Performance Evaluation of Computer Systems
Prof.Krishna Moorthy Sivalingam
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module No. # 01
Lecture No. # 01
Introduction to performance evaluation of computer systems

Good after noon again. So, this is the c s six two one zero (( )) in the class and… So, the first few minutes I will spend discussing the course, the objectives and what the expectations for the course are in terms of assignments tutorials and so on. So this course is in the r n t slot which many of you probably know I have taken. So the theory part is the four or four credit course. So, the theory part for this course will go from two to three fifteen on Wednesdays and Fridays then, we have another fifty minutes total for a for a tutorial, tutorials would be normally problems which we work out in class there is no lab components for this course, what will be tutorials? Will be primarily problems as we go into all the math part later on, there will be some examples and so on. We will work out in class.

So, the course objective is to understand basically performance evaluation. Performance evaluation of systems, in general performance evaluation applies to any kind of system like any system which we build we can measure performance and see whether one is better than the other, you can applied mechanical systems, electrical systems and so on. This course will try to focus on techniques most relevant to computer systems. So systems we will talk about examples of the system; so computer systems and, because I am also computer networking person there will be more emphasis on examples from computer networks. And, in terms of the prerequisite for this course I expect that you have basic background in as a basics of probability and statics which I think most of you have some of you have had, but think that you do not have. We will see as we go along, if you think that your background is adequate or not, but there will be the fair bit of mathematics involved, but I do not think it is of the order of like a statistics course there were you have to… I get going to that much detail, but there will be still math, especially for the first half to one one-third to one-half you will have deal with lot of the mathematical concepts. So, be prepared for that and that is why all this tutorial problems are primarily for the mathematical problem so, that is the first part and the this is ideally course taken in the seventh semester.

So I think some of you are in the seventh semester b tech students or dual degree students. So I prefer that you have had an hours course completed as well as networks course completed, because some other examples will makes sense only when I talk about, it does not need you to remember all the details of your (( )) system course or computer architecture course. But sometimes if we say I am going to explain how to derive the performance the started aloha you should know what I am talking about right it just can say what is started aloha? That point you take. So, text books there are primarily this will be the Raj Jain's book which is this one perform art of computer systems performance evaluation; this is a good book it is easy easy to understand there are still some errors, I hope all those errors have been fixed, because it is twenty years old now and, but that is of the problem twenty years old. (()) he is not really updated this book not that the techniques are changed or whatever he written that is not really changed too much, but we just like to see some more recent examples and so on, and for that there is the alternate (( )) buy this other book it is a it should be there in Tata.

Because, but this will a 1982 edition I do not have the latest edition I hoping to get that this is a book by Kishore Trivedi which was the second book on (( )). So also in equally good reference very similar kind of topics are covered and what is not covered on Raj Jain I will cover with special notes or some; I will give a hand notes for those as petrinets and so on might come later on we will look at that at the time. There are several reference books to which have listed most listed, let you know that are all also other books are talk about the same topic, but very rarely we have to go to those text books also.

So, the topics that we are going to cover in rough order, first is the basics of what is Performance Evaluation? What are some of the important techniques for Performance Evaluation? And, what are things to be considered There are two parts, one is how to do Performance Evaluation in the sense what should be end result? And the other thing is techniques, specific techniques like Markova chains and Markova modeling probabilistic analysis those are the techniques part, but there is another solid part which most of us tend to forget.

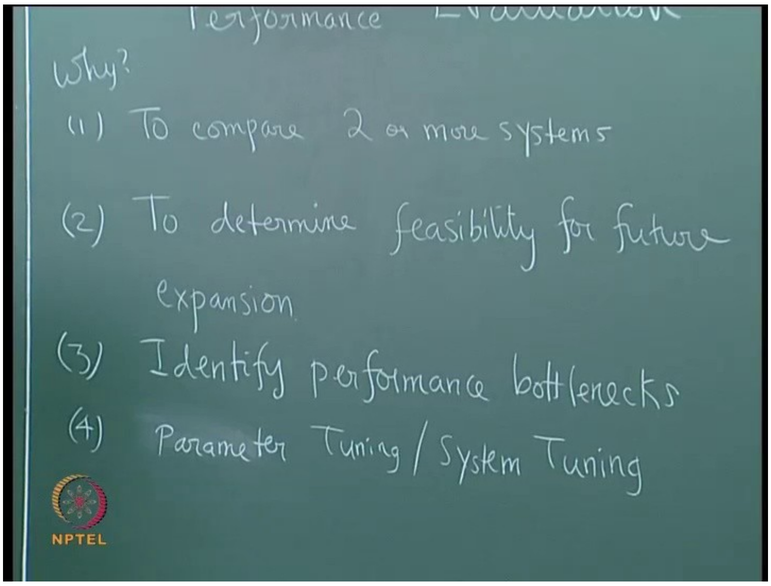
If you look at many of the B Tech projects M tech projects, you develop a system and then at the end of it you say, I have implemented the system done and then is the any measurement done on it. If you do a measurement then how did you measure? How did you compare? These things will talk about the rest of the class this week also, but that is the basics of what is performance evaluation? What are the expectation of a good performance evaluation. That is first part, one that like couple of chapters in Raj Jain's book easy reading we can quickly go through those points in two three lectures also, then we will start with the little bit of random variables.

So the initial random variables stochastic processes little bit of overview and then we go into queuing theory which is the actual derivations in you might have seen six zero four. The results of queuing theory which we briefly apply, but now it is will start with Markova chains and virtual processors and so on, which are they have generic Techniques; you can use this in many cases and we will see examples of how we can use that for in this context queuing theory and if time permits we will see example from networking outsource. So queuing theory single server network of servers, mean value analysis all these are tied together. So, that is the first major chunk of this course that will be probably ((   )) one time, then then stochastic (( )) is also another moral module which is... So these are the main theoretical techniques that I have listed and if I end up covering these two very fast they will add more, they were lots of tools and techniques either you will do more techniques or look at more case studies, how these techniques are used in terms of analysis. Then will move on to discrete event simulation.

So there are 3 types of (()) Performance Evaluation techniques one is analytical modeling, mathematical modeling which will again this only is scratching the surface do not expect to know all the different tools in the world and as to, how to do Performance Evaluation. It gives an indication some where knowledge, you can start with you can feel confident that later on, some other technique comes up becomes to still go on use those techniques. So simulation is the second part which many of us do and there are networking particularly lot of this have use n s 2 or have use n s 2 and so forth. So that will be simulation part again what did you simulation? How did you simulation? What are the basics of d e a system?

And one should know some of the stopping criteria all those things will come later on. Then once I have this basic techniques - two techniques covered third is experimentation which I want to get in to (()). You have to simply implement systems then we talk in actual terms of what are the things to measure? How did you identify those things to measure? What is the meaning of work load? How do you characterize workloads? And then little bit on data presentation, how to compare two systems? All these are based on the confidence intervals and so on. So that is when we do the implementation or do the experimentation implement the simulators then we have results. And how do you process the results? And how do you present it? And some of it simply heuristic, but at least you should know what the meaning of when I say zero, what is the zero conference interval. Sorry it is the conference interval contains zero or not; those thing we will talk about later on and then finally, case to case applications and most important. So, then I was planning on 2 or 3 programming assignment, I just want to introduce discrete event simulation to you and therefore, will implement some protocols also some mechanisms in that. So, this course deals with performance evaluation of systems in general and computer systems in particular, networks or architectural systems and so on.

(Refer Slide Time: 08:01)



Performance Evaluation

Why?

(1) To compare 2 or more systems

(2) To determine feasibility for future expansion

(3) Identify performance bottlenecks

(4) Parameter Tuning / System Tuning

So, why do you want to do performance evaluation? What are the goals of performance evaluation?

Conversation between student and professor not audible

So one is to compare two systems; two or more systems, the system goes in hundred users and in future the system except some thousand users and next scalability should be possible. So credit the scalability while performing.

So to I guess to plan for future expansion to see if your system will be able to... So scalability is important, especially when we talk about systems today where we have thousand users use satisfactory, a searching in those days meant searching in relatively small database; now it is a global database effort, pretty much everything. So, comparing systems, determining feasibility for a future expansion or else you read chapter one. Yes, so the second is to identify a performance bottle necks or third. So to identify... so bottle necks in any system, it can come from different places.

If you look at computer system like there are several sub components several substance; we have memory sub system is there - within memory you have your cache management subsystem, when you have a hard disk system there is a file subsystem also. So there are several subsystems and then you find that as you add more users or the or the c p u gets faster, you are still not able to get the performance improvement that you are expecting. So, therefore, you need to find out which of these sub systems is actually causing the problem? In a typical system, for example, we know by now that your i o system is the major bottle neck you are hard drive system hard disk system is also is also bottle neck, because your file system is not as fast as your main memory system and so on. So, there are these also despite having a faster c p u you find that some of these other systems contribute your system being slow, you need to find it out. So, that is one of the is to find where the bottle necks in the system performance are?

Then, single system. So if you have some problems in that. Is this performance problem or functional problem? The functional problems your performance evaluation will not necessarily tell you.

We will doing scalability analysis, yes you will do performance evaluation for a large system you find that whatever work from hundred system hundred user and not working for million users could be, because your algorithm is running very slow or could be because of some hash define that says hash define max users hundred and then your system fails when you go to million users that is more of functional issue than a performance issue. So that, you assume that your system is functionally tested and you know you know that it well for x number of users as to how well it works is what we try to find out.

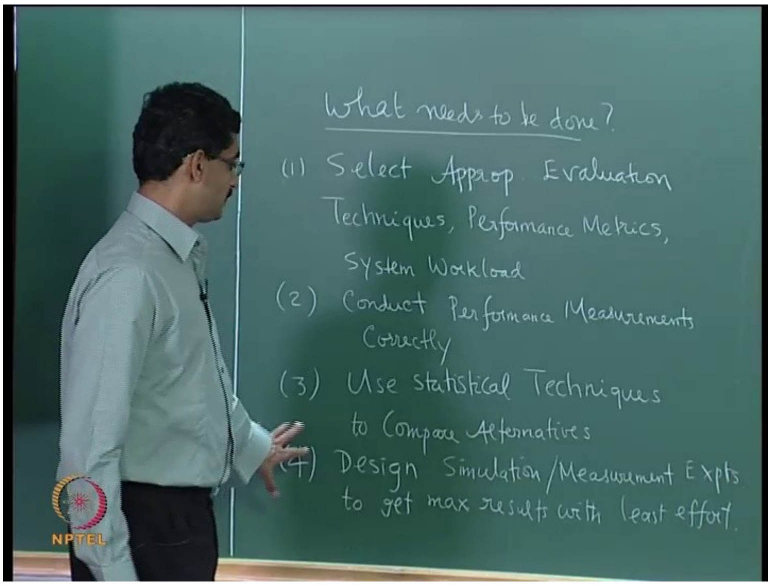
Any other uses.

You also would like to for any system, there are several parameters that you have to tune your system depends on many parameters it can be set. For example, operating system page size, how do you know which is the best page size? Is it four kilobytes or eight kilobytes? It will be have thirty two mega bytes for the page size; some way we have to figure it out. You know that I can build a system for any page size, but for a system to operate correctly optimally you need to find out so called, this is what is called as parameter system tuning are that is it look. We will edit this. So parameter tuning or so called system Tuning. You should find the right operational parameter for any system and sometimes it is not possible to simply guess as to which is the right one you have to try out several combinations of the system. And then you say that this is indeed something that works. But in the case of o s you have cache again we look at the hard drive also what is the best block size? Block size is determined by the hard drive, but there is also something the o s can decide. What is the minimum number of bytes I will transfer to and from the file system? That also will determine your system performance. So, there are several things to be tuned and set correctly if you look at networks again even if you t c p in t c p there are also several parameters to be tuned.

Can you think of some windows size yeah windows size. So, how did people design decide the sixteen bit is adequate for the window size. They thought that randomly sixteen bits therefore, the large size of sixty four kilobytes; you find that sixty four k b is not enough, how do the receiver and the sender agree on some appropriate window size? There is one. And so, like this you will find several system parameters that you have a series impact on your system performance if it is said, if you say this the conjunction (( )) very small value then you will find that your performance is not satisfactory you will not able to get the maximum efficiency from your t c p or link level protocol from that is operate.

So, these are some of the reasons anything else that comes to mind as to why you should performance evaluation? So fine, we will start this; we leave this as there are adequate reason to look at Performance Evaluation about seriously. So, then what needs to be done, before you go to how? Is either I use Markova Techniques or may be theoretical models or mathematical models such as simulation or implementation and so on. Other are few set of steps that one needs to do, we will just go head and list those.

(Refer Slide Time: 14:20)



What needs to be done?

(1) Select Approp Evaluation Techniques, Performance Metrics, System Workload

(2) Conduct Performance Measurements Correctly

(3) Use statistical Techniques to Compare Alternatives

(4) Design Simulation/Measurement Expts to get max results with least effort

So, that is the why, then we will look at the what. So this book actually has six parts and each part roughly covers one of the six different things that has to be done. So we will kind of write that in , it is kind of self explanatory. So I do not need to… So first is to select the (( )) appropriate evaluation techniques.

So, assume that you have you have somehow built the system, let us say implementation is being done. So you need to define certain things, one is what kind of experiment which you going to build? And then what are the performance metrics as well as system workload and so on? So one has to define what is it that I am trying to optimize? A system can be measured the performance of a system can be measured with several metrics in different in several different ways. Its simple example is if you look at network, delay is one metric, response time is the one metric, other is throughput - how many packets per second are being handled by a particular router and so on.

So those you have to first of all quantify, which are these are important that make sense? You can also take delay and say is only main important or should its variance also important, that is another extra or you have look at the third moment, fourth moment and so on, or see all these important or only some of these are going to be important. And the next is what kind of work load. If you look at our wireless systems today, your cellular systems it was mostly voice data. So, mostly people optimized on making sure that the voice calls went thorough. You had large number of voice calls accepted with good quality of service. So we have do not care all these now noise and so on, then slowly data started coming peoples start using more of data calls then you have; now your work load start changing.

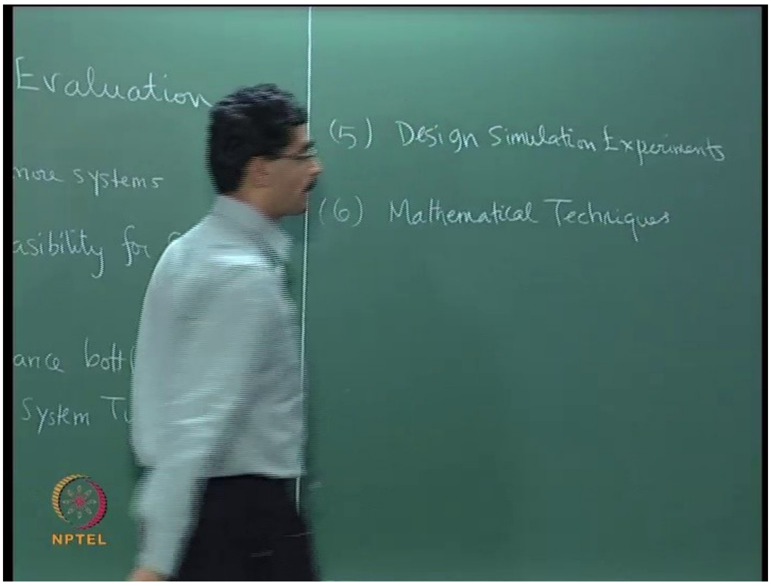
So, now I have eighty percent of the calls or voice calls twenty percent is data. But over years that may change where most people are simply using regular land line. So cellular is mostly for wireless, now it is only the primarily internet access device. So, it becomes eighty percent data twenty percent voice in a work load starts changing. So you have to sit down and think what are the possible ways in which your system work load can be characterized? And then it gets even worse when we say people are using the internet for g p r s, 3 g whatever it is and then thus the large bunch of people simply watching TV on their mobiles. Now, you have now you have to look at and say thirty percent voice, thirty percent data simply downloading, another forty percent is for video traffic with which requires much higher fidelity and you know people will not tolerate all drop (( )) some things like that. So, therefore, your system workload characteristics is also very important step, that is the first part you would say you have to work out several such scenarios, and then say these are the possible scenarios that problem is that this can be infinity - the number of scenarios can be infinitely large.

So you have to use your judgment as a as a performance evaluator and say I know how the system should behave? And these are typical scenarios we will present those cases to you for evolution. So, that is your first step. So, we look at much these is the little bit later in terms of these workload characterization and so on. So, once I define what experiments to conduct. So, then I should do my performance measurements correctly. So, now you run your experiments, collect the data and this is more or less routines stuff that will do we have seen that in labs as to how to collect data from the performance measurements and so on? Then, once I get the metrics, the metrics that I measure then you need to do some statistical analysis it is simply cannot say that, the system a is better than system b based on a few test. So there is to be little more rigor in the way that you will compare two systems that is where you will have your use of statistical techniques to compare alternatives.

So, as an example it is you have your own fancy sorting algorithm you know that there are so many existing sorting algorithms or you would see them in a d s a in other courses anywhere. And you have your own sorting algorithm and then at the end of it you have implemented you have designed implemented you want to see whether yours is better than the others system, then you should have conclusively statistically you should be able to say that your system is definitely better than the all the other existing systems and it sometimes it will be possible - sometimes you will be not able to say that there is statistically different this system is statistically much better than any other existing system will be good in some cases, but bad in some other cases it depending on how your algorithm is designed. So, you need to have some proper techniques that really tell you that, either it is definitely good you cannot say a most of it is more or less good that is not really adequate.

You have to have rigour when you say that I am comparing one system versus another system, then the… Actually some this is in little bit out of order, this design of experiment should have come before, but the book is kind of organized this way. So design simulation experiments or whatever it is, simulation, measurement in such a way that, you get maximum results with minimum effort. We will see what this means later on? For example, if you trying to measure different metrics in the same simulation 1, you should be able to get all the metrics at the same same time rather and have to have multiple runs of the same system each for an different metric different points in time and so on. So, how do you Identify the important factors either affect the system performance? I will define those later on what are the factors that affect the system's performance, which of those factors to vary? And then how do you get the system performance analysis? So this is another part that tries to tell you how to design? How to identify what are the factors? And, how to combine factors together? And so on.

(Refer Slide Time: 21:39)


Evaluation
nore systems
asibility for
ance bott
System T
(5) Design Simulation Experiments
(6) Mathematical Techniques
NPTEL

Then, we will come to five, here this is the (()) the second two parts are the actually relating to the techniques themselves. So, if use discrete event simulation you can use simulation also; its tool it should be used correctly you can use it incorrectly also and when we go later on will see how people (( )) using this incorrectly. So, how to design design of… So, the classic question that people ask is I have a simulator and running the simulator, let say you are processing some request, the packets coming into a router you have implemented a new scheduling algorithm for a router and then you run the system for several packets coming in a different points of time. And then you are measuring the system itself, then the question is, when should I stop the simulation? In the case of a mathematical model, you simply you have to apply your formula you run it and then you will get a number, you say this is the main delay this is the main throughput you was tell this is the value you take that and go home. But in the case of simulation you can run it for thousand time units, million time units, twenty million time units. When do you know when to stop? How do know that your simulation has reached some sort of steady state where there is no point in running simulation any further. if you top is if you stop its simulation too soon then you end up with mostly a transient behavior in the system if you run it for very long, my students will tell you that simulation will run for hours together and then we vary several parameters and then you have multiply that by the number of hours. So we can quickly run into (( )) when you have several factors each factor techniques several possible values and then it takes a very long time you will never finish your B Tech or M Tech project.

If you really let everything run for a billion seconds and so on, then you, then how do you that you actually achieved some sort of steady state. So these are things that one has to know that you design your system efficiently, otherwise you will simply run it for run it for ten thousand units I hope my advisor will not see that and therefore, there are some results coming out with same adequately that is the fifth part of this book and little bit of course also we will talk about what simulation is? So you understand when your when you simulator that there are some things to be dealt with the simulation itself. Then, the last part is actually the mathematical techniques and there are several mathematical techniques we will only focus on some that gives as an inkling of what is going to be there? What kind of systems are there? So, this will be largely queuing systems to certain extend to understand queuing systems better; we need to use some little bit of Markova chains and Markova processes understanding that and all the other things I mention before mean value analysis and all those things will come. So this is roughly six major components of this text book and roughly what also you will try to follow?

Sir what is the difference between first one.

This this is actually defining what you want to measure? Yeah. That is whether you want to use simulation tools or mathematical tools or experimental tools that is the first step that you try to do. Then you decide what kind of the performance metrics such you want to measure.

((how they decide))

It is matter of its usually what you do is, you try to do all three the classic thing is you never trust your simulation. We will see examples ore see case to what particular characteristics are… Simulation will give a good abstraction, but simulation will the problem is resource constraint it may take a very long time to get good results with simulation. Analytical techniques will be able to get quick results you can put it mat lab and get a formula you know, even with iterative you can still get the results fairly quickly, but mathematical techniques will make lot of assumptions. So you will abstract the system away. So much that you miss out on the final details your system will mostly black boxes put together simulation, you can get into you can also get more detail performance metrics in the case of a simulation. You can you can measure everything that you can that it when the simulation in the analytical model you probably cannot do all the kind of metrics that you want to measure. And third is the implementation where there is depending on the system complex to you might implement you might not implement. But ideally what you do is you start with mathematical model which are actually suppose to be what to next part of this lecture. You start with math modeling to get asymptotic understanding of what the system behavior should be? Then, you say you know with for example, queuing theory all mean delay I can measure with my queuing model that it is tend we saw that the networks it is ten seconds for packet and so on.

But you want to know more about how about the packet which have this is the particular packet length you want to know for the different packet lengths. What the corresponding delays? you do not only you do not only you do not care about just the mean alone, you want to know more then you have to going to simulation, because your analytical model will not help to do that. Then, you go into implementation then you start suddenly seeing that your system operating system or whatever it is will start imposing serious constraints on your own on the performance itself then the implementation is the best. But before you go there, so usually what happens is you have say fifteen possible techniques look at sorting as the classic example, there are several ways to sort.

You first do your analysis says time complexity that says that you know these insertion sort is probably not the best to do the bubble sort is not the best to do, because of its high complexity therefore, I should go to something which is closer to your and log in complexity there it is vary or theoretical model comes in then implement then you find that in the implementation two order n log n algorithms will be totally different, will have different behavior when it comes to actually running on various system. So, it is not the theoretical complexity gives you reduces your search space from say fifteen algorithms to four or five algorithms, then there are empirical issues that come in that we will understand with simulation. Then you really put into the system in then say what happens in the system where there is a more realistic work load where there are slight look at searching, for example for query is on the (( )) there you have you are on searching algorithm then that is a totally different domain here is a very control domain when you put it in simulation environmental. Where as in a real system, because in a simulator environment you vary the set of parameters, the number of know the size of your data set (()) based on what you decide, but your real system might work on totally different to work load. So you want to see on a realistic system work load how it happen. So you keep reducing your search place. So when you go to implementation will be probably have only two you start with say fifteen bring into five in simulation then you come down to these two seem to be most promising. Therefore, I will do it in that particular case. So that is the way which you normally scope down to what you want to implement? And finally, you what you will implement you will turn on to be the most simple one, because the more complex is the system is harder it is to actually implement. So you find that many time simpler solutions work better than complex solution, but the there thus depends on the problem sometimes complexity is divided so on. So that is what this is and the these are how do I compare two systems? We will see this later on we saw this on networks class also. Is waiting time the best a thing or is it the total time spent in the queue which is the most important metric you choose the wrong metric you might say one is better than the other.

So, therefore, the choice of metric is also very important and not just mean I said mean, variance and several other factors it resource utilization from the system point of view link as to be utilize to the maximum possible extent when a may system utilization is very high what happens to the user; we will see this later on the user will see very high delays in the system. So, if we want to add as many peoples hostels are now fully loaded, because you want to give the utilization very high, but then what happens, people are not happy. So, there is always is global utilization which system wants to happen I wants to see. And you have the local utilization user perspective you want only the whole floor to yourself. So there is no travel whenever you there is nobody around you, you have all the freedom of you know the enjoying this nature. So you have always this classic (()).

So choice of performance metrics, what will make a difference? And then work load again. So, how we characterize that is really more of an art then anything else, you should know what a how the system is going to behave. And if you took talk to some of the students in our labs this work load characterization is so hard, we might come up with the set of scenarios then usually we fall back on… So, what do you do normally?

We look at bench marks you say that here is the second bench mark, here is my (()) whatever system simply run on it and submit the second bench mark; whatever it is then say that on this bench mark my system is beating all the other guys that is what you will see typically in all the newspaper magazine adds why, because of the convenient (( )) out somebody has said this is the possible work load run it on the systems. And then you show that this is my one number whatever you compute or may be several numbers you show and that is where work load characterization becomes challenge is simply depend on some system experts to know the system for fifteen years are so. That will tell you what to do? That is what this will (()) comes a new.