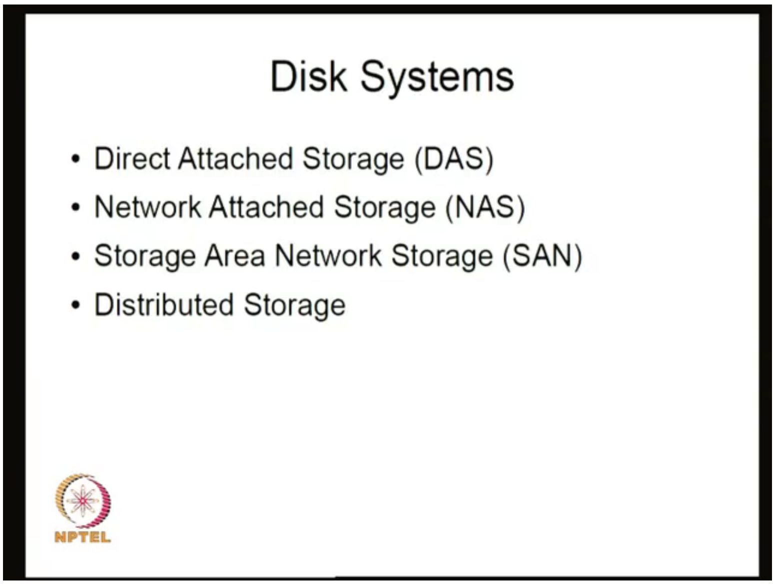
Storage Systems
Dr. K. Gopinath
Department of Computer Science and Engineering
Indian Institute of Science, Bangalore

Types of Storage Devices and Systems, Long-term Storage
Lecture – 12
Storage Models, Storage Devices, Tiering Concept

In the previous class, we looked at USB systems; I will briefly mentioned something about disk subsystems in this class. And a few issues relating to how we can mix and match different types of devices.

(Refer Slide Time: 00:39)



Disk Systems

- Direct Attached Storage (DAS)
- Network Attached Storage (NAS)
- Storage Area Network Storage (SAN)
- Distributed Storage

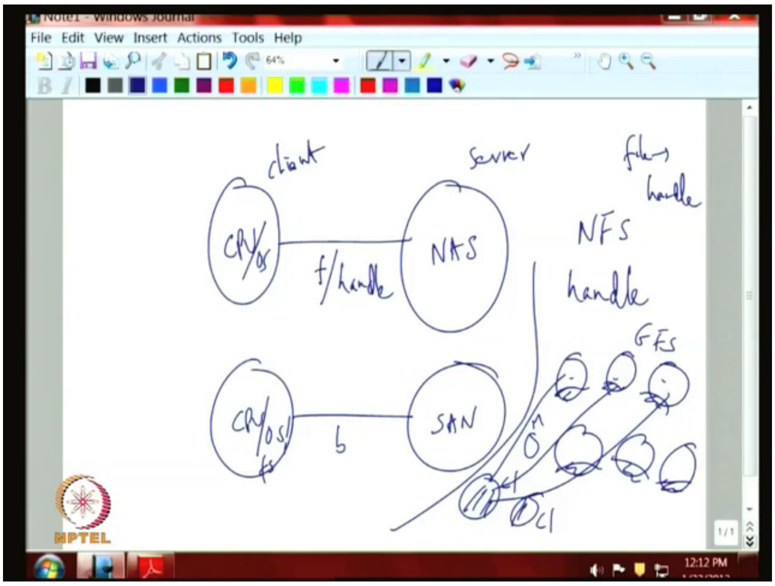
I think in disk systems there are varieties of them you can see one is called the direct attached storage, what is this? This is a typical pc kind of situation in a pc you have a disk attached to the bus pc bus pci bus and this bus called direct attached storage that. Of course, in larger systems, you will have many more disks attached to the same bus.

For example even on your pc typically you can put in about easily 2 days or 4 days easily do it you want to put more than that you need some additional hardware tribute do such a thing, because usually space is not available and you might need what is called enclosures external enclosures.

So, basically direct attached storage means, it is it storage is more or less connected electrically through some buses and it is very near the CPU itself. You can also have what is network attached storage what is called NAS, this is the one which uses what is called a file protocol what is the file protocol mean? It means that you are the CPU or the processing side most about file names and it asks about ask to the NAS system saying that please give me this particular file, etcetera, etcetera ok.

So, this is at the file level, the protocol is at a file level, it is also a something called a storage area network storage and these at what is call the block level. Here the requesters will only ask in terms of block numbers not file; typically a file is composed a many blocks. So, storage area network basically it is accessing things at the block level not at the file level, but this was accessing at the file level ok.

(Refer Slide Time: 02:51)



So, essentially if you look at it, you might see the following in the case of; these are the; what let us call the CPU right and this is a let us say the NAS. Then here your only ask work files in every things in terms of files and somebody has to map it here from files to blocks or whatever the way it has store, but the CPU does not know anything about the block structure what knows on you about fineness.

Means the other one this is basically the NAS system in the case of san kind of system this is the good block level. So, there are a plus and minus points both of these things, this one is somewhat simple basically because most essentially CPUs a most of the CPUs right basic most of the operating systems let us call it; that means, most of the CPU should call it operating system actually. Most of the operating systems have file systems or the come part with that. So, basically the file system all the basically what is happening at the file system is no composed of 2 parts, what is called a client the client on this side and then a server on this side.

So, for example, we have a system called NFS network file system there is a client NFS portion on this side in the server NFS on this side. The client NFS takes care of all the aspects written to issuing the request and the server NFS it figures out given the file name or actually what is called a handle. Basically there is a conversion between file to handle it happens. The beginning you give the file name and then the NAS system gives a handle and then later on we can you use the handle as a way to talk both the particular file.

So, I will to make it more accurate it is file or handle, that is how these 2 interactions are; here in this case it is strictly block level; that means, that in this particular OS actually should keep the maps about where the blocks are. So, responsibility of this particular various slash file system, there is a file system out here right it is there responsibility on this side to figure out where the blocks are.

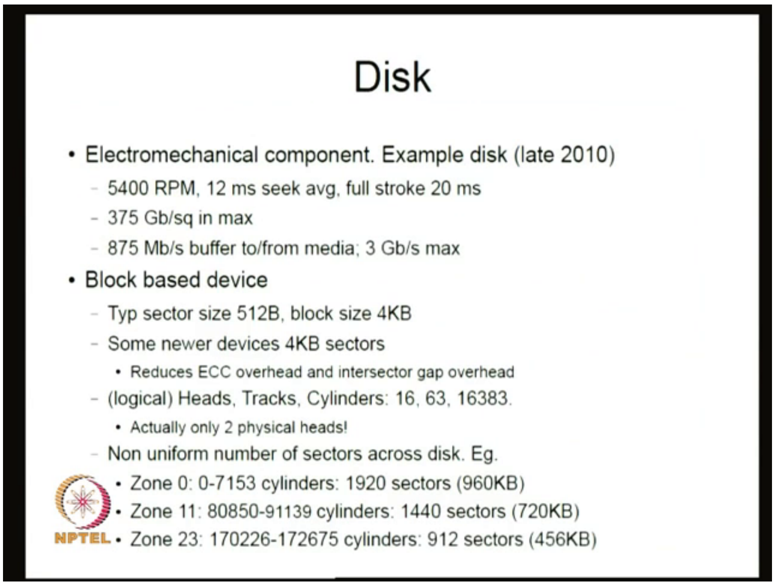
So, that allows for careful management of the resources, but its more work on this side and it also turns out since it is you on lower in the layering, this san does not have any idea about how these blocks are being used; because it is does not you know that is a file is there, basically file is composed of many blocks san has no idea about which blocks are actually part of one single file whereas, here the NAS knows which files have which blocks.

So, it is more information on this side. So, NAS can do some more straightly more intelligent here whereas, san might not be able to do that much here with respect to where the blocks are in which files etcetera because all that mapping is kept by here. So, the sans there is more interagency in this side and this have to be geared for some other types of operation, which emphasizes its capabilities better not at the level of intelligence about whether they are the blocks are in which files.

It is also as something called distributed storage; basically if you look at Google file system they do not use either of these things. They use a distributed storage in this model what we are talking about is, you have a full nodes each other have some storage, let us call this some storage and a client there could have many clients, this could be clients and they basically go through what is called a metadata server let us call it M, the first talk the metadata saying that I am introducing some piece of information and then metadata says your information is present here, here, here, and then the clients will directly talk to this guys you know after that they do not go through metadata.

That means, that initially the storage is the distributed across many nodes and the metadata keeps the information about where they are a clients have to contact the metadata to figure out where things are first, after that they can directly access there. This turns out to be very scalable that is why the Google file system which will talk later, you should this model of course, that many models, but is in a good example of that distributed storage.

(Refer Slide Time: 08:25)



Disk

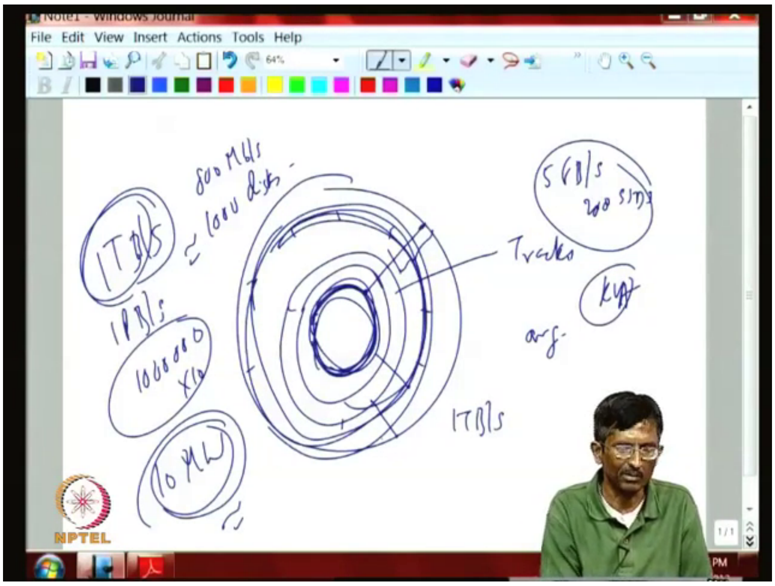
- Electromechanical component. Example disk (late 2010)
  - 5400 RPM, 12 ms seek avg, full stroke 20 ms
  - 375 Gb/sq in max
  - 875 Mb/s buffer to/from media; 3 Gb/s max
- Block based device
  - Typ sector size 512B, block size 4KB
  - Some newer devices 4KB sectors
    - Reduces ECC overhead and intersector gap overhead
  - (logical) Heads, Tracks, Cylinders: 16, 63, 16383.
    - Actually only 2 physical heads!
  - Non uniform number of sectors across disk. Eg.
    - Zone 0: 0-7153 cylinders: 1920 sectors (960KB)
    - Zone 11: 80850-91139 cylinders: 1440 sectors (720KB)
    - Zone 23: 170226-172675 cylinders: 912 sectors (456KB)

Now, with that introduction let us look at a disk, first thing important about is that a disk is an electromechanical component this is different from.

What we looked at before the flash which is a semiconductor based technology these are electromechanical and disk is. There is both electrical aspect in magnetic aspect in the system. So, if you take a disk of late 2010; it is 5400 RPM revolution per minute its making, 5400 revolutions per minute. So, you can think of it per second how much is it making you can divided it by 60 how much is that? It is about 90 revolutions per second it is spinning quite fast.

And if you call a what is called as average seek of 12 milliseconds; that means, that if you have a disk and go from half the distance on the disk your head is let us look at the picture again.

(Refer Slide Time: 09:46)



If you look at a disk it has got from concentric tracks these are called tracks and your head can be from anywhere here all the way to outside. So, the average half distance any half distance from here to here wherever from here to here that is usually given by the term average seek time half. An average you could any were right and you want to reach any other place and average you have to do half the number of tracks into cross. So, that basically what is the 12 milliseconds; what is call full stroke means go from one extreme went to the other extreme went that is full stroke that is call a milliseconds.

There are of course, some minor differences actually terms out read takes slightly lesser than right for example, example in this disk right takes 13 milliseconds even take 4 because it has to writing is a slightly more involved operation. So, for that reason it takes one extra millisecond every point thing that is interesting for us is you will see that there is a number of bits per square inch now in this disk of 2010, it was 375 gigabits per square inch and you take one square inch you are packing in about 375 gigabytes.

Now, days you have terabit per square inch; people are saying that beyond that is going to be tough, but there are other new technologies that are coming in as you probably give you pack it in better.

So, you can just remember it has one third of a terabit per square inch of approximately one third of terabit, but currently already have terabit per in square. Respect to the throughput you can get about 2; 1 from media, now inside the disk not only is the magnetic medium there, there is also some electrical buffers may it be (Refer Time: 11:54) economy. So, now, what happens is that? You are on a bus you transmit to the electrical component first and from electrical component to get to the magnetic media that is why we talking about from buffer inside a disk to the media that is the magnetic media wise back and forth that is about 875 megabit per megabit per second.

But if you are talking from the CPU side, which is coming from the typically from the electrical side and then you can talk to the electrical buffers, semiconductor buffers electrical buffer whatever we call it. That one is much higher basically, the media is the bottleneck the magnetic media. That is coming that you can do for only 875 whereas, if it is only electrical kind of transfers we are doing about almost plus 2 4 times faster 3 to 3 and half times faster ok.

Now, I can this what is called a block based device is typical sector sizes 512 bytes block size is 4 kilobytes typically again as we discussed last time, in Google file system the block size is 64 megabytes so that you can get effective throughput. Nowadays this sector size in for a long time has been 512 bytes, nowadays some since 2010 we are getting your devices with 4 kilobyte sectors of course, most people still use it as 512 byte things only, but basically why are you going for higher sector size, this like the way we went from 4 kilobyte to 64 megabyte for Google file systems, it also there are reasons why people want to go with higher sector size. It also out it reduces what is called error correcting code overhead, again example if you take 512 byte sector size you need about 40 bytes or ECC reed Solomon code ok.

So, it is all most talk about 40 by 512 it is about 12 percent approximately, you lose 12 percent just in the ECC overhead. Whereas, if you go from 512 byte to 4 kilobyte it turns out you need only 100 bytes. So, since a (Refer Time: 14:16) by 100 bytes you are losing about 4 percent. Efficiency in this case 4 kilobytes if 96 percent in this case at 188 percent or 87 percent.

The real is that the amount of ECC Reed Solomon code a require is not linearly proportional to the sizes the it actually turns out that with slightly larger ECC size you can do much better, you can handle bigger sized sectors. It was always here this is it something called intercept of gap overhead. Basically if you look at a track every sector of to the completion of every sector there is some markers so that you can position the head you need to forget the sector and sever the sector begins. So, the things there is some specific markers on the disk and that is usually that is where the inter-sector gap is. So, the more number of sectors the more gap has to be kept ok.

So, that also eats into the space. So, if you go with bigger sector size it turns out you reduce that overhead also again. I think some of you all must have looked at all these things; there in the disk we have something called heads tracks and cylinders. Now these are all turns out to be logical conceptual mean because physically they are oftentimes different. Real why it has happened is because in very very old disks 1980s approximately you had in the bias for example, there were fixed number of bits for heads and tracks and cylinders for example, there were something like 4 bits per number of heads etcetera and it turned out that because of the fix size they could not access more than something like 500 megabytes in the beginning sometime in the 1980s. So, there re interpreted these bits as in some different ways, so that even if you had only 2 heads, we still call it 16 heads and still were protect ok.

So, for example, on this disk it says that there are 16 heads 63 tracks and so many cylinders. It is completely let us say this will correspond with reality, it turns out this particular disk has such will only 2 physical headers heads. You need to physical heads and you can look at the number of cylinders that also is way of. Basically it has if you look at it has got 172675 cylinders, it is about a 10 times as much is this it is; it place that has got 16383 naturals got 17267 something all that kind, and basically what is happen is that head says when a 2 is importing as 16 and it has compensated by going with larger number of sorry smaller number of cylinders when actually more number of cylinders.

So, if you multiply all these things if the same number of bits we can terms of the size something all that kind, basically the number of bits have been reused in a different way.

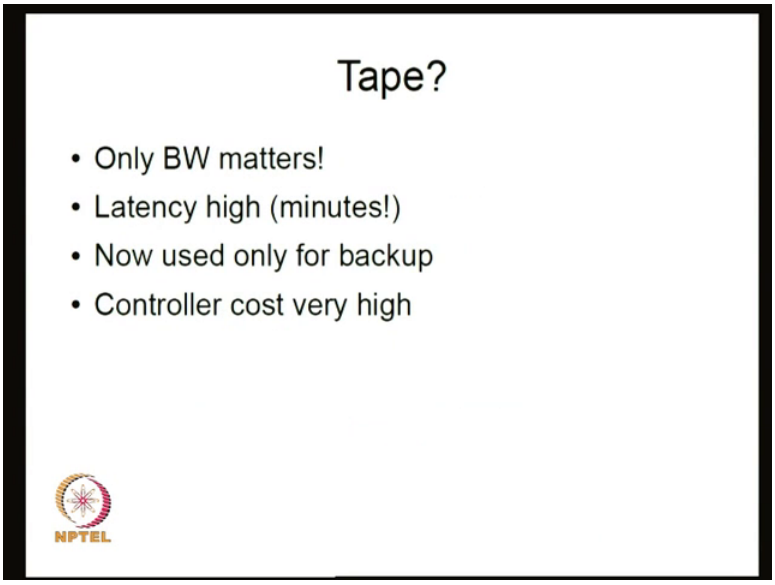
So, that is why the there are logical concepts physically they are different. There is also interesting about disk is that, it is the number of bytes on each track is different. But thing as you can imagine if you look at this picture for example, right this area is this linear dimension is quite different from this linear dimension. Look at this, this is much bigger than this one. So, if you are going to have constant density there should be more bits here than on the track than on this track. Now this is when the past in the very earlier devices they somehow did not like that. So, they would actually have the same number of bits whether here or here by making the bits slightly less frequent outside and keeping the density that they wanted inside.

But nowadays they do not want to lose the density. So, there are many many more bits outside than inside. We will see that it has been categorized into multiple zones 0 1 0 0 1 1 23 for up till 23. So, basically 24 zones, on this particular disk I am talking about there are 24 zones the first 0 to 753 cylinders have 1920 sectors this is outer most essentially you have almost a megabyte on each track not almost. But if you go in between you will find that it reduces it going to 720 kilo bytes, and also come to the inner most zone you will find that it has got these number of cylinders it has got 912 sectors, only half about a half a megabyte. So, inner most only half a megabyte ok.

So, now it turns out that this is all this is highly device specific. So, when you are talking out file system 0 11, but all these things because; that means, that for every disk that they see they have to configure the algorithms again. So, they do not really worry about all these things, there is assume that a disk is bunch of a sectors and they do not worry about usually the track or cylinder etcetera you just do not worry about all these things. Let us my design should be robust to; however, these things are divided. But some device some file systems actually do take some of this into talk out, if you look at them well known first file system FFS which is there also in Linux as ext 2 file system and ext 3 file system, they actually try to allocate blocks in nearby cylinders. They may not take into account the track size per say, but they will worry about the fact that it should be near whatever blocks are for a particular file is being allocated, they should be on nearby cylinders. That way they mean access time for when you want to move from one block in one cylinder to the next block, which happens to be the next endure the want of moment of the head can you image.

So, the basic problem of what we have what we are facing this is upper level software, does not make me want to does not make going to too many of these details. You wants a get too much into details then it means that it is highly specific to the disk; that means, that I cannot take some other disk can install put this one make it, it makes a difficult. So, they had to be slightly abstract a ways some of the details. So, some file systems just say that my abstraction is only blocks, some of the file systems will say that at least I will think about in terms of cylinders, but I will not think in terms of tracks and other things.

(Refer Slide Time: 21:14)



Tape?

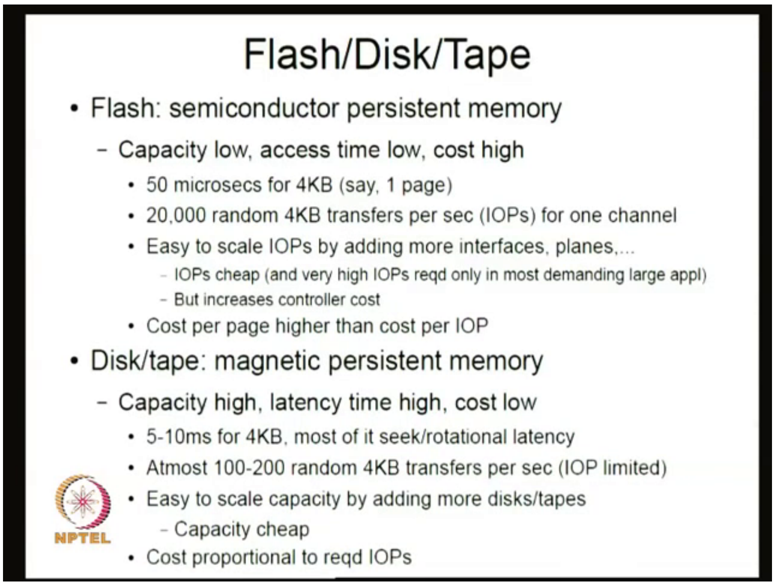
- Only BW matters!
- Latency high (minutes!)
- Now used only for backup
- Controller cost very high

So, that is a big introduction to the disk issue, let us look at tape. Here tape is a nowadays not that very common, it is to be common about 20 years back because you are we are the is to also use it to store our things, but nowadays disk has replace tape for storing place. The thing about tape is that it is used now only in very high end systems for example, if you talk to people in CRN the people who did that boss boson side, who are discovering some new types of bosons you might heard about in some atomic say nuclear facility right research facility.

So, they generate. So, much data that anything they had to back up, they cannot afford to back it up in disk it is too costly. A tape is about a factor of about 3 4 times cheaper. So, tape is being used in very high installations, but in regular installations what we normally use tape is disappear it is only for very large capacity systems very very large capacity petabytes 100 of petabytes. The problem with the tape is that it is got a very high latency, normally the latency is in terms of minutes sometimes it can be even the half an hour because somebody has to physically go get the tape and install it on the device and so, it call somebody on the person has to move physically and go and get it.

Now, only important thing about tape is that it is call a very high bandwidth. Once it starts moving it moves at very high speed, it can be some of the high speed tapes we can be have better bandwidth than you on disks they are very good at bandwidth; that means, that the minute you started, if you want a write a few chunk it can write it quite efficiently. But you cannot afford to stop it and rewind all this things, if you do not added you do not have efficiency. So, it is only is for backup also the thing is because of all the controller cost also is very high, basically is of very big systems. So, the electronics that control these things is very costly. So, we will not be discussing tape in this particular course that much or, but I want you to be aware that tape is a viable medium still because this is a very low cost for extremely large capacity storage.

(Refer Slide Time: 23:41)



Flash/Disk/Tape

- Flash: semiconductor persistent memory
  - Capacity low, access time low, cost high
    - 50 microsecs for 4KB (say, 1 page)
    - 20,000 random 4KB transfers per sec (IOPs) for one channel
    - Easy to scale IOPs by adding more interfaces, planes,...
      - IOPs cheap (and very high IOPs reqd only in most demanding large appl)
      - But increases controller cost
    - Cost per page higher than cost per IOP
- Disk/tape: magnetic persistent memory
  - Capacity high, latency time high, cost low
    - 5-10ms for 4KB, most of it seek/rotational latency
    - Atmost 100-200 random 4KB transfers per sec (IOP limited)
    - Easy to scale capacity by adding more disks/tapes
      - Capacity cheap
    - Cost proportional to reqd IOPs

Now, now we already seen three types of technologies flash disk and tape. Flash is semiconductor persistent memory, what is a capacity what is it characteristics? Its capacity is low compared to disk for example, an access time is low compared to disk cost is high compared to disk. If you look at a disk or tape it is a magnetic persistent memory, capacity is very high compared to flash, here we talked about access time here we talk about latency time because it can involve seeking also ok.

So, here is just that is not on this hardly any seek or misses things you do not have those aspects here, but here we have not only the access time, but also the c can rotational latencies and the cost is manageable it is not higher compared to flash of course, your talk of memory itself dram. In dram the capacity even much lower, access time is even much much lower and the cost is much much higher. Remember that we said that when you are designing storage systems, we have three aspects capacity latency and cost you can usually have only 2 at a time you can get all 3.

So, again we have to figure out what has to be done, now just keep into look at a what is the way we can try to exploit the characteristics to our advantage. So, if you look at flash for example, you can read about to read one page let us say point pages for kilobytes 15 microseconds. This is much better than for disk it is 5 to 10 milliseconds, is it because it includes seek and rotational latency. So, this guy is very bad disk compared to because of this you can essentially know about 20000 random 4 kilobytes transfer per second ok.

Now, it turns out if you take any SSD solid state disk, you have multiple channels that is multiple independent units within a same SSD. So, that you can essentially multiply the number of channels will get that much number of Io operations. These are called IO operations because typically in data base, you have to update 4 kilobyte chunks typically this is the your record for your bank account recorded these some small amount 4 kilo bytes whatever right it has because it has basically has to have information about your current account balanced our that is not much.

So, usually small numbers. So, the number of transactions should depends on number of 4 kilobyte chunks you actually you can update. So, if you are able to do fifty micro seconds for 4 kilobyte and you can 20000 random 4 kilobyte transfers that IP operations, Here so many k number of channels you multiply by k and 20000 and typically you can easily get 8 channels 8 things of that can 4 to 8 without trouble without (Refer Time: 26:41) 16 also easy depending on whether it is what is call enterprise SSD or consume a necessity. Consume a necessity is do not have too many channels, because every channel incorporates in they are independently operated; that means, that you need independent electronics for all this ok.

So, for enterprise flash you can go much higher people routine we talk about one million IOPs per in some of the more let us say expensive SSD devices you can get that much.

So, what is important about this IOPs are cheap here? Getting IOPs is cheap what is costly the actual the capacity, cost compared to disk the capacity cost is high that IOPs are cheap. So, you can as you mention you can scale IOPs by adding more interfaces planes on that things. So, IOPs are cheap and what is also important is that, very high IOPs are required only in most demanding large applications typically most applications we deal with do not required that kind of IOPs.

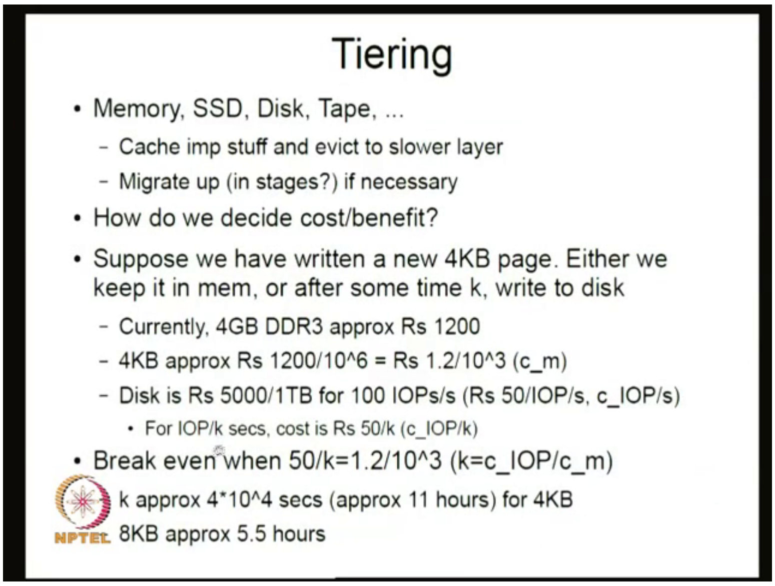
So, its only specialty systems that required very high IOPs if good example is from web service for example, suppose there is world football or Cric-info those kind of sites, there are thousands of people or millions of people accessing it, then for those kind of guys they have a very high rate at which information put full dot further IOPs is important. Unless you are one of those very big sites eBay, Google, Yahoo, Cric-info, World Cup, football, all those kind of pieces; you do not have to worry about IOPs in most of systems can handle IOPs. So, basically what the important thing you have remember about flash is that cost per page is a more important parameter than cost per IOP as per that way. So, the case of disk tape all those things it turns out because its 5 to 10 milliseconds per cool 4 kilobyte, you can at the most do 100 200 random 4 kilobyte transfers per second. So, this IOP limited.

So, if you are talking about some eBay kind of system, if they want to satisfy some 10, 000 request per second a disk cannot do it one disk cannot do it any to have if you have suppose you want to support 10,000 access per second. So, you have pullet say 104 kilobyte transfers per second is happen let us say; that means, you need at least have 100 disks there that are working concurrently then they can get to that currency.

So, your IOP limited. So, when you have to scale this is by putting more disk and tapes you scale both the capacity as well as IOPs. So, the cost is proportional to the required IOPs as I told, you if you have an eBay kind of system and somebody conceive that how to handle one million IOPs per second; that means, that you had have 10 to the power of 4 disks there is no choice you have to say if you are going to be disk base system, you have had power 10,000 disk appear then only you can handle it otherwise you cannot handle it that increases cost tremendously, it increases power it increases space requirement all those things that is where flash actually has benefited about. Basically if you put a flash there it handles IOP requirement quite well under look with low capacity itself.

So, for big capacity you put in the disks. So, given that the question for us is; how do we figure out how much of this flash should be there how much should be disk, how much should be tape. We will mainly look at the flash memory flashing tape issue, the tape issue will disregard for the time being because we can tape this as a only for backup notice it is not being used for anything else.

(Refer Slide Time: 30:43)



Tiering

- Memory, SSD, Disk, Tape, ...
  - Cache imp stuff and evict to slower layer
  - Migrate up (in stages?) if necessary
- How do we decide cost/benefit?
- Suppose we have written a new 4KB page. Either we keep it in mem, or after some time k, write to disk
  - Currently, 4GB DDR3 approx Rs 1200
  - 4KB approx Rs 1200/10^6 = Rs 1.2/10^3 (c_m)
  - Disk is Rs 5000/1TB for 100 IOPs/s (Rs 50/IOP/s, c_IOP/s)
    - For IOP/k secs, cost is Rs 50/k (c_IOP/k)
- Break even when 50/k=1.2/10^3 (k=c_IOP/c_m)
  - k approx 4*10^4 secs (approx 11 hours) for 4KB
  - 8KB approx 5.5 hours

So, let us. So, these were in the Tiering concept; comes in your memory SSD disk and tape. So, what is the basically I think most of you familiar with this? You cache important stuff in more expensive memory and elected to slower layer when they are not needed. More important stuff in the high speed memory if possible, where ever is possible whenever you cannot keep it around because there is a contending request to store that place in that same high speed memory you push that unfortunate gray out to the slow of memory right I know you can migrate up also, sometimes what you push down to lower slower layer, it may have to be it may be accessed again. So, it has to be migrated you can either migrate in stages form one to another for example, move from disk to SSD to memory or you can directly go from disk to memory it depends on your system of course, if you go step by step its more latency and cheaper possibly, but if you go from disk to memory is faster, but costly system. Basic fourth for us is how do you decide what is the cost benefit when do we how do we do this things.

So, just like a tricky and subtle issues. So, will just try to understand it carefully. Suppose you written a new 4 kilobyte page why is it happening you have an application? An application is generating pages all the time new information, the applications writing something is computing something right. Now if you compute a page, you have to store it in memory typically in dram it has to be stored first now it may be temporary memory after a few minutes of time it may longer be necessary it may can thrown of you or I compute something I need it for some time later much later. So, either we keep it in memory or after some time we write to disk, because when you want to retain to write to disk because we want to reuse that memory locations again for something you on computed because every so often I am computing new things.

So, now let us look at what our options. Let us look at I sometime back bought 4 GB of DDR3 what is DDR3 essentially it is a type of dram with multiple channels. So, what; that means, there are let us say 2 channels; that means, that I can independently access and there is this part of the memory or this part of the memory that in senses I can have 2 requests going at the same time. So, you can have multiple channel systems DDR3 is a multichannel system. So, you have multiple CPUs, multiple cores they can be pumping the memory simultaneously. So, 4 GB DDR3 are brought some one year back I think for rupees 1200 what means that 4 kilobits approximately, this much 12 00 rupees by 10 to the power of 6 right because 4 kilobyte its 4 kilobyte and 4 gigabyte are factor of one million way this is somewhat.

Let us call it also c cost memory c on the squadron is cost a memory. Disk is you can get a 1 terabyte disk now a days for 5000 rupees of course, if you are using what is called enterprise disk is a consumer disks, enterprise disk will be about a factor of 2 by 3 times more expensive all right we will just use the I think I should use the enterprise disk here because mostly I am talking enterprise disk, but anyway I have used the consumer grade devices here just had multiplied factor of 2 whatever I am talking about if you want to go for enterprise disks enterprise disk will basically more reliable better disk compare to consumer grade disk for example, you know consumer grade disk if something it dies you know. So, data it is suppose great loss, but you do not lose information like your bank account information also (Refer Time: 34:55) that is why they can it generally believe the itemization will be much much more higher quality than consumer devices.

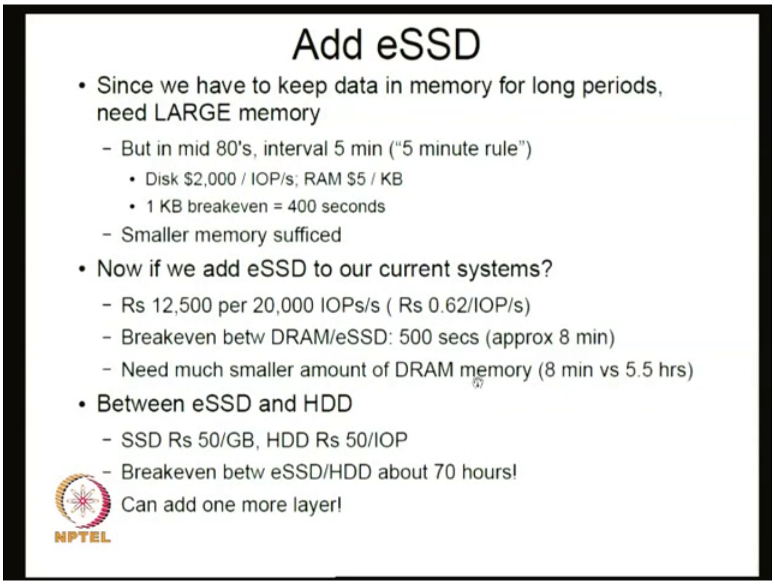
So, disk is rupees 5000 per terabyte for 100 IOPS. So, if you think in terms as mention we are IOP limited what is our cost of resource IOP not space so; that means, that I am paying rupees 5000 for 100 IOPs. If somebody want a 200 IOPs had played 2 device 2 devices 2 disk at to by is, because I cannot supply it with one disk and they most I can apply hundred IOPs somebody wants 200 IOPs there is observe away do it that I have to get 2 base then only I like I can do it. So, my cost a proportional to IOPs; that means, it is rupees 50 per IOP per second, let us call it c under square IOP this is the cost per IOP per second.

So, a this is a cost then it may be that I do not need IOP per second kind a cost I need IOP per k seconds; that means, a cost rupee rupees 50 by k that is instead of doing it in per in 1 second this is per 3 per second cost rupees 50 for IOP per second. If I am I do not want to do it in that smaller time, I want to do it much I want along it for be can I can more (Refer Time: 36:11) slower devices let us say. So, for IOP per k second cost will be proportionality less 50 by k. The question for this is I have a because I have to write to the disk right; that means, that the cost is just about the same then if you take this cost of that 4 kilobyte if it is about the same cost if I am going to write once every k seconds because I need to get a disk of that where is IOP writing. So, these causes are about the same an 50 by k is equal to the cost of 4 kilobyte this one ok.

So, this I am going to on a breaking on a this point. So, if you calculate it its basically if you in this using symbols only, it is basically k will be equal to c under squared and equal to c IOPs by k. So, k will be equal to c IOP by c under square. Now say look at these numbers it turns out if you calculate for k it is 410 to the power per second approximately 11 hours 4 KB. If you do it for 8 kilobytes this is 5.5 hours, what is it mean? It means that if you have generated some pieces of information, you want the cost of the IOP cost you are paying compared to the memory cost per 4 kilobyte and paying the costs are equivalent only if you store the data to generate for 11 hours approximately. So, if some information is needed I compute something and if needed after 11 hours, there I should put in the disk I am going to generate some information I needed within the first 11 hours there should keep it in memory ok.

Now, this is a slightly tricky thing here why because; that means, that I generate stuff I have to keep stuff what I generated 11 hours in memory in dram. It 11 hours have been generate amazing amount of temporary and other data I have to keep it around for 11 hours; that means, the memory dram requirement is very very high extremely high that is why there what I will see that nowadays, most laptops etcetera are putting more and more dram. Firstly, because if you want to have your system respond quickly you have to keep most of in memory let you have.

(Refer Slide Time: 38:50)



Add eSSD

- Since we have to keep data in memory for long periods, need LARGE memory
  - But in mid 80's, interval 5 min ("5 minute rule")
    - Disk $2,000 / IOP/s; RAM $5 / KB
    - 1 KB breakeven = 400 seconds
  - Smaller memory sufficed
- Now if we add eSSD to our current systems?
  - Rs 12,500 per 20,000 IOPs/s (Rs 0.62/IOP/s)
  - Breakeven betw DRAM/eSSD: 500 secs (approx 8 min)
  - Need much smaller amount of DRAM memory (8 min vs 5.5 hrs)
- Between eSSD and HDD
  - SSD Rs 50/GB, HDD Rs 50/IOP
  - Breakeven betw eSSD/HDD about 70 hours!
  - Can add one more layer!

Take into account let us say, let us say I had SSD. Basically whatever sayings since have to keep data in memory for long period 11 hours, etcetera, you need large memory that makes a system very expensive, but may be is that is in the immediate is it turned out that disk was much more expensive ram was remember expensive because.

For example in those days it has 2000 dollars per IOPs something I that and RAM was 5 dollar and if you do the breakeven for 1 kilobyte in those days one kilobyte was often the block size, it was 400 seconds; that means, that you kept stuff in memory 400 seconds which about approximately how much about 6 and half minutes.

So, if you access something beyond 6 and half minutes, then only keep in disk otherwise you keep it in memory; that means, you need to store what is generated every 6 and half minutes approximately in dram and those disk whereas, in our case it is 5.5 hour whatever you generated 5.5 hour. From may be is also that those CPUs were slower. So, the amount of (Refer Time: 39:55) generated to going to smaller where is in 5.5 hours our machines now with so many gigahertz and so many course, if generating tremendous amount of memory developed that they have generated right. Some of the this temporary get dropped some or temporary things which cannot be need not be stored across, but still a good portion of it has to be stored, that basically means that in those days things was slightly more balanced the disk memory was straightly mores it is not balanced, but there was some similar of balance. Nowadays it is total out of back, completely out of back because you have to have such a big memory to keep things when then only you will find the disk system means actually make sense.

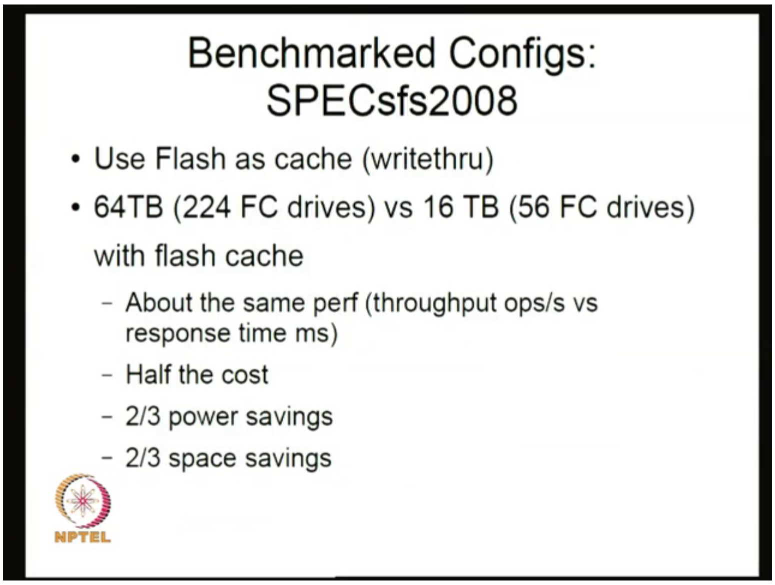
So, because of this we can start thinking about can you put in some intermediate kind of device, that is what we can do eSSD enterprise SSD. If you put it often what will happen is that, you will find that we are talking about some rupees 12500 for 20000 IOPs, it is our only rupees 0.62 per IOP. Previously what was the number it was may be 50 per IOP, this is very expensive compared to 62 paise per IOP. It we can see that tremendous difference in the cost. So, if you calculate the same thing breakeven between dram in eSSD it terms out be five hundred seconds how one the calculating breakeven and using the same formula here sorry see what is a breakeven c IOP by cm ok.

So, if I do that I put 0.62 divided by cm is what? This much 1.210 to the power and do that it turns out it comes out to 500 seconds approx 8 minutes. Because the difference because you know how to store approximately what you have generated your CPUs have generated what is 8 minutes that is all the storage that is all the memory dram memory require. So, need much smaller amount of dram memory 8 minute versus 5.5 hours and that being everything for 8 kilobytes. So, basically this was 11 hours for 4 kilobytes, it is a 5.5 for 8 kilobytes the calculated the things; that means, that you can reduce your dram requirement grammatically, you have SSD that is even by you will see that lower people are playing around with SSD. These are the between dram and SSD and this is corresponding that can dependent eSSD and HDD. So, what you done is you have memory followed by SSD followed by hard disk?

So, between dram and SSD this is the situation, between SSD and HDD it turns out the breakeven is 70 hours, even this also is tremendous one can imagine put in even more one more layer. So, people are talking about I have mentioned earlier, you have what is called single level cell SSD and then you have multiple level solid SSD and basically you can use this kind of additional layers, then you can start finding that the amount of stop that you need for example, what this means is that you need a very big SSD because that SSD has to observe all the rights for 70 hours. So, you might find that by adding one more layer formally you can candidacy to some more arrangement, because the more is bigger the SSD the more costly it is.

So, I hope this is clear? Such the people have started the because of the obvious advantages of these kind of systems because you can see the tremendous difference.

(Refer Slide Time: 43:56)



Benchmarked Configs: SPECsfs2008

- Use Flash as cache (writethru)
- 64TB (224 FC drives) vs 16 TB (56 FC drives) with flash cache
  - About the same perf (throughput ops/s vs response time ms)
  - Half the cost
  - 2/3 power savings
  - 2/3 space savings

People have tried to see what they can do with this kind of systems and what I mentioned earlier was that you can use the flash as a cache. Basically what is the situation here what we are doing is you have memory dram, SSD what you do is you keep it as what is call a right through cache what is a right through cache? Right through cache means you write it to the cache and also it goes to the slower area also; that means, there are 2 copies because there are 2 copies even if one copy fails you can pick it further one in right back there is a single copy unit.

So, if you use a write through cache what happens is, then you can essentially not worry if the flash fails and that is how people are using for quite some time for example, if you look at netapp other companies they have a flash what they call a flash cache and this flash cache are quite big there are about minimum 1 terabyte, 2 terabyte, 8 terabyte and this is things. So, if you use that flash cache, it turns out that for the reasons we discussed earlier right basically you can essentially match each of those layers instead of it being. So, totally out of whack right we can essentially match it better and people will also discovered have actually experimented this consistence for example, there is some see some what is called a benchmark, around some organization called spec and they were using a particular benchmark and it turns out to get the kind of IOPs, because they are talking about this particular benchmark actually measures through put member operations per second versus response time ok.

So, operations of very important in through put how many operations you do for getting a kind of operation require the only way to do it is by scaling number devices scaling the number of drives what is called further channel drives. A further channel drives are the most expensive kind of disks and they will run at 15000 RPM, you will first look at the previous on right we are talking about this is the consumer grade disk is 5400 RPM. So, further channel drives typically run at 50000 RPM. So, there are three times later; that means, that place about 12 milliseconds seek coverage, that there will be a factor about three times slower and 4 milliseconds. So, he is in because the 4 milliseconds, he is able to about 250 he can do about 250 IOPs per second and you are paying about a factor about 10 times a cost like approximately and your IOPs just increased from 100 to 250 that is all nothing more be that ok.

So, if you really want to get the kind of through put that you want operation, the only way to do this by increasing the number of devices. So, if you go to 224 drives, you get the desired through put versus response time in this case, and you get it tells out because the fiber channel rise coming the particular size. If you increase the number of drives for the kind of IOPs require corresponding the disk such also the total capacity increases it goes to all 64 terabytes; and take that kind of system you compare with the system with only one forth a number 56 for the sunrise, but with the flash cache with 16 terabytes ok.

Now, this flash cache probably be about 1 terabyte or 2 terabyte. If you go with this kind of flash cache you were it could even the 512 gigabytes, it is not work out it will depending on application it might work out that is just sufficient. You find that the performance about the same, what is very interesting as is it is about how the question include all the cost because I have mentioned to you that I use in flash cash, flash is expensive. Even if you take into account the cost of the flash, because of the reduction from (Refer Time: 48:43) for device 56 device, that custom cost quite a bit. Even in if you add the cost of the flash in spite of all these things it turns out the cost is about half and you get about most important put that small savings.

Student: (Refer Time: 48:57).

Why is that basically if you look at this 224 forbids all rights, we are basically talking about something like our 10 to 15 watts per drive and if you are talking about if you take if you take 15 drives for example, that means, it is about approximately 4000 kilowatts and I think you know about something called power factor. So, it is 400 kilowatts, it becomes approximately possibly 600 kilovolt ampere and. So, basically we are talking about a fairly big power system for this, and where as if you go with this you will be with a much smaller probably 1 kilowatt ampere kind of system ok.

So, power sense is there very important space savings also, on there because just imagine storing a truly for power savings right. So, there is a big aspect relating to the space requirement also that is why you notice that if people of building very very big systems. Some for let us say somebody says that I have to device a system with 1 petabyte per second bandwidth how much some people starting to think about all these things now ok.

So, for example, aggress let us not go with 1 terabyte per second or you will say if you say 1 terabyte or 1 petabyte per second kind of systems, this is a very big systems you will just take simpler a one terabyte system. So, how many disks are required for this? You will find that the number of disks required will be because typically you get about I told you 800 megabyte per second approximately maximum you can get. So, if you do this it will about approximately 1000 or more whereas, of course, is one petabyte it will be one million disks. If you put one million disk not will be the situation you will get? Multiply by 10 what is the power required? Is 10 megawatt. Now this is this like an (Refer Time: 51:31) but actually 10 megawatt you had cool it also; that means, you need air conditioning also.

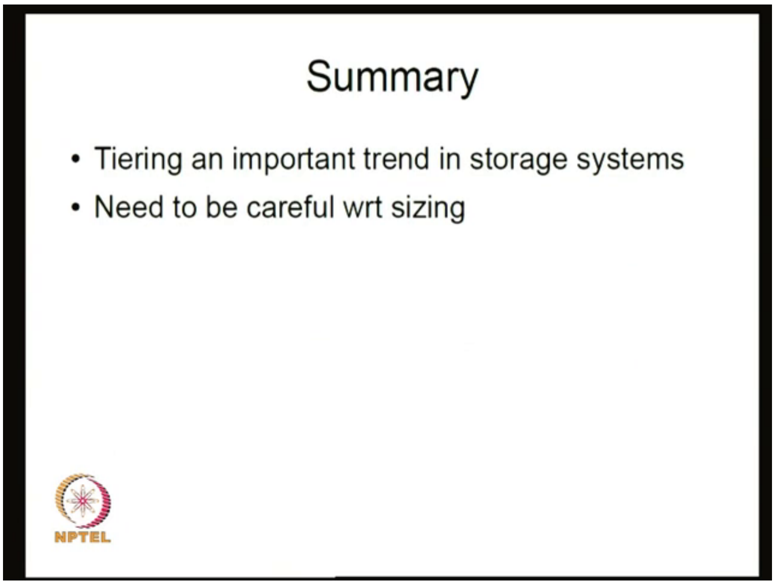
So, you add the air conditioning because with air conditioning also requires power, we have all those power air conditioning all this stuff it turns out to be that 10 megawatt it will be some multiples of megawatts; 20 megawatts sometime you need that kind of power to run this. Now one terabyte per system is very big point, but the thing is you can essentially get the same thing with SSDs with not too much size very costlier, but you will not have to invest in a power plant. For example, if one terabyte per second right how will we get this? Your bandwidth it is your SSD will be some kind of semiconductor memory, it will be approximately giving you about easily I have 5gigabyte per second kind of bandwidth second with multiple interfaces in other things ok.

So, and if you. So, the number of devices you need, if you calculate it will be about let us say we said one terabyte basically you need about 200 SSDs, I need to SSD will be about consuming power about you do not (Refer Time: 52:56) right now it will be in the tens of watts multiple tens of watts because (Refer Time: 52:58) not everything is being used you cannot you can only pump that product which is being used. So, you are talking about something which will be at the most kilovolt ampere not megawatts.

So, this something that people have to think about when you are designing very lascar systems. So, there is power savings and space savings and. So, for last systems it is almost inevitable nowadays so, but the important thing is to remember how to think about it how to clear it. So, I gave you some outlines of a method and this was first done by jean gray. Jean gray was you got a twinge our sometime back and she basically came with this what is called 5 minute rule. So, in those days it was when this power 400 seconds I told you right in those days and 1980s this is approximately its 6 and half minutes, but rule is called fine minute rule; that means, in those days the thumb rule was you keep some stuff if the interference in trevallies less than 5 minutes keep it in memory, it is more than that put it in disk ok.

So, this kind of five minute rules updated to the current system speed that is what you have to do. So, if you do that; then your system is reasonably well balanced and you will have a more cost effective system ok.

(Refer Slide Time: 54:37)



Summary

- Tiering an important trend in storage systems
- Need to be careful wrt sizing

So, basically in summary Tiering is an important trend in storage systems and it is going to become much much more important the future because you are going to have different types of storage available and you basically have to be careful with respect to the sizing of each of these components, then you will get essentially a well balanced system. And the reason why this is important is because you will find that different technologies progress in different at different rates. Semiconductor technology that is there in dram that is there in slash that will be there in what is future called phrase change memory or racetrack memory.

(Refer Time: 55:20).

They will be got particular rate whereas, magnetic medium will have its own rate.

(Refer Time: 55:25).

And now they are they are doing different things, if you look at the semi conductor semiconductor memory they are going towards access time optimization, and the magnetic memory they are going towards capacity. So, there will be finally, there will be going different ways.

So, finally, you have to put some intermediate technology, it will actually gap which will follow this is a gap. So, you will have intermediate technology many of them, these are all fill the gap so that there is better balance. With that I will conclude today’s talk and we will continue this class on some our some other interesting aspects of storage desire (Refer Time: 56:06).