

Storage Systems
Dr. K. Gopinath
Department of Computer Science and Engineering
Indian Institute of Science, Bangalore

Types of Storage Devices and Systems, Long-term Storage
Lecture - 13
Types of Storage Part 1: Mobile/Personal/Organizational -type Storage, RAID
concept


Welcome again to the NPTEL course on Storage Systems. In the previous class we looked at how to manage a system with multiple types of storage. We discussed this 5 minute rule and then we tried to see how you should balance, capacity versus the number of I O operations per second unit.

So, we will continue a bit on discussing the types of storage systems around there. So, in the beginning this was a illustration we looked at memory stick, and the memory stick what will be goes on right we had a brief look at that particular thing.

(Refer Slide Time: 01:17)

Types of Storage

- “Mobile” Storage
 - Memory Stick, Camera, Smart Phone, Laptop
- Personal Storage
 - PC, “Home” RAID systems, “Home” NFS server
- Dept/Organizational Storage
 - NFS/CIFS server
- Highly Available Storage
- Parallel Storage
- Web-scale storage



Today we will start looking at again in the all the types of storage that are available and I think first of all you should know that storage is everywhere right. You can use it in mobile storage for example, it is there memory stick, camera, Smartphone, laptop etcetera right memory stick can get laptops with SSD only. So, they are expensive, but

they are lighter and they are also somewhat better with respect to you should drop it most likely the laptop is still use.

Of course even disks can be designed so that if you drop them they survive how did I do it? They do it by they have some kind of you know standard accelerometers and then it like finds out that it is dropping and therefore, it can park the head. Normally what happens; is that there is a head and which if it crashes into the medium then you will use it.

So, the thing would be to as it is falling, detect that it is falling and then park it safely in a place. So, in for falls the head does not hit the medium. So, that is the idea. So, many laptops also have that feature. So, if you have that feature, then even if you take a fall itself.

But it generally turns out the new types of storage based on flash, they are much more convenient, but they are expensive example the camera for example, it all uses this kind of memory only. As far as to my knowledge nobody has tried to put camera under hard disk put it you might be some models, but I am not aware of it. So, basically because if you think about, it may be that especially if you have a video camera for example, right like disc will be make some noise which will also be captured we will shake all those things are problems. So, you do not want to incorporate this.

So, cameras you shall never do that, but that is what they have solid state disc as the most important part and you notice that there are differences in the way these systems are used camera for example, most of the times you are writing songs sorry I made a mistake. Camera most likely you are putting big files and these are usually megabytes right multiple megabytes. So, a system that is geared for camera ideally it should have not a 4 kilobyte block sized system, it should be slightly bigger possibly.

But what is the current situation because you want to be able to take the camera and put it into any pc, it has a fat file system. So, as a fat file system essentially decides what the block size. So, even know the camera should have a much bigger block size, why we will help you if you have a big block size? It turns out the metadata that you want to keep right its proportional to number of blocks right.

So, if you have a very large block size, there is essentially you are reducing the metadata by that factor. So, you are going to reduce it by a factor of thousand almost, if you use a larger block size, but it is not done because of interoperability reasons. Basically your machines are what you have they are able to handle fact that it to whatever that is what you goes ok.

So, that is even why ideally you should design storage for the specific application, but that often creates interoperability shows. So, you need to had your part. Same thing about smart phone also because you think about a Smartphone, nowadays we come with 8 GB 16GB, 32GB kind of stuff right.

Now, usage pattern of the Smartphone what kind of thing could it be we are updating, you are storing music, you are also updating your contacts, you are if it is a Smartphone it also is doing lots applications right. So, what could be the right model for Smartphone. So, it is almost looking like a regular desktop kind of system right.

So, because there could have variety of there could be varieties of applications and it turns out Smartphone's do not have a discounter. So, that only two level storage dram memory and you have SSD memory the flash memory. So, the thing is that instead of the disk situation you have SSD ok.

So, you can see that there should be some difference with respect to how these are being used. Again what is the kind of file system that the Smartphone uses? Again it has to be interoperable with; because you can take your Smartphone and attached to pc, again it should have a way in which you can read those things right again most likely what will be the file system it will most likely. Again a fact that the file system most likely you can design things in a which is different manner, but that will make it non interoperable

So, this interoperability is a big issue for consumer devices very big issue, you have to handle it somehow. So, we are not able to handle it then nobody will touch it. So, that is why all this things essentially mostly using fact 32 and it may or not be very good. For example, in this systems there is not much serious attempt at reliability or for example, if there is a bit error somewhere let us know it recover for it because that kind of stuff has not been designed it.

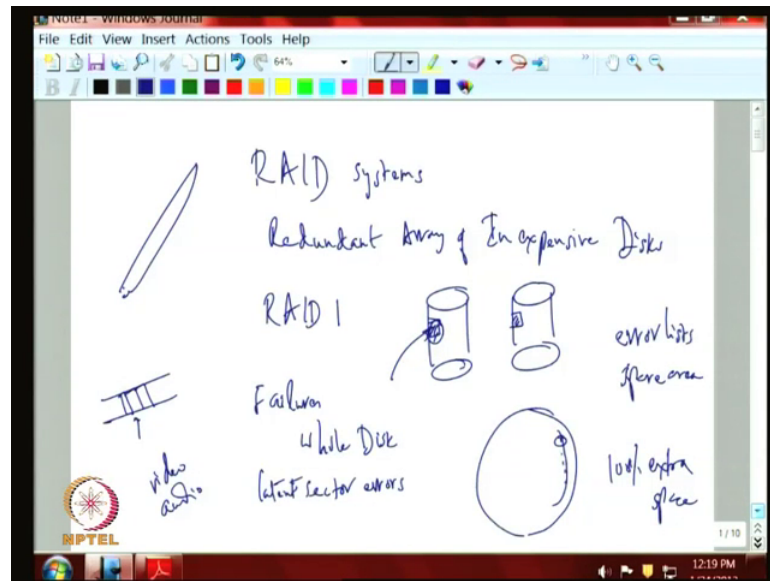
With laptops I think things could be slightly different, because it store critical information of course, it store critical information Smartphone's nowadays also, but still not as critical as what your problem is storing laptops. So, some laptops actually have a let us say ability to recover from errors, but that is very minimal state because for example, most laptop memories are not ECC. So, memory itself is not ECC there is no point in worrying about disk failures.

Basically what it means is that the memory itself can be introducing let us say errors in it, no pointing trying to recover from it memory itself because that is a larger generator of information because its computing all kinds of things it is a cache of all the important stuff and part of it will come to come to disk. If that generator stuff itself is junk you do not worry too much about other things that is an why most laptops there.

But in principle one can imagine future generations, where this laptop memories become very big, nowadays you can get laptop memories for about 4 GB, 8 GB right and if you call 16 GB or etcetera then people will start worrying about there was a large amount of memory, the possibility of error is now much higher. So, then probably most likely bigger in the future laptops with let us say 128 GB or such memories, they will incorporate things like error correcting codes ECC. In that which case then, we start worrying about yes I have taken care of memory, but what about my disk. Show to something about that disk also suppose it is silently corrupting things can handle it ok

So, in those kind of situations you might find that you may want to put protect, it also again certain errors that is where what you can call as there are some systems called RAID systems RAID; RAID what is RAID system; let us just look at this RAID systems.

(Refer Slide Time: 09:35)



So, we have something called RAID systems, what are these? This is that it is redundant array of inexpensive disks.

So, there are varieties of RAID systems for reliability, the one which is most common is RAID 1. What you do is, suppose you have a disk that is call symbol for disks this typically. So, if one block here you have an exact copy here once again. So, there is any if this particular disk fails, now the disk can failure multiple ways one way is what is called whole disk failure models, what is it? You can have failure models one is whole disk; what does it mean? It means that somehow something has gone bad so that the nothing in the disk is readable especially the something happened to the arm, something has happened to the electronics, the control controller security that something is there just not taking commands ok.

So, the whole disks failed its totally loss particularly what is called sector errors, often called latent sector errors what is the sector errors? The whole disks not failed, but parts of the disk are failed I think some of you might be familiar disk flop is in the past you will notice that some files are readable some files are not readable right.

So; that means, that most of electronics, most of the mechanical components are working, but some small media part of the media has fail it is gone back. So, it could be because of dust particle, it could because of something which cost that particular small section of the medium to go back ok.

Now, there is also another type of error called latent disk what I call latent sector error; that means, that you are able to write to it and then you do not touch it and then after about a month you try to go back to it. In that intervening period it has gone it was while you wrote it, but it very bad after one month somewhere. These in some size the sector failures latent it is somewhere there, but you cannot predict when it will happen. So, if you have a latent sector error you can lose a bit, but suppose have RAID 1 the idea is that if there is a latent sector error here very unlikely to have the same error displace here. Therefore, I can if I am unable to read from here, and look at a corresponding area and other disk and use it as a way to fix it. How do I fix it? I can try to here again because that part of it is gone ok.

So, normally disks have what is called error lists and they have a spare. So, they have something called error list and spare area spare area what is; that means, that in a disk for example, you have some sectors some tracks at the outside for example, which are for this spare areas. So, there are some sectors out there and when you try to recover from this error which happened here, you want to take the copy and basically you will remap it so that this becomes this one.

So, basically what happens is that, the hardware the folder of the system for the disk when somebody asks for this on the disk. It will be have been remap to this spare area, so that the software does not know anything about it. Internally what is happened you ask for the same place, but there is some table inside will this is say that these things are all bad it hit one of the things. Basically some sense you might do not associative match, it says that this thing as you asked for the block, I look at in my error list, if hits there then you know that you have to look up the corresponding place in the spear area, ok.

So, that will what happens is that you can recover from sector errors. So, the red one gives you the ability, but the problem with this is that you need to have 100 percent extra space these are problem with this solution. We can use what is called coding to reduce the amount we will talk about its slightly later ok.

So, what is one problem with this kind of solution? The solution the problem the solution is that, normally I am expecting you know that a disks they are good only when you are accessing things sequentially. So, what has happened right now because on error is suddenly went to you are thinking things are nearby right logically went from this block

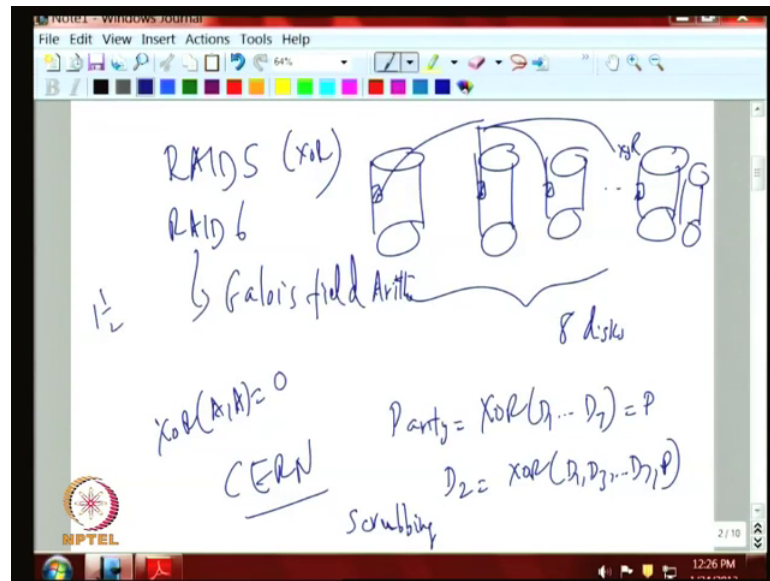
to this block, but this block actually is not here its actually somewhere else and then we want to go to the next block. So, what is happened? You read this block here go somewhere and I can come back here next, ok.

So, this creates what you might call as sudden latency variations luckily not much, but if you are talking about video streams or audio streams all right audio especially a serious problem because audio is basically what is called a real hard real time system. Audio why is that because if you do not because you need some samples every twenty milliseconds you need a sample we do not get it then you get a here click or something you there is you can do some adjustment, but still typically inspire of a best of our often times you hear, a click that click is basically is a lack of information because your system is not able to give that sampling in time ok.

Video is slightly better because there is a lot of redundancy, I think some of you might know that in video what I called predicted frames from the call I frames b frames and predicted frames, there is some redundancy in the system. So, even if you drop some things it is. So, if you do not get the thing in time you just drop it for the kinding and continue surveying the next ones, but usually you cannot really see it or my or I also can sort of adjust to it with adulates a problem ok.

So, you will notice that disks solutions work, but sometimes they can be some things, it could because of terms like this. If you hit a click I am not saying it because of this it might be bad recording itself, but it could be because of the some of the disks also. So, if you are thinking about laptops in the future of who knows you might have RAID systems also coming here and RAID 1 is an expensive way it can be done it is simple and there are better ways more cheaper ways for example RAID 5.

(Refer Slide Time: 17:14)



So, RAID 5 is often used there is also something called RAID 6, which is even more is better, but it is more complex. In RAID 5 what happens is that you do not want to pay 100 percent cost. What you do is you have a disk, you have so many disk, you can have a for example, 8 disk for example, I will just put I will let us say there are 8 disks; what you do is, you same block you take these ones and do an XOR and store it here ok.

So, basically these is what is called a parity this is the parity. So, parity is XOR of let us say D 1 to D 8 let us say D 7 because of total 8 disks. So, you have an XOR of these things and if one of them fails let us say D 2 face, we can reconstruct D 2 by let us call this parity. So, D 2 will be given by XOR of D 1 D 3 up till D 7 comma p. So, this equation is valid I think some of you if we know that properties XOR you can see it is very clear, because it comes out what is this you are XORing D 1 with XOR of this one. So, D 1 D 1 will cancel out and this I will cancel out and d three also will cancel out only D 2 will remain that is why you will get final D 2 because what is XOR of XOR of a comma A is what is 0.

So, basically you are cancelling when you do XOR of this things will P right. All the common things will follow up only that is not there here is D 2 that something what you come out. Now what is good about this? Good thing about this is that, you have some efficiency here because you are talking about for 7 disks I have to have an extra disk. So,

your efficiency is about you will use only once 1 of every 7; 1 of 8 disks basically you are wasting it in some sense because of redundancy ok.

But the problem that you say, the problem is that if I want to recover from this I have to go and read 7 disc then I will take a long time. In the sense the recovery is slightly heavy in the case of RAID 1 right you look at RAID 1 they are only two things you just have to look other disk you just have to read it is all are simpler ok.

So, that is why it turns out RAID 1 is most often used, the RAID 5 is used only in extremely large capacity systems, where this last that is there in RAID 1 is to in turns of a. So, you only have some petabytes of data, I think as I mention petabytes of data can be involve megawatts of power right. So, if you are worried about power consumption probably you want to go to RAID 5 we have to be careful about this right

Now, you cannot have RAID 6 where instead of one parity disk you have two parity disk; and they use some slightly more advanced techniques they want to this XOR they use something called Galois field arithmetic. RAID 6 is this, this is a mostly XOR ok.

So, just like a more expensive method that people have used it for example, many commercial systems used RAID 6 sometimes RAID 6, but what is happened now is that there are issues of reliability unit RAID 6. So, people have gone to the next step for example, people at places like cern for example, use only RAID 6 they still have some amount of errors. So, they will want to go for things which have not just two parity disk also, another one third parity disk triple parity triple redundancy some people are starting to look into those things ok.

Now, you will notice that going to more does not necessarily mean things are ok because you can have what I called a rush during reconstruction itself. For example, you have a RAID 5 right if I have to figured out that this thing failed. So, this failed. So, what did you start reading from here also reading it from the discovers that we discover that, this thing is unreadable that disks not failed, but this sector error is there on this particular thing.

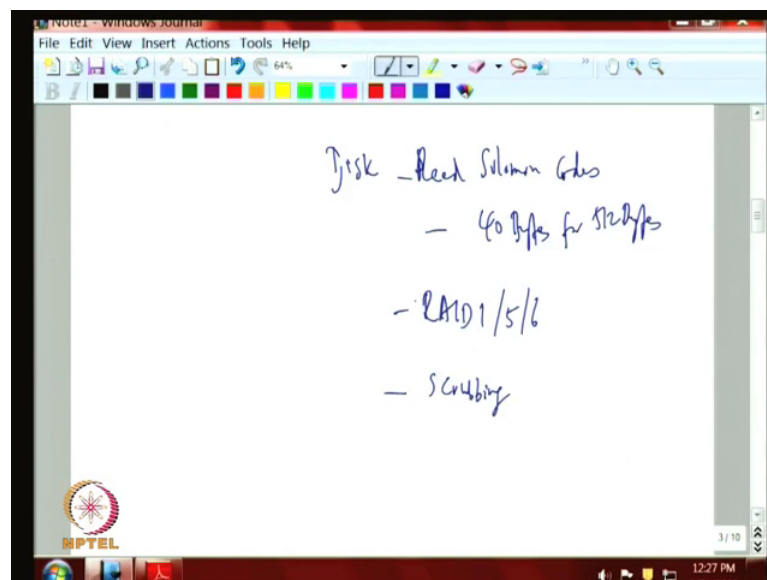
So, I really cannot reconstruct the (Refer Time: 22:34) this is often one called one and half error one and half error, that is a full failure followed by a sector failure because we are treat that particular thing and somehow it was not there right. Sometimes it happens

and people worried about this also. This is to be our data is very large you have to worry about this and the very worried you elevate this problem is what is called scrubbing what is scrubbing? Just like you periodically check is it readable. So, essentially instead of waiting for one month and discovering of one month that you are could not read it, every day you go through the whole system in the background in the same size when the system is not very fully utilized, you keep reading it to ensure that we are able to read all the sectors.

So, before that major error occurs you are essentially keeping everything readable. It is just like some aspect of you be paranoid about data things can go anything things can follow up at anything. So, what to do is every day you keep doing it. So, this is done actually this is called data scrubbing and it is widely used still large systems and very large systems terabytes of stuff you have to do it there is no escape ok.

So, you have let us see the kind of error mechanisms that you can see if you think about it you can see how many levels of things are there.

(Refer Slide Time: 24:12)



On the disk itself we have what is called read Solomon codes and these are geared for handling things like 40 bit codes for 512 bytes. So, sorry I think I should make 40 bytes for 512 bytes sorry this is a kind of extra informational skipped. So, grid calls is there what does do you need? You often have things like a RAID, RAID 1 or 5 or 6.

In addition to this you have scrubbing. So, you do lots of these things then there is some chances that you could actually be there when you want it. So, I think I covered some of this part of it. So, let us get back to. So, when you when you go see we have the mobile storage, suppose I am done the mobile storage I am pretty sure you will find storage and there is different forms so on right and it will be are everywhere just like computing is everywhere you will first storage is also coming around and I think mobile storage is the reason why you are starting to see lots of interesting candidates.

Camera being the first one that you must have seen ok, and Smartphone's, laptops all these things are somehow enable because of mobile storage. Now again can also have what is called personal storage much bigger amounts of stuff on the PC for example, or you may also build a home RAID system what is home RAID system? You decide that nowadays in many homes you will have multiple laptops and at least one big desktop and you will try to keep the big desktop as a place where you have a RAID system. So, that will be the backup place that will be the place where you keep another copy just in case your laptop gets stolen lost or whatever else right.

So, you can even build home RAID systems and nowadays most of the chips that we get have support for RAID 5 or RAID 1; that is you can put a controller which has support for RAID 1 or RAID 5. And therefore, you can have multiple disks on the pc and then you can have a low cost RAID system. But what is the good thing about this? The good thing this is low cost and it does not change your model that much and what is the bad thing? The bad thing about this is that if something happens to that disk something happens to that PC why should something happen to the pc? There might be a flood whole pc is under water which is not something unusual.

So, in Bangalore also it happens may be it happens many places. So, you can have flood you can have fire; fire is very common, you have a building the fire take place both the disks are there that doing all redundancy everything, but what is use of a redundancy in the whole disks fried in a fire. To these are all both together because that is why I need to have home. RAID systems are useful to some extent, but they are not good enough right with if you are worried about data, that is where this new trend about what is called cloud storage which are not written here has also come what do you do in that case? You not only store it locally, but it store a copy in the cloud also.

So, what is the reason why you want to go to the cloud? It is purely for disaster recovery here the issue that I cannot keep everything in the cloud because cloud becomes it is too slow. Because normally I can be dealing with multiple gigabytes of data, accessing gigabytes of data through whatever current technologies we have it is not exactly fast it is actually quite slow.

So, because of that you cannot really afford to keep everything on the cloud and if some of few might have notice that if you want to upload something like your all the pictures taken right it will takes some time for you can upload it right. So, on this thing they are not exactly use because you might get them was upload some 40 or so, pictures whatever 40 into 4 megabytes that is about let us say not 4 megabyte or a 10 megabyte also (Refer Time: 29:05) 4 megabytes it will at least some amount of time, but you can even have much bigger sets. So, cloud storage may not be that effective as regular access. So, you might use it for recovering purposes ok.

So, basically nowadays I think all of you are probably using it some form of them either or the other solution that across storage is part of the application itself. Like one using Gmail that cloud storage is part of application. So, that you do not stored it locally its always up somewhere out there, and the people are giving that application service the reserve business to keep it intact. And they might use wrote RAID systems or whatever system there they might use RAID they might be use distributed storage models etcetera which we will discuss soon ok.

So, that is possible. So, one is to just store the things other thing also you may also have what is called a home NFS server that is you your laptop etcetera they do not directly use, they do not have storage that much on the disk, they directly use it from the homes NFS server. This also some model a lot of people are starting to use and this can be helpful in many cases. There is in why that is good is because if I have multiple devices I do not back it back it up each one of individually. I back up this home NFS server and I know that everything is there.

Now, the biggest problems we have is that we have multiple gadgets, we forget to backup one thing or you do not back you back it up at different times. So, there is always some let us say some incoherence about the data, what you have you do not know you have to manually figure out what is the latest copy etcetera ok.

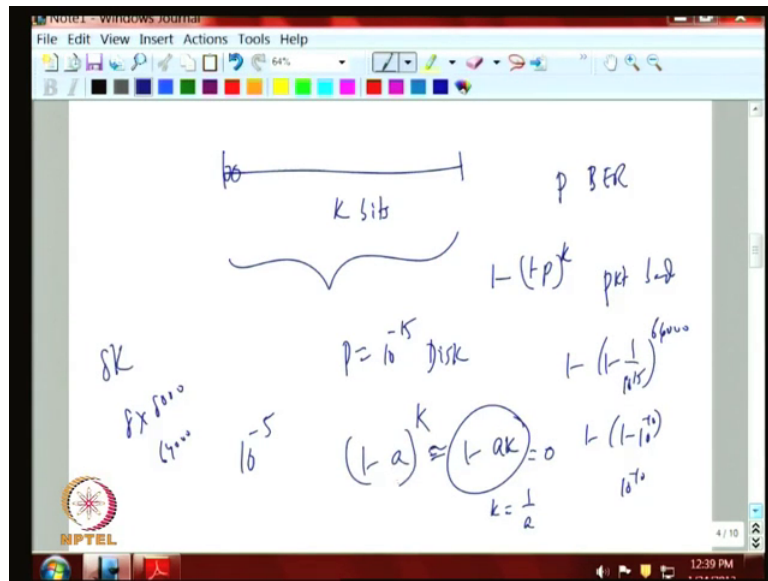
Whereas you use this NFS server kind of model you will know that everything is as per what is the latest thing is sitting NFS server and since it is a single big system you can afford to put RAID another things RAID another redundancy models out here much better. You know and do it on every single is small piece small device we have. So, this also is becoming starting to become common.

So, similar to this home NFS server or a department level and organizational also you have storage of this kind right this is very common. I think for example, you are and you are in the department you are using NFS server right and NFS is the Unix based model there is a windows based model called CIFS they are also widely people used ok.

So, now the difference here is that as I mentioned before after this point it is all local to this, it was basically local to the system basically direct attached storage most of the times all this while here and from here onwards you are starting to see some kind of networking also in the picture. The home NFS server and NFS server say department level organizational storage for example; these things all depend on networks.

Now, you can access these things through ethernet or through wireless (Refer Time: 32:21) especially we will take home NFS server right most likely you are accessing it through wireless, you are accessing it through Ethernet. Now there are some issues here also because it turns out if you use NFS servers, typical NFS servers have what is called 8 kilobytes as the block size. Now 8 kilobytes is a slightly big packet for wireless why is that? You can start thinking about that this way, if you look at what is the bit error rate of a some packet?

(Refer Slide Time: 33:04)



Suppose I have a let us say K bits alright.

So, what is suppose there is a bit error rate of P what this is what you call bit error rate. Now that means, that any bit has a probability of being bad let us P let us P prominent the question is if I send K bits what is the probability that the packet is damage some way. So, it turns out you can easily analyze this. So, what is the probability of? So, if you look at this right this is the probability that this is characterize 1 minus t right this also is 1 minus t all right. So, if you have K bits the probability of everything may characterize 1 minus t to the power of t the probabilities of thing is bad is this all right the packet is damaged in some way packet is bad right,

So, now if you think about it even if P is very small, let us say P is luckily for us P is 10 to the power of minus 15 for disk. So, if you do this 1 minus, 1 minus, 1 by 2, 10 to the power of 15 to the power of let us say K is 8 kilobytes, now a 8 kilobytes 8 into 8 bits into 8000 around this 64000 bits.

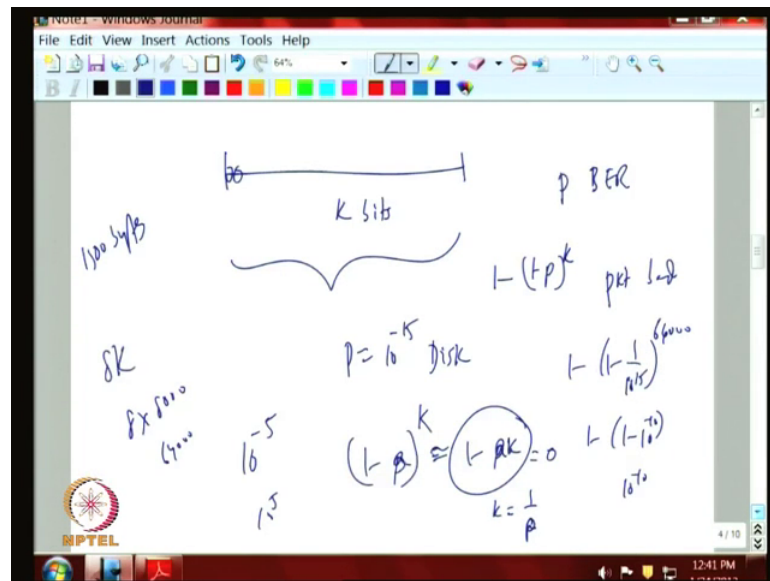
Now, you have a root sign you have a approximation 1 by K is what is it approximately equal to 1 minus a k. So, this is how much? It turns out to be very small its 1 minus approximately 10 to the power of minus sorry I made a mistake here what is that this one is what? This is 1 minus 1 minus 10 to the power of approximately you will make it also as 10 to the power 5 to the power of minus 10. So, the probability that the packet is

balanced 10 to the power of minus 10 that is very small with the case of disk is really small.

Now, suppose in the wireless case what happens in the wireless case you know that the wireless is a bad medium what is that meant is a bad medium? You notice that the signal keeps changing on real time you go up here suddenly you can get something, you go somewhere else you cannot even get a signal. The bit error rates are much higher there as much as 10 to the power of minus minus 5 , minus 4 depending on situation minus 8 (Refer Time: 36:06) it is highly variable.

That means that if I send a packet let us 8 K packet what will happen? You use the same formula again. So, you will see that at a particular rate right especially again we will take the approximation $1 - ak$ at of course. So, if this will be almost 0 when K is equal to 1 by k right and I am sorry I used the P here and K here let us make a P here again ok.

(Refer Slide Time: 36:45)



So, let us call it P here right ok.

So, if it is a 10 to the power of minus 5 it 10 to the power of 5 bits there is definitely about mean error and we are talking about 64000 bits; that means, that packet even if sent most likely will be dead a good chance because a probably so high; that means, that if you think about NFS may be it sense it in 8 K byte chunks, it is not a well suited protocol for this kind systems ok.

So, that is something you have to worry about you have to its true that you can always use it, but you have to be aware that wireless of different mediums. So, if you are building your whole NFS all right sorry you may have to change the protocols are bit you may not have to have 8 K right sizes, you may need to do it small. Luckily for us the issue is that you are going to send that you cannot send 8 K as such you know wireless packet you will send it 1500 bytes data

But a problem is you send 1500, 1500, 1500, 1500, but there could be because you are essentially talking about ability to sends so many of them intact. Only then the stuff is got on the server side and it has to rights it. If any one of them falls apart then you are dead. So, again it has been very transmitted. So, to reduce the retransmission rate you had to figure out a way or in which you can reduce the size of the NFS packet; that means, that instead of NFS which standard size 8 K you should start writing in smaller size probably if you knows 500 bytes, 700 bytes (Refer Time: 38:45).

So, that is you sending one single packet and the chances of your error is not let us say almost one, but if close to one in 10 to the power of 3 or 1 in 10 to the power of 2 that is 2 manageable. Let us one of the things I should remember that everything in a system when you think about it carefully, things are all varying depending on a technology parameters you have to be very careful that you choose the right parameters and a it you put some technology in a new setting you have to start thinking about what was the way it was what was the design for. I am using it in a different context, am I violating those design assumptions I am violating it most likely you have to again redo the thing there is no way to get it right. So, that is something which I wanted to do (Refer Time: 39:32) clear about, ok.

So, again if you look at all the stuff, you will find that as I mentioned I used wireless here it is a bad thing for it, but if you go to departmental organizations etcetera what do you have? You have either nowadays gigabit ethernet sometime in the positive 100 megabit per second and even earlier 10 megabit per second, but 10 megabit per second NFS is very bad it is not very nice. 100 megabit it is reasonably good little bit is quite good and you can people are going to move to high speed, but one gigabyte is quite good you can actually look some very good essentially network is no longer a bottleneck 800 1 gigabit, ok.

So, there are issues here also, basically when there are NFS has some time outs. So, if you write something and somehow it does not make it other side, it makes of 5 minutes or three minutes and then again retransmits it. So, in a sense what happens is that your timeouts are sufficiently large because the medium is especially gigabit ethernet is a very good, we were talking about wide medium we are not talking about wireless medium. Reducing wide medium you can keep timeouts to the substantial you can keep it for 2 minutes 3 minutes etcetera.

Whereas if you have a spec if you are doing gigabit ethernet wireless which also is going to consume, then you might want to keep the NFS parameters slightly smaller because the chances there are high therefore, should be able to respond to it a bit first you cannot keep waiting for it. You wait for you wait for 3 minutes, in the 3 minutes you could have sent so much right.

So, you should detect errors earlier and therefore, we should if you find some error you should ask you to retransmit a bit earlier not wait for three minutes that is a bit loss. So, there are some issues have that kind of you have to work out. Again here also again you will notice that you will start changing with respect to that kind of technology use I mentioned here NFS home NFS server you are going to use things like wireless medium most likely because wiring is a big issue and you do not know do all those things whereas, here most likely you are in organization where everything is wired.

Now since it is wired you can have different types of cables here. You will be doing using what is called typically if you use gigabit ethernet you will be using what is called some fiber or basically the you can in the past used to be what is called coaxial cable, and basically use the principle of total internal reflection so that it is essentially no attenuation that is how it was used in the past now fiber are chatter to the something.

Coaxial cable is basically you know very good for transmitting electromagnetic waves and now it has come to fiber optics, fiber what is called optical fiber and used optical fiber essentially you transmit also using total internal reflection. So, that you can drive bits using optical you know wave forms, you can do it with very low error, but there is a problem here what is the problem? The problem is that you have to move from convert from electrical signals to optical signals somebody has to do that ok.

So, normally what happens is that in low cost systems since things are nearby you do not want to take the trouble of doing it using the conversion from electrical to optical and vice versa it is too costly. So, usually what you do you what is called gigabit on copper that is you still do it using wires only.

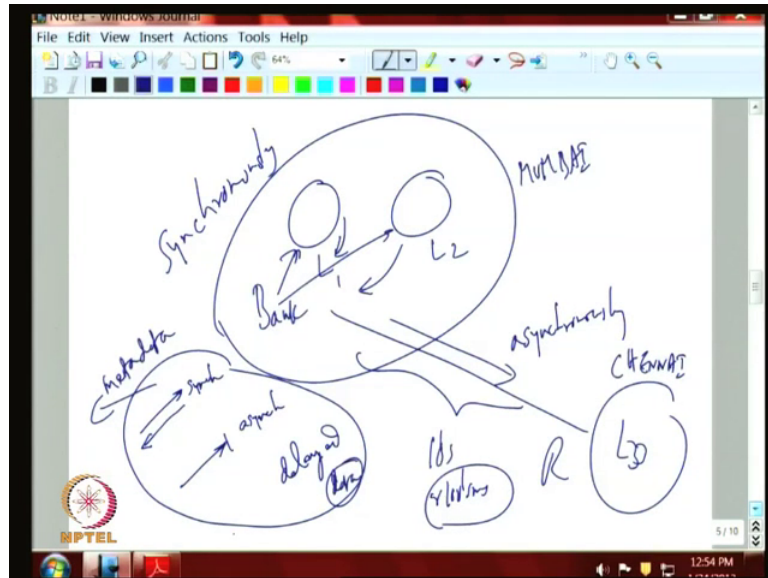
So, right now 100 megabit etcetera are goes on copper, all what you have everything what you have goes on copper. That means, is that cable are of copper if you look at the cable right there are multiple lines and one of them the conducting thing is copper. So, in gigabit also as long as the distance are not too long you can use copper and that is usually a low cost solution; that means, you do not have to have conversions whereas if you are going to large scale organization storage, the distances can be substantial. Now it turns out you cannot really drive the signals well on long copper it is not possible. And why is a needed? Because if it is an organizational system you need to worry about availability of the system; that means, that you no longer happy with just RAID systems; that means, that you need to also have a system which if the locals storage falls apart fire or whatever right a copy is sitting there at least some kilometers away. So, that is the fire happens here very unlikely for you happen there also right separated by 10 kilometers, 30 kilometers, 50 kilometers some forever want to keep this storage.

So, if you use that kind of model right then you cannot use copper for it what you have to use you have to use a fiber optic cable; that means, that you need to you will translate from electrical signaling to optical signal to transmit it . So, this thing becomes slightly costly. So, the various ways to convert it this standard thing is something called SFPS and that cost to it and on, but it also all high end storage especially if it has to be driven from distance has to use optical means only, ok.

So, this is something very critical for disaster recovery. I think you might have heard about some hurricanes and this thing that which essentially wipes out things within a space about few kilometers. So, you have to keep it at least across some if not hundreds of kilometers 50 kilometers at least you have to keep it right then only there is some chance that you can; which have a big buzzer realtor phenomena right suppose you have a Mumbai, Downpour in 2005 I think right it to dropped something like 910 millimeters in 24 hours right something else right some of you might proven with this. Means whole of them Mumbai is flooded it is no escape. So, you cannot escape a local copy of Mumbai I have to keep it somewhere else.

So, for that you need this optical transmission because that zone, but there is a problem there also let us just quickly look at what that problem is. So, you have a copy.

(Refer Slide Time: 46:55)



Here right let us say error bank. You keep some piece of information here is a local copy you keep another local copy let us call it L 1 L 2 there are two local copies this is in a single city and then for disaster recovery we will keep a copy L 3 just some protestant for example, this could be a Mumbai and this could be Chennai; that means, that any transactions we do you want to make the transaction the results of it available L 1 L 2. And this let us call it remote r should L 1 L 2 R, but you notice that if you want to keep everything consistent, then this takes a long time it can take tens or hundreds of milliseconds it takes hundreds of milliseconds; that means, that you can (Refer Time: 47:51) must do 10 transactions per second which is tolu, whereas you can probably do some thousands of transactions here, because the things are in the same city.

So, that is why you need to have different models and the model that often people is this called you do it transactions here synchronously. I needed to this what is called asynchronously; that means, that here whenever the transaction you do it here and do it here and wait for the confirmation from both sides, only then you say the transaction is committed. Whereas, here you do (Refer Time: 48:39) send it do not wait for it and you try to come up with a model by which we can in case something happens some failure occurs let us get a fix it up based on copies here you figure out what to do.

Because this one is going on synchronously only this kind if it fails right you may have with something about it. You sent it did not overwrite it; that means, you have to figure out what happened here and reapplied here in the sense this is going to strictly like a backup situation. So, various models like this are done and this idea about synchronous and asynchronous very common. I think even in the case of many systems many disk operations or not synchronous then synchronous what happens disk operation? Idea synchronous writes means, I write to the disk and I will wait for the disk to say I finish writing it that is synchronous.

A disk also have asynchronous mode you just initiate the initiate the request say please write, but do not wait you proceed immediately asynchronous there is also one more model which is called delayed writes. The thing is you do not even cue it you just keep it in memory, here at least you went to the disk and said write it, but I am not waiting for it here what you going to say is that I know this case too slow it is 10 to the power of 4 times or 10 to the power of 5 times slow right. So, what do you want go to disk, I will keep in memory I take my chances ok.

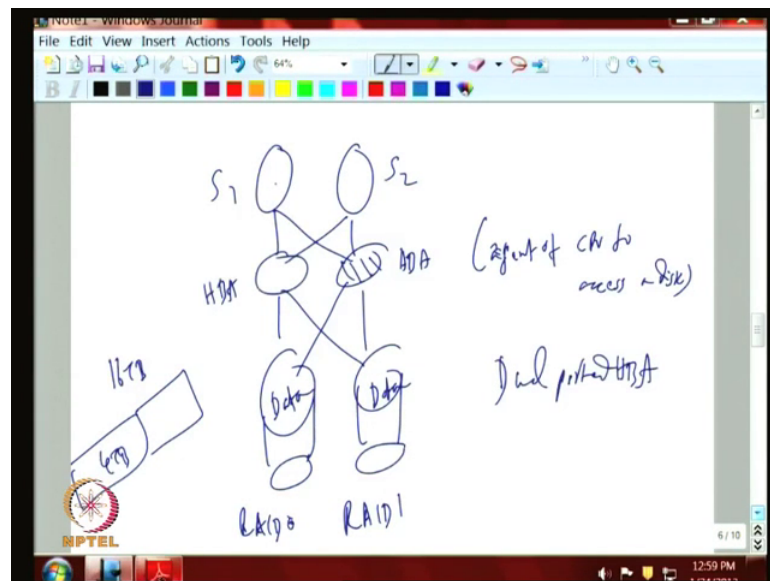
So, in case something happens? I will have somewhere to recover from that situation. So, I will use some stuff in memory and I will make sure that I keep some check pointing information. So, that even if something bad happens now able to recover from it. So, these things are commonly used in file systems; synchronous writes asynchronous writes and delayed writes. Typically delayed writes service for data in file systems, synchronous ones are usually for metadata typically, asynchronous in many situations for both metadata as well as for data since it depends. If you look at a data base they are not comfortable with these two models, they usually do synchronous because usually I am dealing with money a database much usually your bank account all the current stock you cannot afford to lose data.

So, usually the data there is typically because in a database data is more important than how the system stores information all right the metadata for data storage. So, data has to be stored synchronously. So, data there is have a different model and. So, there are issues of this kind latency issues that are very critical in the design of the systems. So, you will start noticing these things once you start looking at more closely ok.

So, that is that part, you can also have what is called highly available storage. Again when you talking about large scales storage this NFS CIFS server they usually can handle some amount of let us say availability, but we you might want to design especially some designs which are highly available like I just mentioned about this bank system all right you want to keep it or multiple copies, but this highly available storage actually tries to do it at every single level ok.

So, what we did previously was a slightly simpler model all right. So, we are talking it a bit abstractly about copies of data, but in practice it can be something more, let us say you need to take a much more detailed look. Because when you talk about a failure you are not talking about a system fill as a whole as a sink of a system, there may be multiple parts to it the question is you have to look at each part and you said what to do with it.

(Refer Slide Time: 53:11)



For example you might have two servers and then the server has what is called HBA is a server, S 1 S2 HBA as I mentioned earlier what is HBA? Host bus adopter it is essentially an agent for agent of CPU to access a disk all right its an agent S 1 HBA and then it talks to data. Now what is the situations possible now? The server can fail HBA can fail the disk can fail right. So, what is the solution to this? In case the server goes away and this data only accessible S 1 there is a problem right; that means, that until the server comes I cannot use it again, but the data is available nothing will happen to it.

So, question is it possible for me to even if S 1 fails can I access data. The way we do it is by probably providing one connection like this what does it mean it means that even if this is dead this HBA is working this HBA is working these two are working. So, this S 2 for example, is still alive it can use a HBA, if it is what is called dual ported HBA what it means is that? It has connections to both these this data.

Basically what we are talking about is. So, there is a discount here basically I put RAID 1 here let us say to make it simple RAID 1 hum. So, there is two disks data here. So, idea is that if this fails I still can S 2 can access this path and get to this data ok.

So, of course, if it is the case that this RAID 1 then a copy is also sitting here; so I can actually access like this say, but suppose its not RAID 1 the different piece of data here then I have a situation that S 1 fails S 2 can actually go and get this data because there is no copy of it here, if it is what is called RAID 0 what is RAID 0? It is basically concatenation you have a some root for some data let us say a 4 terabyte its 16 data 16 terabyte let us see.

Now, the maximum biggest size disks 4 terabytes currently, you cannot put a 16 terabytes and 4 terabytes. So, you have to join them. So, disk is RAID 0 that is basically what you get is you get a 16 terabyte disk, which looks like a single disk that is RAID 0. So, if it is RAID 0 that means, this data and this data there is no commonality. So, if something happens here unless you have this path there is no taxes are digital ok.

So, now there are two ports one port is this data and one port is this data. Similarly HBA also can fail this also can fail all right; that means, that you need to have you might have dual ported systems here. So, you will find that there are multiple paths around. So, that in case if either a failure server HBA are data you can survival thinks, this is what is called a highly available storage.

And I think I already discussed previously about web scale storage the Google file system is a good example I am not discussed parallel storage which I will cover a bit later. So, I just wanted to give you a high level introduction to this store. Now we will dive into each of these things more systematically. We will start looking at this NFS and those kind of systems and then we look at how to get these things it one about high level storage parallel storage and this one.

Thank you.