of business you are in, this lower and upper bound can be set. Let us say, somebody decides that server side response time should be less than 2 seconds or and throughput should be some 30,000 requests per second.

By deciding this really does not help, you have to plan accordingly, you have to plan your infrastructure accordingly so that your response time is within the bounds which your management decides. So here one important thing can be, you can think of one important thing, that is your customer's expectations. Now this customer, there is an industry rule called 8 second rule. The customer is your response time is beyond 8 seconds, they are not going to stay, they are going to go away. So depending on the uniqueness of your product you can now set this time.

So in this example the set is, the time is decided to be less than 2 seconds, that is a good number for a server side, because there is a response time, server, as we have discussed to building infrastructure, this response time is not limited to the activities of the server alone, this includes the network delay time, this includes the client side, the capability of the client as well. But at least from business perspective, we should be clear that our server side response time is not very less. Okay, so for this purpose you need to do some kind of performance modelling.

(Refer Slide Time: 10:46)



Quality of service requirements

- Deciding on site availability
- Finding the cost of unavailability
- Determining single point failures
- Determining the minimum configuration possible
- Analysis of Mean time to failure (MTTF) and mean time to repair (MTTR)

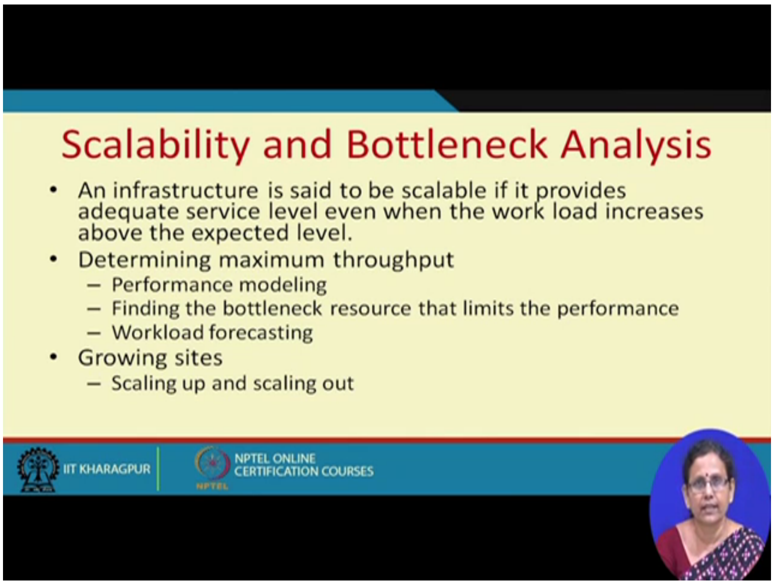
Then next is your quality of service requirement. What is your quality of service requirement? Deciding on the site availability, while deciding the whether the site will be 100 percent available or not, in fact this example we have discussed in earlier time as well, in early lectures as well where we took probably one example of again this shopping site and digital goods, digital, your digital library. In a digital library, like that of your science direct or IEEE or strangers which publish research papers and books, their products are unique.

That paper will not be published elsewhere, so if it is a very important paper, if the site, if you find tonight that site is under maintenance, they write, from time to time if you are visiting such sites, you might be finding that site is under maintenance. They know that even if they put complete 12 hours, their site under maintenance for complete 12 hours, their prospective customers who are interested in that unique products are going to come back. But think of a shopping, online shopping site.

Today if the product is, if the site A is not available, its server is down, you are going to go to site B, because their products are not unique. And same product is available in both the places, so while deciding the site availability, estimating the cost of unavailability, that is the cost of losing the customer is very important, so that is again a modelling situation. Then finding the singlepoint failures, those are your bottleneck elements. Let us say you have one web server, multiple database servers. If one of the database servers goes out of service, the same data is replicated in other servers as well, so you are, that is not a bottleneck.

But if you are web server, if you have a single web server goes down, the site is not available. So therefore finding such singlepoint failures is very important, which are those elements, if they fail your service will be interrupted. Then to have this , whatever performance parameters you decided, what should be your minimal configuration to have these kind of expectations fulfilled, expectations of the management fulfilled, that is again a decision-making situation. Your, for doing all this, analysis of meantime to failure, analysis of meantime to repair, etc., in fact, quite a bit of reliability analysis can go into it.

(Refer Slide Time: 14:51)



Scalability and Bottleneck Analysis

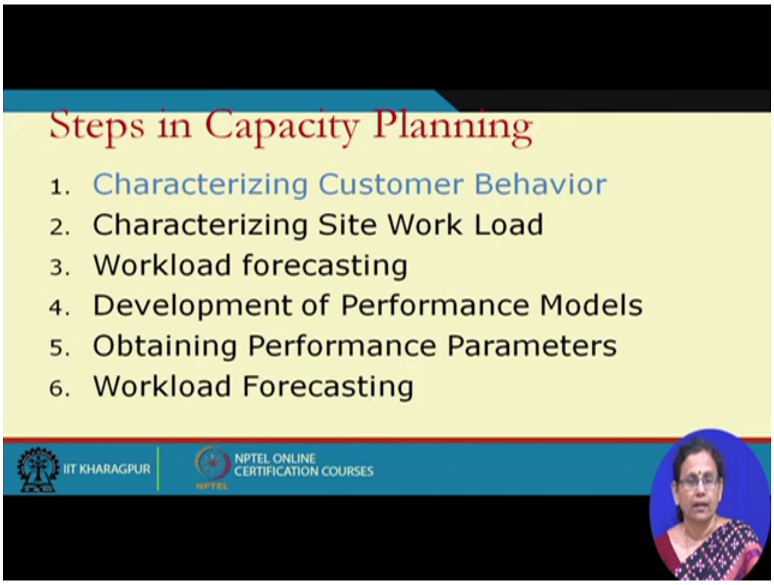
- An infrastructure is said to be scalable if it provides adequate service level even when the work load increases above the expected level.
- Determining maximum throughput
  - Performance modeling
  - Finding the bottleneck resource that limits the performance
  - Workload forecasting
- Growing sites
  - Scaling up and scaling out

Scalability and bottleneck analysis, an infrastructure is said to be scalable if it provides adequate service level even when the workload increases above the expectation level. Let us say in your website, a particular section is going to be accessed more in a particular selling season. Let us say for example, for example let us say during winter the chance of winter clothes getting sold is more. So which means if you are now devoting, I mean if you are actually dividing your load among a number of servers, more number of servers will be devoted for the search pertaining to winter products.

Okay, and maybe in Summer you can change that setting, so you should have adequate capability for understanding these scalability issues and taking appropriate actions. So an infrastructure is said to be scalable if it provides provides adequate service level even when the workload increases above the expectation level and determining the expected throughput during that period through performance modelling, finding the bottleneck resource that limits the performance and workload forecasting.

In fact about finding this bottleneck, while discussing, if you remember, just to give you little bit more insight, if remember, when we were talking about security issues, that time we saw, because of these security protocols how how the performance can be degraded and how it can be improved by adequate arrangement. If you have not gone through that lecture, please go through to understand the, to relate this concept better and workload forecasting. And you should have adequate if your workload is increasing, you should have adequate capacity either to scale up to scale down scale out.

(Refer Slide Time: 17:34)

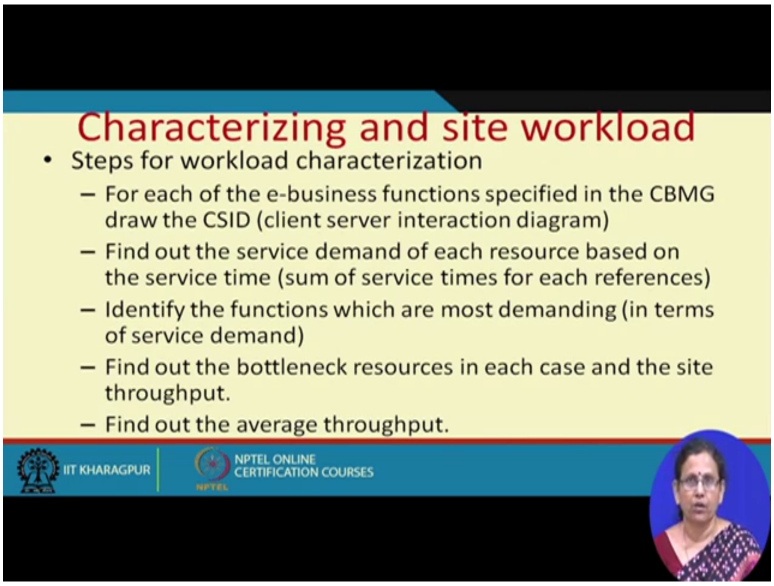


Steps in Capacity Planning

1. Characterizing Customer Behavior
2. Characterizing Site Work Load
3. Workload forecasting
4. Development of Performance Models
5. Obtaining Performance Parameters
6. Workload Forecasting

By scale up we mean by increasing your server's capacity and by scale out, by scale out we mean by adding more number of servers to your cluster. Okay. So these are the steps in capacity planning of which 1st steps we have already understood, that is characterising customer behaviour. Then characterising site workload, forecasting, etc., those modelling details we are not going to go, but we are going to see one example that how those, how calculations of those can help in taking a decision.

(Refer Slide Time: 18:23)



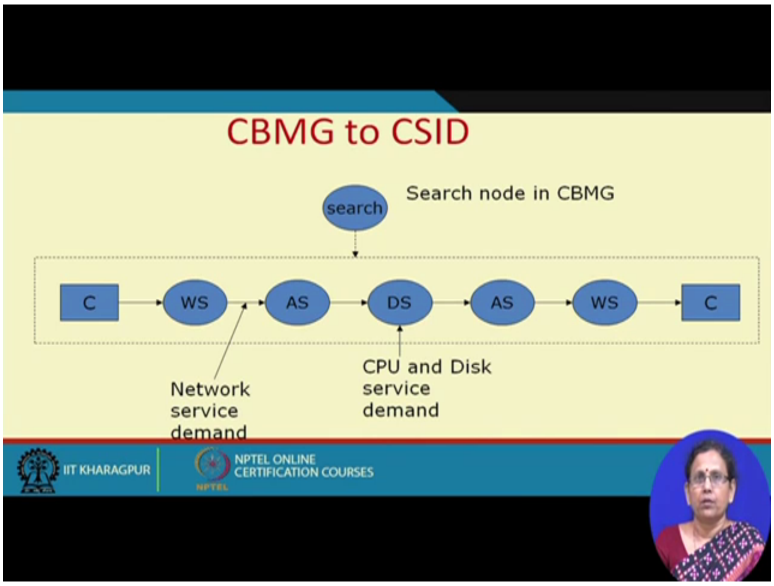
Characterizing and site workload

- Steps for workload characterization
  - For each of the e-business functions specified in the CBMG draw the CSID (client server interaction diagram)
  - Find out the service demand of each resource based on the service time (sum of service times for each references)
  - Identify the functions which are most demanding (in terms of service demand)
  - Find out the bottleneck resources in each case and the site throughput.
  - Find out the average throughput.

So these are the steps involved in workload characterisation. For each of the e-business function specified in the customer behaviour model graph that we did in the last class, you can even go deeper to see how each of the function is get executed. So in fact we can draw a client/server interaction diagram, this client/server interaction diagram we have also used at the time of, while discussing about the infrastructure, specifically your dealing with security issues.

Then we can find out the service demand and each of the resources based on the service time, then we can identify the functions which are most demanding in terms of service demand. Then we can find out the bottleneck resources in each case and the overall site throughput. And then we can find out the average throughput of the website.

(Refer Slide Time: 19:33)



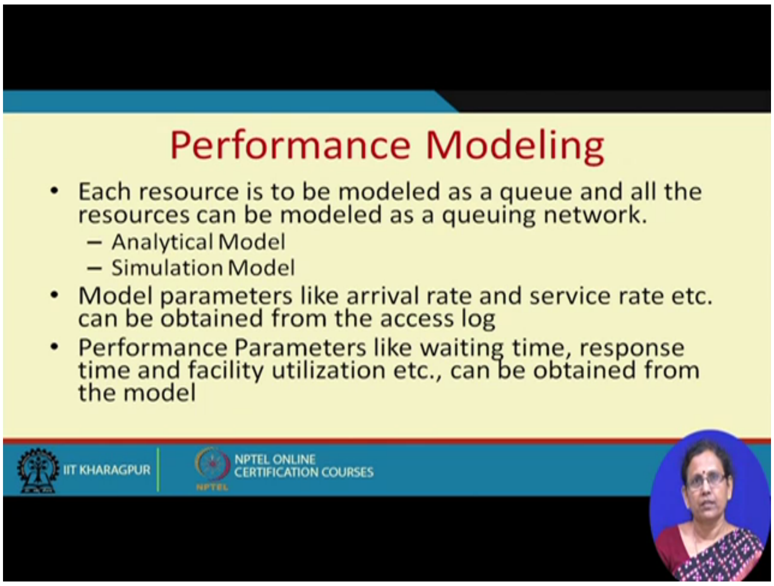
CBMG to CSID

Search node in CBMG

search
C
WS
AS
DS
AS
WS
C
Network service demand
CPU and Disk service demand
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So we can go little deeper, from customer behaviour model graph, let us say you are thinking of a node called search, searching the product. Now searching the products involves interaction among many entities, web servers, application servers, database servers and so on, this is a whole set of client/server interaction diagram. For a search operation, client 1st connects to the web server, then web server connects to the application server, application server connects to the database server, database server gives data back to the application server, it is sent back to the web server, the web server again constructs the,combines all these details, reconstruct the page and sends to the customer.

(Refer Slide Time: 20:46)



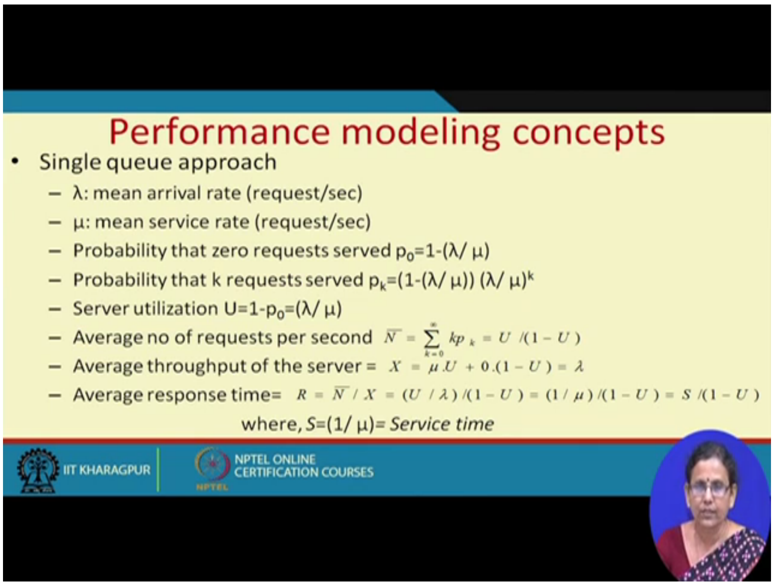
Performance Modeling

- Each resource is to be modeled as a queue and all the resources can be modeled as a queuing network.
  - Analytical Model
  - Simulation Model
- Model parameters like arrival rate and service rate etc. can be obtained from the access log
- Performance Parameters like waiting time, response time and facility utilization etc., can be obtained from the model

So now each of the element involved in this process, they have their own capacity. And depending on the load, they behave in different manner. Now modelling this process is called the performance modelling. Now each resource is now is to be modelled, specifically for this particular purpose people use queuing model. So each resource is modelled as a queue, can be modelled as a queue and all the models together can be modelled as a queueing network. Now see, we are not going to learn about queueing theory but what I am trying to tell here you is to have performance modelling or modelling of this kind of situation, you can use these queueing models.

We are not going to learn about the models, like last time we did not learn about, learn about the Markov process in mathematical details but at least we knew some of the formulas which can be used for modelling purpose. So here also we adopt the same philosophy, we will be, from the model will be finding the performance parameters like waiting time, response time, facility utilisation, etc. and can be can use it.

(Refer Slide Time: 22:03)



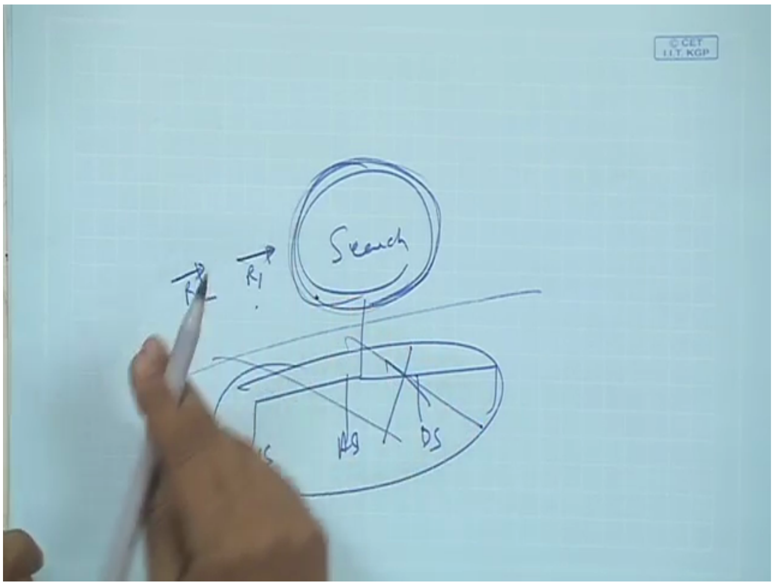
Performance modeling concepts

- Single queue approach
  - λ: mean arrival rate (request/sec)
  - μ: mean service rate (request/sec)
  - Probability that zero requests served $p_0=1-(\lambda/\mu)$
  - Probability that k requests served $p_k=(1-(\lambda/\mu))(\lambda/\mu)^k$
  - Server utilization $U=1-p_0=(\lambda/\mu)$
  - Average no of requests per second $\overline{N}=\sum_{k=0}^{\infty} kp_k = U/(1-U)$
  - Average throughput of the server = $X=\mu.U+0.(1-U)=\lambda$
  - Average response time= $R=\overline{N}/X=(U/\lambda)/(1-U)=(1/\mu)/(1-U)=S/(1-U)$

where, $S=(1/\mu)=$ *Service time*

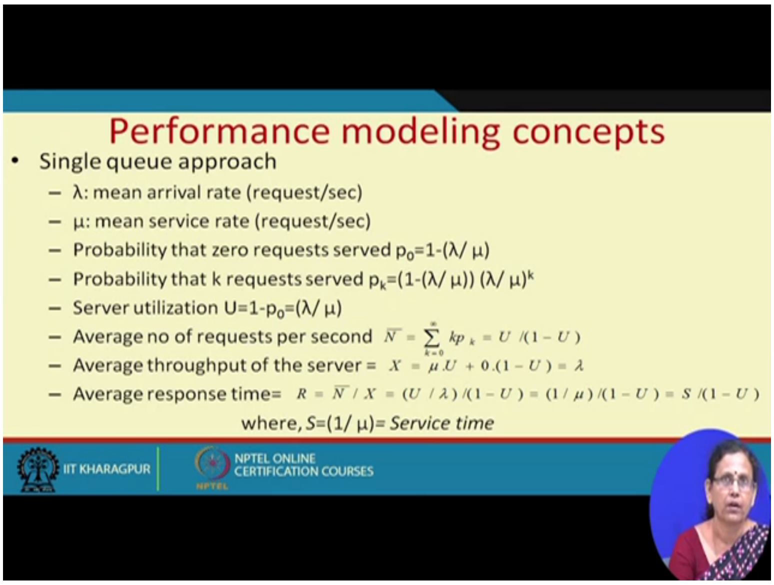
Assuming that your whole setup for let us say for that search operation, instead of dividing it into into the parts like web server, application server etc., we can go to that deeper as well. Why that deeper, even you can go to the processor level detail but it is not required in this case, it is after all modelling, modelling works under a lot of assumptions. So our assumption is we will not be breaking the operations. For a specific operation search, internally whatever may happen, we will be considering the search as the operation as a single node and we will be considering all the requests coming to that node as a queue.

(Refer Slide Time: 22:55)



So your search is the operation and the requests are coming in a queue. Now when they are solved by this search operation, this search operation maybe involving interactions between web servers, application servers and database server, but we are not considering these in this detail. What we are considering is that search as one operation and these requests are coming. Assuming that it is a situation of this time, where we are not considering these details, this is a model with a single queue.

(Refer Slide Time: 23:50)



Performance modeling concepts

- Single queue approach
  - λ: mean arrival rate (request/sec)
  - μ: mean service rate (request/sec)
  - Probability that zero requests served $p_0=1-(\lambda/\mu)$
  - Probability that k requests served $p_k=(1-(\lambda/\mu))(\lambda/\mu)^k$
  - Server utilization $U=1-p_0=(\lambda/\mu)$
  - Average no of requests per second $\overline{N}=\sum_{k=0}^{\infty} kp_k = U/(1-U)$
  - Average throughput of the server = $X=\mu.U+0.(1-U)=\lambda$
  - Average response time= $R=\overline{N}/X=(U/\lambda)/(1-U)=(1/\mu)/(1-U)=S/(1-U)$

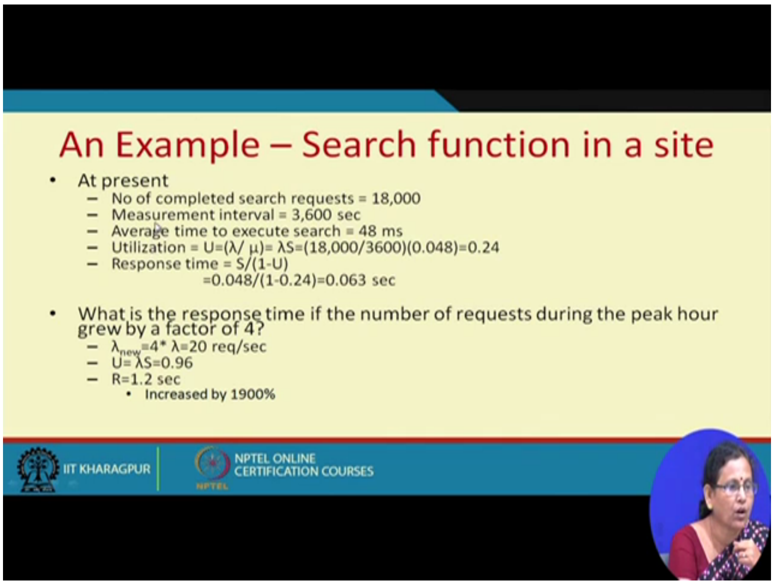
where, $S=(1/\mu)$= *Service time*

So considering this single queue approach, if lambda is our arrival rate, which is number of requests per second and mu is the service rate which is requests served per second. And where from we will be getting these parameters, how many requests are coming per second and how many requests are served per second, these will be again obtained from access log data. Please once again go through the access log data to find out that how much, how these arrivals, number of requests and the service rates can be calculated.

Then we have the probability that the 0 requests are served per second is given by the formula, then probability that K requests are served per second, this is given by the 2nd formula. Now this formula, 1 - lambda by Q etc, these are not coming, I mean, I am not as proposing this like this, they are actually output of a single server queue model. This is a very basic queueing model but if you would like to have a performance parameters, the performance model which is much more detail, you can even have a queueing network, even you can have a queueing network.

So therefore, let us say the formula for, formula that sum, the probability that K requests are solved per second is this much. Then your server utilisation is given by this 1 - P0, that is lambda by P, average number of requests per second is given by another formula, you can find out the average throughput of the server and average response time off from the server. So from a queueing model, for a specific search operation, you can find out these parameters.

(Refer Slide Time: 26:09)



An Example – Search function in a site

- At present
  - No of completed search requests = 18,000
  - Measurement interval = 3,600 sec
  - Average time to execute search = 48 ms
  - Utilization = $U=(\lambda/\mu)=\lambda S=(18{,}000/3600)(0.048)=0.24$
  - Response time = $S/(1-U)$
    $=0.048/(1-0.24)=0.063$ sec
- What is the response time if the number of requests during the peak hour grew by a factor of 4?
  - $\lambda_{new}=4*\lambda=20$ req/sec
  - $U=\lambda S=0.96$
  - $R=1.2$ sec
    - Increased by 1900%

Now suppose from the access log you found the number of completed search requests is 18,000 and the measurement interval, the period during which you took these observations is 3600 seconds. Again from the data you found out, from the access log you found out the average time to execute the search has been 48 millisecond. Now what is the utilisation of your server, server utilisation is 0.24, which means server is 25 percent utilised. Now what is your response time, on an average is 0.063 second, it is a good number.

Now suppose the next question comes what is the response time is the number of requests during the peak hours generally grows by a factor of 4, which means you, some selling season is coming, you are announcing a mega, you are planning a mega sales event and expect that your number of customers who generally arrive will become 4 times more. So if there is fourfold increase, then your new arrival rate now becomes now becomes 4 into lambda, that is 20 requests per second. What was your original, it was 180 by 36, 180 by 36 it was.

So now it has, it is, there is fourfold increase. So now your utilisation, server utilisation is 0.96, lambda by mu, your server utilisation is 0.96 and response time is 1.2 seconds, so which is some 1900 percent increase. By the way if you have learnt wheb, I expect that all of you after going back, after seeing such a beautiful application of queueing model, all of you will go back and learn about queueing theory. But when you learn about queueing theory, you will understand that this utilisation that we are talking about, this lambda by Q in this particular situation where you expect any number of arrivals and you do not have any bound on your arrival, the utilisation of this type 0.96, it is not 100 percent utilisation.

But it is an indication that your server is going to crash, in fact beyond let us say 50-60 percent utilisation, if, because after all you have modelled your, the queueing model that we have considered, we have the basic assumption is the arrival, the people will be coming to your site following exponential distribution. And by the nature of the exponential distribution, sometimes very few people will come and sometimes very large chunk will come. So the value that we are considering in terms of lambda is just an average of that. So which means when that large chunk of people actually come, your site is going to fail. So therefore with this modelling process, we got the indication that if we are increasing a fourfold increase in our traffic due to a specific selling season, our server is not adequate.

So may be temporary be we make some arrangement because this selling is going to stay for let us say 4 to 5 days or 15 days let us say, beyond 15 days it is not there, so why to invest so much infrastructure. So after looking at this model, you can now decide whether I should be actually investing in the infrastructure because if I am going to frequently have this sales like every month one day, every month one day, I can have, it is worth investing. But if it is a let us say yearly phenomena for let us say 2-3 days, it is not worth. So it may be possible that I should be hiring some outside servers for this period to maintain a situation where the response time and throughput is within my expectation. With this we finish this lecture, thank you very much.