

**Linear Systems Theory**  
**Prof. Ramkrishna Pasumarthy**  
**Department of Electrical Engineering**  
**Indian Institute of Technology, Madras**

**Module - 01**  
**Lecture - 03**  
**Part 02**  
**System Models**

Welcome back. Another interesting example to learn is that of a computing system or also what we call here as a web server system ok.

(Refer Slide Time: 00:30)

The slide is titled "Webserver Control System<sup>2</sup>". It features a diagram on the right showing "Clients" (represented by a laptop, a smartphone, and a tablet) connected to an "Internet" cloud, which is in turn connected to a "Server". The diagram uses dashed lines to indicate the flow of information and resources.

- ▶ Web servers are programs that run on the remote machines and process web-requests generated by multiple clients and provide information in the form of webpages.
- ▶ The objective is to provide a fast response to requests from users
- ▶ At the same time ensuring that resources (CPU and memory) are not overloaded.

**Why is this challenging?:** The resources might be shared [contention], the amount of available processing power and memory is uncertain, and feedback can be used to provide good performance in the presence of this uncertainty.

<sup>2</sup>KJ Astrom & R.M Murray, Feedback Systems, An Introduction for Scientists and Engineers, Available Online

Linear Systems Theory      Module 1 Lecture 3      Ramkrishna P.      8/13

Now, not really surprising or what I mean by here much of the time we are trying to access some information from the web could be in terms of news portal something which will tell me score of a cricket match. Of course, here I will not do things like IRCTC which are more transactional that you have to log in and provide your credentials captcha, not a little bit just simple information like how do I retrieve information from a web page.

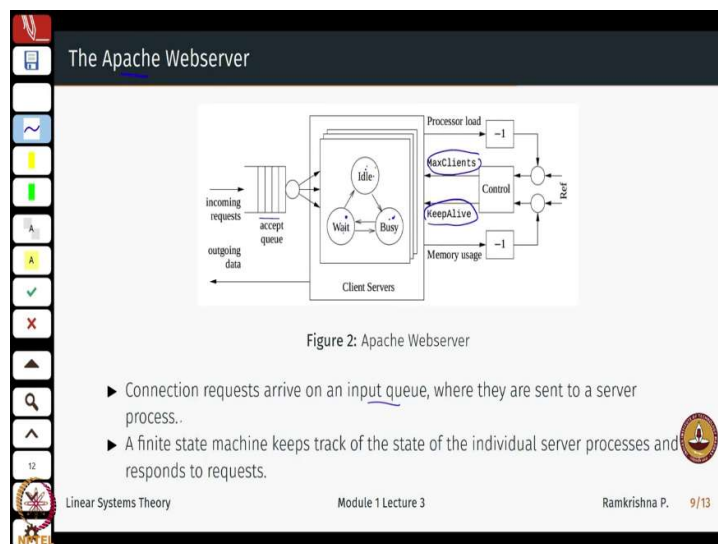
So, what are the typical setting that we are looking at right. So, these web servers are programs that run on some remote machines generated by some multiple clients, it could be from your cell phone, tablet, PC or laptop and several others ok. So, as a client so, my objective is that as soon as I hit the enter button, I just have to see all the information instantaneously right. From the service provider point of view what I am also interested is

I should be able to serve the requests at a faster rate, but at the same time ensure that my resources especially if I am talking in terms of a server, my CPU and the memory are not are not overloaded ok.

So, the objective here is well I have to provide a service without myself or if I am the server without myself not being overloaded ok. Why is this challenging? Right so, one way to say ok if there are more users just provide more resources and the problem would be solved. Well it is not as simple because, most of the times what happens is that the resources that we have might be shared by some other processes, which in the computer science, terminology is also called contention. And therefore, in many cases the amount of processing power and memory is uncertain, also limited.

And, the question which we can ask being a learning a control course is can we use feedback control theory, to provide good performance in the presence of limited resources and also uncertainty ok. Before I answer these questions, I should know well if I say feedback, then I should be able to measure something if I am measuring something where is that coming from. Or, in other words can I write down a model which describes the dynamics of this entire system starting from the client to the server via some internet ok.

(Refer Slide Time: 03:16)



So, what happens is if I here is a typical scenario of the Apache Webserver. So, whenever you request information from a web page you are first placed in a queue. So, and I request is in the queue until it is assigned worker right.

(Refer Slide Time: 03:39)

The Apache Webserver

- ▶ How does the entire process work?

Queue  $\xrightarrow{\text{idle}}$  Worker  $\rightarrow$  Busy  $\rightarrow$  Wait  $\xrightarrow{\text{KeepAlive}}$  Idle

- ▶ How does one respond to client requests at a faster rate? We would also like to serve as many clients as possible?
- ▶ How long to "KeepAlive" a connection? How many simultaneous requests "MaxClients" can be served?
- ▶ Trade-off between performance (response time) and resource usage (amount of processing power and memory).

Linear Systems Theory      Module 1 Lecture 3      Ramkrishna P. 10/13

So, first I will be in the queue and I am waiting for a worker to be assigned and a worker is assigned if and only if it is in the idle state right. So, what happens once a worker is assigned this a little finite state machine here starting from idle, busy and these three wait states which will deal or which will handle the requests ok. So, the connections are on an input queue and then they are served to server for processing ok. So, what happens once? So, once a worker is assigned it changes its state to a busy state.

So, it just automatically goes to the state ok. So, once a request is served, it does not close the connection or in computer science terminology the session is not lost. So, what instead it does is, it waits the worker waits if there are any more to see if there are any further requests from the client, it waits. Now how long does it wait? Well there is some predefined quantity called the KeepAlive time, if there are no say if the KeepAlive time is 2 seconds, if you do not respond for 2 seconds the connection is terminated and the worker goes back to an idle state.

So, what is also essential to notice that, the worker cannot take any requests any further request when it is in the busy state and also in the wait state ok. It just doing nothing, but still it is not allowed to take any other request unless the connection is terminated. So, it waits for what is called the KeepAlive time to expire and then goes back to the idle state only when it can take or only in the idle state it can take new requests ok.

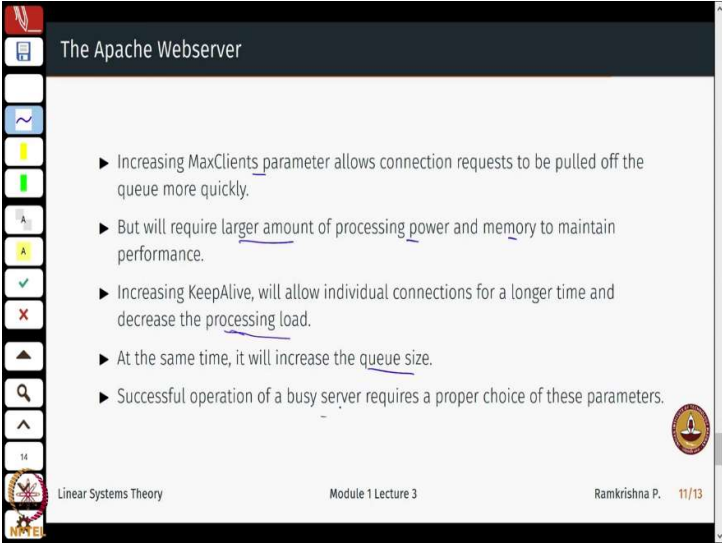
So, now how does, how do we look at this entire process and ask our self how can how the service can be made better right or how can the clients be served at a faster rate? Not only that we will also like to serve as many requests per hour or per day as possible right. So, the question would be the two things right. So, can I limit the number of people accessing my service? Well, I can define it by MaxClients say not more than 600 clients can access my server at any given point of time.

Anybody else will just be in the queue waiting for a worker to be available? I can also look at changing the KeepAlive time; should I keep it say 2 seconds or 20 seconds 2 milliseconds and so on? Now, what is the impact of KeepAlive for a longer time, what is the impact of increasing or decreasing the MaxClients? So, here we are we are also looking of a tradeoff between performance which is the response time like how long does it take to access or to download the webpage and the resource usage.

One way to look at it is if I were to have better service, I need to have more resources right that is now is that always the case. Say for example, if the traffic situation in any big city is really bad at 9AM, does it really mean that we just double the number of roads or double the number of flyovers or the public transport or any other infrastructure is that really a wise option or not?

So, we always look at it as at a tradeoff right. So, if I know that if I travel at 9AM, then I might take a little longer time to reach from point a to point b. Then if I am actually maybe going at early afternoon or maybe even late in the night ok.

(Refer Slide Time: 07:42)



The Apache Webserver

- ▶ Increasing MaxClients parameter allows connection requests to be pulled off the queue more quickly.
- ▶ But will require larger amount of processing power and memory to maintain performance.
- ▶ Increasing KeepAlive, will allow individual connections for a longer time and decrease the processing load.
- ▶ At the same time, it will increase the queue size.
- ▶ Successful operation of a busy server requires a proper choice of these parameters.

Linear Systems Theory      Module 1 Lecture 3      Ramkrishna P. 11/13

So, what happens if I increase the MaxClients? Right, if I increase the MaxClients I can actually pull off more request from the queue and therefore, possibly serve more requests in a given amount of time. Now, if MaxClients are increased, I have to serve the request of all those people and that will require a larger amount of processing power and memory to maintain performance. Sometimes I can say well let the service be slow, I can just take in as many people as possible well that is not typically allowed right ok.

Now, if I increase the KeepAlive time. So, KeepAlive is like almost like an idle state right. So, the worker or he is actually not processing any job. So, that will reduce the load on the processor. So, my CPU utilization and also the memory will usage will go down right. But if I am just waiting for a request the good thing is my resources or my processing load decreases, but at the same time my queue size increases because there are other people waiting for the request to be processed ok. Now, the therefore, we need a smart choice of these parameters for successful operation of any web server process ok.

(Refer Slide Time: 09:08)

The Apache Webserver

- ▶ The dynamic model (discrete-time) consists of states as the average processor usage  $CPU(k)$  and the % memory usage  $MEM(k)$ .
- ▶ The control inputs to the system are the KeepAlive time  $KA$  and the MaxClients  $MX$ .
- ▶ How does one obtain the dynamics? Around an operating point?

$$\begin{bmatrix} CPU(k+1) \\ MEM(k+1) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} CPU(k) \\ MEM(k) \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} KA(k) \\ MX(k) \end{bmatrix} + B u(k)$$

▶ A linear model around an operating point

$CPU = 0.58, MEM = 0.55, KA = 11s, MX = 600$

is obtained as (via System Identification Techniques)

$$\begin{bmatrix} CPU(k+1) \\ MEM(k+1) \end{bmatrix} = \begin{bmatrix} 0.54 & -0.11 \\ 0.0026 & 0.63 \end{bmatrix} \begin{bmatrix} CPU(k) \\ MEM(k) \end{bmatrix} + \begin{bmatrix} -85 & 4.4 \\ 5 & 2.8 \end{bmatrix} \times 10^{-4} \begin{bmatrix} KA(k) \\ MX(k) \end{bmatrix}$$

Linear Systems Theory      Module 1 Lecture 3      Ramkrishna P. 12/13

So, when I look at modeling this system do I know any equations from any physics textbook or any maybe even computer science text book? Well, the answer is no. How do we look at models for this? So, first is can I identify the states? Well, I can identify my states as the CPU and the memory like the average processor usage.

So, the CPU (k) CPU at any given time instant k is measured as the average processor usage and memory is usually on the percentage of memory usage, again at time k and for obvious reasons we will model them as discrete time systems ok. What are the control inputs that can alter the performance of the system? Those are the KeepAlive time and the MaxClient. So, that KeepAlive time will reduce the processing load, right one of the objective was not to overload the processors. Increasing the MaxClients I know that increasing MaxClients I can serve more request, but that actually requires a larger amount of processing power to maintain performance.

So, now how does one obtain this dynamics? Again I am looking at a linear model. So, I would have to pre-define an operating point. So, essentially to say that I know that these are the states, these are the inputs this and this and I just want to. So, in a simple discrete setting I am looking at  $x(k+1) = A x(k) + B u(k)$  gets x are the states, u are the inputs, this is my A matrix and this is my B matrix.

Now, I just somebody has to tell me what are these numbers?  $a_{11}, a_{12}, a_{21}, a_{22}$  and similarly over here ok. Now, what we do is we set an operating point could be whatever

let us say in this in this example. So, I just plug this or get this numbers from this book again. Good thing that much of the things which I am quoting so, far are freely available online. So, you can just access that and there also bunch of other examples exciting examples in this book ok.

So, what happens? Ok I just do a set of experiments which are called system identification techniques to give you a little indication of what this system identification is say suppose. So, these are models which are generated based on a series of a input output data right. So, if I say as simple as a I know that I have a resistor R, that it is actually a linear resistor in the operating region which I look and I ask you to find out may be just make use of some equipment in the lab to find out what the value of R is.

So, what I do? I just look at start from 0 volts till some 20, 30 whatever volts I just see how the current changes. So, here I have the voltage and I just plot it against the current, it well it should readily give me a straight line and this slope here is the value of the R right. So, this is a very simple identification experiment, identification experiment by just looking at some input output data. So, the problem here is to find out the value of R. So, what I know is that these they are actually most likely related by a linear relation  $V$  equal to  $I$  times  $R$ .

Similarly, I do a bunch of experiments and exploit lots of statistical tools, again it is a big subject on this course called system identification we will not go through that. But so, those experiments give us a model which looks like this around this operating region ok. So, models are not essentially always by some kind of physics or some kind of equations we know from elsewhere, we can also determine models with help of some experiments ok.

Now, once I have this model, I just have to check well are this informative enough? Do they actually capture the effect of increasing as clients, do they actually also capture the effect of increasing the KeepAlive time? So, let us have simple observation here so, from the B matrix so, as the KeepAlive time increases. So, this is captured by this column of the B matrix, you see that as the KeepAlive time increases because of this negative entries it decreases both the CPU and the memory usage.

(Refer Slide Time: 14:01)

The Apache Webserver

- ▶ What can we infer from the model?
- ▶ From the  $B$  matrix, increasing KeepAlive decreases both the CPU and MEM usage.
- ▶ The MaxClients increases both the processing and memory requirements.
- ▶ The  $A$  matrix tell us how the CPU and MEM evolve around the operating region.
- ▶ The diagonal terms describe how the individual resources return to equilibrium after a transient increase or decrease.
- ▶ The off diagonal terms define the coupling between the CPU and MEM. A change in one causes a change in other variable.

Linear Systems Theory      Module 1 Lecture 3      Ramkrishna P. 13/13

Similarly, increasing the MaxClients well the effect of it is captured in this column, you see that increasing MaxClients actually increases both the processing and memory requirements ok. The  $A$  matrix with this entries tells us how the CPU and the memory interact with each other or evolve with each other around an operating point ok. Well as usual the diagonal terms will tell me how the individual resources return to the equilibrium after certain transient or after initial perturbation. The off diagonal terms will tell me some kind of inherent coupling between the CPU and the memory usage right.

So, these are other ways of looking at model. So, once I know a model of this I can exploit tools from feedback control theory, I can check if the system is completely controllable or not and then design controllers for various performance matrix right.

So, so we end the examples section by this example of course, the literature I cited has, lots of other interesting examples, you can read through those maybe we could discuss those things over the forum. So, in the next lecture which will be like a concluding module or concluding lecture of this of this week 1 of lectures, we just look at some general formulation of systems, that we will be studying through the rest of the course.

Thank you.