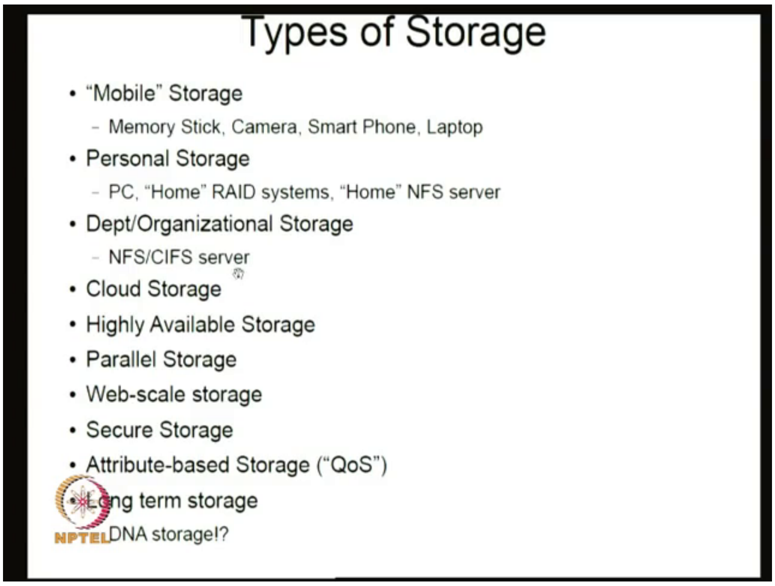
Storage Systems
Dr. K. Gopinath
Department of Computer Science and Engineering
Indian Institute of Science, Bangalore

Types of Storage Devices and Systems, Long-term Storage
Lecture - 14
Types of Storage _Part 2: Parallel Storage, Cloud Storage, Highly Available Storage, Web-scale Storage

Welcome again to the NPTEL course on storage systems.

(Refer Slide Time: 00:25)

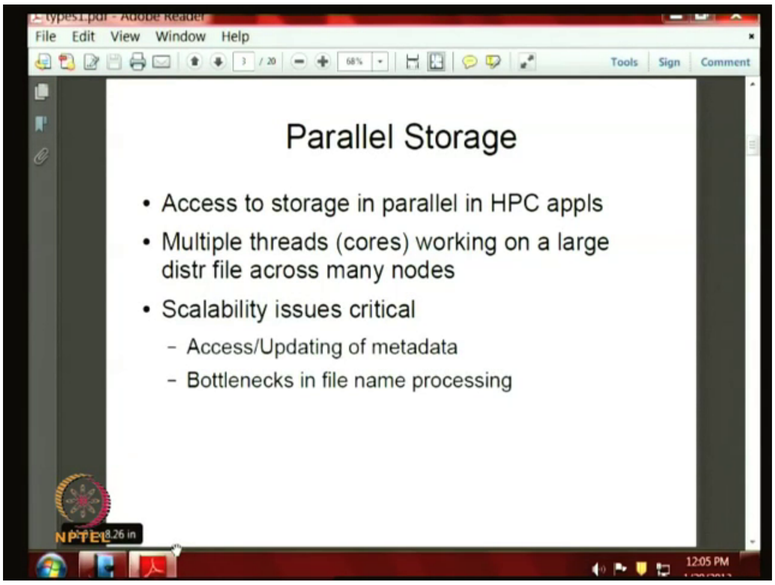


Types of Storage

- "Mobile" Storage
  - Memory Stick, Camera, Smart Phone, Laptop
- Personal Storage
  - PC, "Home" RAID systems, "Home" NFS server
- Dept/Organizational Storage
  - NFS/CIFS server
- Cloud Storage
- Highly Available Storage
- Parallel Storage
- Web-scale storage
- Secure Storage
- Attribute-based Storage ("QoS")
- Long term storage
  - DNA storage!?

Last class you are looking at different types of storage, and I was giving you some details about each type of storage, and we did not completely go through or types of storage, because it is quite vast. So, I think we did up till highly parallel, highly available storage. Now, we will have a few more types of storage, I want to just discuss. So, that we give an I, to get an idea about various types.

(Refer Slide Time: 00:56)



Parallel Storage

- Access to storage in parallel in HPC appls
- Multiple threads (cores) working on a large distr file across many nodes
- Scalability issues critical
  - Access/Updating of metadata
  - Bottlenecks in file name processing

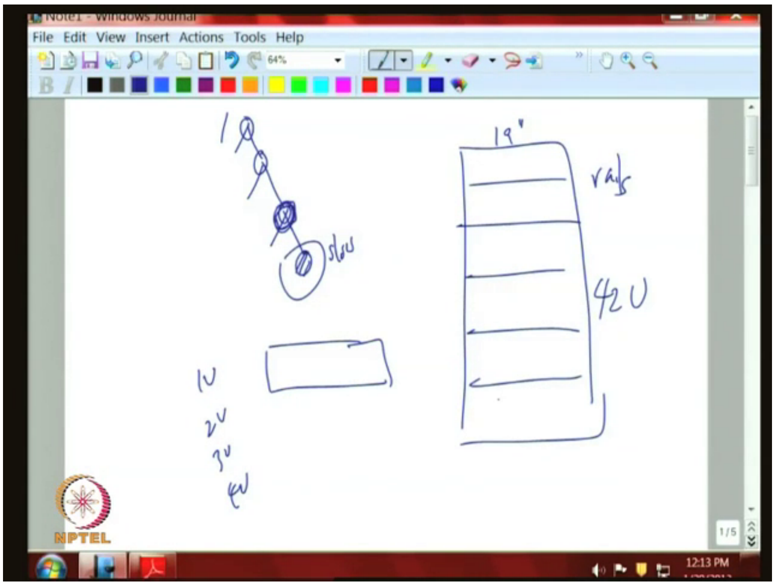
So, what is parallel storage? Access storage in parallel in High Performance Computing applications.

Usually what happens in high performance, computing is that we have lot of independent processors or course hardware threads whatever, but call them, and they are working on a large distributed file, across many nodes. Typically, the information is a single file or a single device and. So, multiple parties are accessing at the same time right; that means, that you need to do some locking or similar issues, some it is axis, is going on and. So, if you do not do this locking or provide simultaneous access, some or fast, then the solution not be scalable. Your high performance computing whenever it has sequential parts right by I am (Refer Time: 01:57) law it slows down right the only way you get 100 percent.

Scaling up is by making sure that there is very little of sequentiality in your code, whatever it is you have to, is it systematical eliminate sequentiality then only you can achieve the linear scale up. So, if you are trying to access storage in parallel, if there is any locking on those kind of issues, there usually the sequential part, because finally, there is somebody.

Everybody has to go to some typically central agent that guy decides, whether to allow you, are somebody else right of course, there are something called lock free solutions, but those things require some more careful thinking, but generally it terms out when you have simultaneous access you might get into scalability issues and also you have metadata, which is critical in a storage system, that access to it or updating to it also becomes an important issue, there could be one thing like bottlenecks in file name processing, what do you mean by that example, I might have a file by the name; slash a, slash b, slash c, when updating something in the lowest level. Let us look at the suppose I have a hierarchy, all right.

(Refer Slide Time: 03:33)



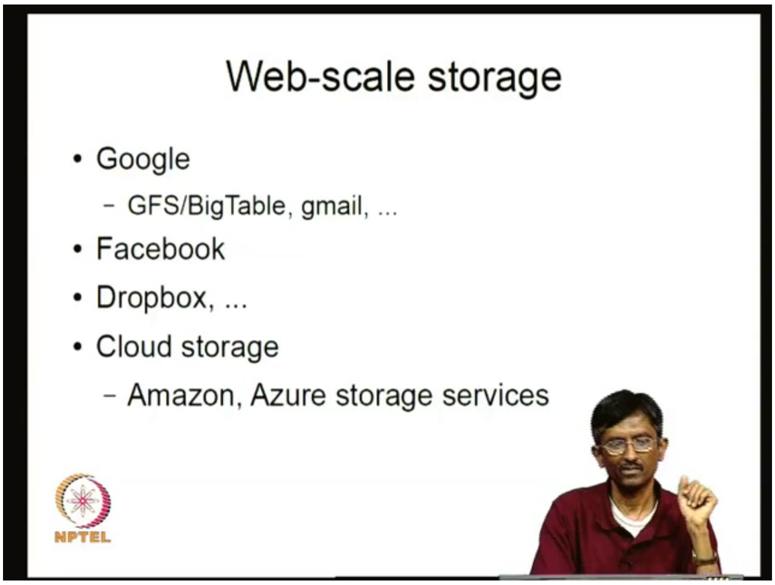
So, suppose I am updating this, let me, something has changed here. Let us say that, this is the directly, which holds information about this object something has changed; that means that, this has to be updated, but is, let us say; there is concurrent operation; that means, that this has to be locked, because this locked you will find that somebody who is traversing it like this he wants to add. Let us say something here to add, this then something changes here; that means, the this also has to be locked right. So, in the sense what happens is that somebody can hold on to some lock and because, this is held up some other concurrent operation, can keep this locked also. So, in a sense what happens is, if there are situations where, if this is very slow and there other concurrent accesses is which somehow lock this up and are waiting for this to finish lock this up and waiting for this to finish.

So, this guy I will be slash let us say. So, this guy is essentially you cannot do any you can any one get to the top or top part of the hierarchy and everybodys waiting for this guy to finish, you can construct examples of this, was another reason why, sometimes you might have seen whenever using your system. Sometime the system seems to be just stuck for arbitrary period of time and slowly starts moving right in what is a same it? It can be 1 second can be 10 seconds all right. I am not saying is, because I exactly, this some of it could be because of this basically, because there a slow operation and you need to update something there and there other concurrent operations which need to update this, but update this the other, each parent has to also to be held constant.

So, that this guys update can be reflected here. So, there is some kind of then which the locking precedes and it is all stuck with this particular we know. So, that. So, that is what of certain bottleneck. So, basically, if you look at parallel storage they are try to avoid this things, they try to ensure that you do not have to go through this kind of processing, there might be use some direct means of naming for example, I have a name slash A, slash B, slash C, I do not go through directory by directory. I just hash it immediately, look up that hash, you need to find some other different solutions for this.

So, that is. So, a lot of issues here, but I am just want to tell you that, there are different issues scalability becomes very critical you have to really of course, it is true for many other situations, but in high performance computing, you are the name of the game is scalability you have. So, many course on. So, much multiple disks should at corresponding through, put if you not getting it something is not right. So, scalability is an important issue.

(Refer Slide Time: 06:57)



Web-scale storage

- Google
  - GFS/BigTable, gmail, ...
- Facebook
- Dropbox, ...
- Cloud storage
  - Amazon, Azure storage services

You can also have a web scale storage all right. I think all of you are familiar with Google and for search applications; they have this Google file system. We talked a bit about it and these good for 64 megabyte kind of chunks, but that is not good enough for smaller things. So, they have some other, another additional storage called another layer of software called big table, which manages smalls pieces of storage and which is sitting on top of the Google file system of course, now all this stuff has been reworked in different new or types of storage systems, but this is the one that has been there for some time consumer also has is a type of storage, you will notice that email is extremely a metadata rich right. Why? Because, you want able to search based on whether who sent something to you, whom did you send it to; when it send it to, whether it have a attachments.

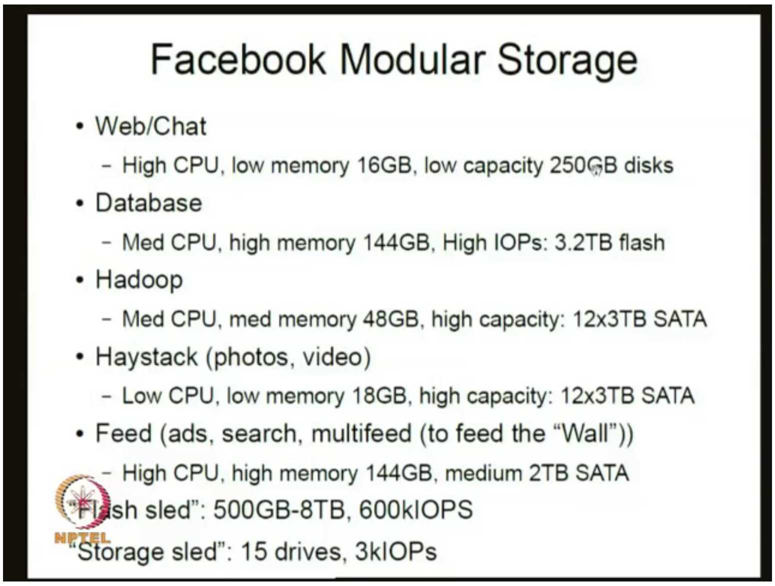
So, all this are metadata for the mail. So, oftentimes you want search based on metadata right. So, the thing is; that means, that mail is typically a highly metadata intensive metadata. Intensive means it is dealing with some small piece of information, we are not talking about gigabytes of data, you talking about small piece of data when was it last modified; that means, that your picking up small piece of data and wanting to look at large numbers of small piece of data large, numbers of small piece of data.

So, this is metadata intensive some mail is typically one of those types. So, you need to design your system. So, that metadata is handled well for example, a GFS is totally unsuited for it, absolutely unsuited for it right. Even bitable may not be suitable for it, you need to something better, probably bitable you have to do it, but I am just saying you have to think about it. Similarly, you have Facebook, it has it is own compulsions. I do not use Facebook, but I am told many of few might be using it.

So, there is something call the wall I am told right, where in you look up other people of posted for your friends something and you want display it; that means, you have to go and go to lots of people, pick out small piece of data and combining together and give it to you right; that means, you should be able to access this is data and also this could be geographically distributed, because you might have friends and locally or internationally also right, multi cities, you have to pull out data formed distributed geographically distributed data, but is one thing that could be so; that means, that your system has to be really geared for this kind of access.

I will just look at let us quickly look at this.

(Refer Slide Time: 09:56)



Facebook Modular Storage

- Web/Chat
  - High CPU, low memory 16GB, low capacity 250GB disks
- Database
  - Med CPU, high memory 144GB, High IOPs: 3.2TB flash
- Hadoop
  - Med CPU, med memory 48GB, high capacity: 12x3TB SATA
- Haystack (photos, video)
  - Low CPU, low memory 18GB, high capacity: 12x3TB SATA
- Feed (ads, search, multifeed (to feed the "Wall"))
  - High CPU, high memory 144GB, medium 2TB SATA

"Flash sled": 500GB-8TB, 600kIOPS

"Storage sled": 15 drives, 3kIOPs

So, the solution, the hand is what is called highly modular storage very highly modular storage, because it turns out this Facebook people to require many different access storage, not one time and not only minute different types of storage, but also different types of CPUs, which can be, which are necessary for doing a kind of processing that is a required for example, if you take web and chat right this sayings, because chatting for example, you are typing a few bytes per second is nothing much even, if you take one whole country is typing chat rate. Let us, a 1 million people are there in some small country, they can barely do 6 million characters per second or 10 million character per second yes.

So, this is not data intensive sources in. Why? But it is high CPU, because you are trying to collect information from one place, because it can multiple parties, can be a attach for same place, same chatter module right. So, you need to figure out where these things are and distribute various places, typically it has got high according to the analysis, these people have done this, requires high CPU, low memory, because I told you do not really type much low capacities you do not need much you can do it with low capacities.

So, what Facebook has got? They have a rack, rack is a set of, let us say if your familiar with the rack, holder is rack look like. Let us just you know I think you all are familiar with it, but let me just. So, basically this could be normally a rack is, what is called 42 u system. It has got various slots, it has some power connection, the back and you can there are somebody else and you can take servers or other specialized storage devices and network devices you can insert into it. This is a big things usually 19 inch. This is 19 inch and it is about each device itself, it is comes in terms of what is call 1 u, 2 u, 3 u or 4 u. I do not recollect, what 1 u means. Typically, these are about few inches. Let us assume that one inch is about 5 or 6 inches and I am just guess, I am do not remember the exact number.

So, these are basically 19 inch wide things about 5 to 6 or 7 inches in height and they go all the way back to the end of the rack. At the end of, at the backside of it and there are connections of the back and basically, what you do is what Facebook has got is, they have specialize racks for each type of application for web and chat. They will have one rack with certain types of CPUs, with certain types of storage, with certain types of network devices and they will replicate this on a massive scale, they might have at least data center, some 100s of these things and they usually the good thing about this, is that it has got integrated power management and network management all those kind of things. So, that and failure management for example, all these things are integrated here. So, all the servers, all the network devices, all the storage devices, they all have a unified model of managing this things.

So, they have some type of racks, of this type; that means, that they would have probably CPU and memory together and then this thing separate racks means separate 4 2 u or 4 u kind of. Let us say capacity; they will have it as some we, a some of the units in the rack. So, web and chat you also in a database, because your authenticating people, you need to keep certain information, critical information, which you go through a data base. Why? Do you know database? Because this has to be consistent across multiple areas, you could basically, just the same reason, why you need a data base for storing money basically, you want to be due, some guarantees with respect to accuracy and even affect the critical thing about a money kind of related matters is that, whatever it is, you should make a mistake about a amount whatever you to do right.

So, you do not mind telling the people that, I cannot give you information right. Now, come back tomorrow, but I will not give you wrong information right, that is critical. Similarly, in database also you do not want. So, in this Facebook also, when you want logging in people, logging out people deleting accounts, (Refer Time:14:51) whatever or privacy management, whatever you are doing right, you want to give some guarantees, you done it once, you should be that way.

So, data base you can, have got certain more stricter semantics. It does not scale, because in that, it really cannot scale to the extent file systems or others things can and can do it. For example, Google file system is a file system, it is not database, if you what, to raise scale tube, the web scale database will not useless, will be useful, will not be useful. So, but you need it first and a specific activities, and the thing about, because of database, being very intensive with respect to resources, that is why you need high memory and you need higher high IOPs, because they going to be making lots of small accesses. They are, updating your account Facebook whatever information a keeping (Refer Time: 15:47) it will not be too much, too many bytes at the most; it will be 4 kilobytes or 8 kilobytes right. Where are you from? What is the last time you will logged in? All these kind of stuff, you barely will fill it in 8 4 kilobyte thing.

So, say. So, many people are using at a same time right. You need to have high IOPs; that means, that what they will have is they will have a rack, which has got a CP, which has got a CPU. Let us say devices, they are not going to be that very high speed, listen, why you need high speed CPU, in the case of web is, because you might be doing JavaScript you are executing some code.

I think some of you, If you are looked at your Gmail, the very, the first time you log into Gmail, it actually picking about huge JavaScript code, and then it is going to executed for you; that is why becomes slow. I think some of you might have seen that it says, do you want fast access or a slow access? If you want, if the network is slow, we can go for the slow access; that means; it puts in a slightly lesser amount of JavaScript and lesser number functionality. Reason why you need high CPU? Because of that lot of interpreted languages are being used and you want execute them very fast.

So, that you get reasonable access that is high CPU whereas, database you will not really possibly doing too much CPU activity, you are not interpreting, typical those kind of JavaScript from that much, there will be some, there are some triggers, etcetera and databases that also might require CPU activity, but according to their experience, you do not need very high CPU capacity, for this kind of access they need high memory and high IOPs. I told you why? Unit high IOPs, why they need high memory? I am not very clear; it could be that they want to keep the metadata. Most of the metadata in memory, I am not very clear, why that is? It is, if you do the processing for example, Hadoop, what is Hadoop kind of activity?

It is about trying to do some batch processing on the data, you need to have, for example, they want to figure out what is the trends? Is there some unusual activity going in a system, or they want to track your trends. So, that you can give you ads, makes it whatever it is. So, this is basically, you might call it usually offline activity. That is, this is not really a real time, there is no user in front a, of it. Demanding quick or instantaneous response or quick responds. This things are basically, analytics. You might call it, those things have done in the background to keep the system in good condition.

So, here in terms out, you are really looking at lots of data right, for example, you might, they might want analyze your behavior or they want to say that we know that lot more activity in India is about certain things, computer something else. We do not know, what is, what that is; that means, that going to go through all the data. So, that is why you need high capacity disk and since your actually streaming through lot of data, you not really doing serious analysis.

So, you just adding things up and things over time, not very serious CPU, that is why you need medium CPU and they find, I cannot, I will explain this things very well, why? High memory for this medium memory, but this is as their experience, I am just reporting what they? What is it? Similarly, with photos and video, you may high capacity that is why they will have in the rack. I say told you there are many racks, each rack will have multiple units, which actually have many of this high capacity disks, like troll up, then in each for all these units.

And again in the case of this photos and video, it is mostly access, an access and retrial, because whatever thing, that is happening is happening on the client side right. You are rendering the image displaying the video, that is happening on the client side not on the face book side, that is all they have to do is to just give you the updates, at the most network is stressed, but not CPU, also that is why these are low CPUs.

Now, if you take about other kinds of accesses feed for example, you want to serve ads to you, you want do some search or you want to feed the wall right. This is high CPU, because you actually have to get multiple things and somehow put them together, you have to create other paging, you have to really take multiple information and give you of web, then present your webpage, which is what to you what to finally, see because a web browser has to be given a page which has to be displayed.

So, as to construct on the fly a web page for you right, getting information various sources that is why requires high CPU and search, because these are all very targeted search ads. All these things are very, you have to do some who knows ads might require some auctioning, some of you who study these things right. You know that they auction the spaces, what has to be presented, somebody is doing who knows some kind of ad auction also going on, that is why you need high CPU for same reason it high memory also, but you are not really doing as much data access as in these cases, because a number of friends, you have a sort of limited hopefully right, it is not some millions of them right. So, that is why you do not really need to look at two minimum it is, there is some usual typical bomb, thus I can all these experience, a face book people.

So, we have really talk to them, why they are saying what they, but I am just guessing, but why there is why they we things are. So, in addition of these, they have what is called a flash sled and storage sled? A sled is basically, that twilit I am talking about that 5 inches, 6 inch, kind of thing. So, many of these things are populate this rack right. So, it terms out for high IOPs right, they will have a flash sled, which is composed of flash only. So, for example, your database right, you need high IOPs, they will have 3.2 terabytes of flash and; that means, is this composed with lot of a flash sleds for example, they need 3. 2 Terabytes and they might need a certain amount of IOPs.

So, in each data center, they might have multiples of these things, because they talk about 600 kilo IOPs probably the need about I do not, from at the most 10 million IOPs as suppose, I do not know I am just guessing, but it could be the reason of 1 million IOPs, that is what they needed. So, they can attach about multiple flash sleds and that will call you taker of it, you also for is storage sled, where you need, you not really looking for high IOPs, but you are looking for capacity for example, then you will go first storage sleds, that is what for example, this kind of things.

So, basically they have constructed a system of interoperable elements basically, I can mix and match I am looking for web, I need high CPU. I have a certain type of design, which has high CPU processing power and then have a storage sled, with this kind of stuff are putting those things. So, in one rack I will populate this with 10 of this guys or 5 of this guys and 5 of this guys or something about guy, I need something which requires high IOPs, I take a medium CPU kind of design I have and then, I in the rack I will put many of these flash sleds. So, is what I do. So, basically, I have by having separating out CPU storage and also memory, they can mix and match in the racks, what you want.

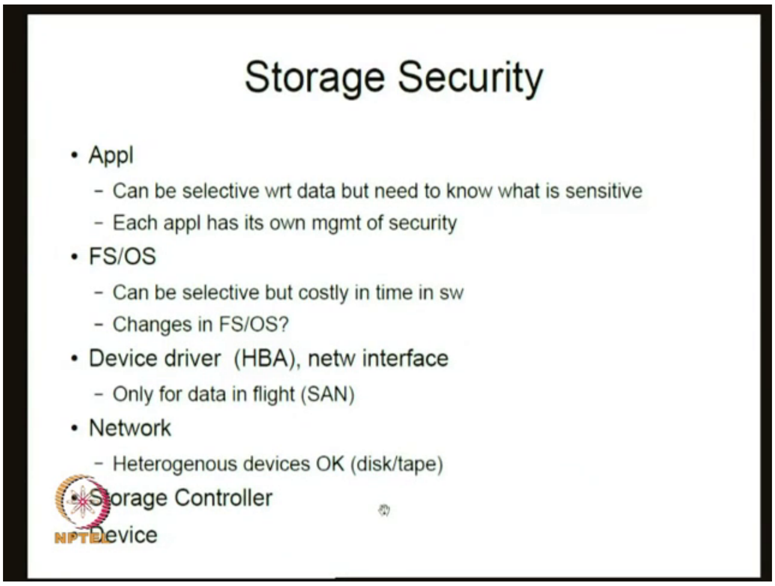
That is it kind of (Refer Time: 24:03) have I just brevier talked about Facebook, there is a similar thing for drop box. In drop box basically, we got survive to share information through the cloud. So, these people also will have to service provide lot of storage and it is not clear. How they, what they are doing, but they must be using possibly similar systems. There is also cloud storage company like Amazon and Microsoft, they sell storage services and they also have to give you second I am also storage, they just tell you that you wanted 1 terabyte or 10 terabytes, they will give it, you on the cloud and they will manage it for you, they give you an interface by which you can talk to it. It will not discussing, you have normally when we talk to storage on our laptops, PCs. It is a protocol, it is hide what is called the

ID protocol, data protocol, SATA protocol or SCSI protocol or any of those kind of things here they have not decided to support, then because when you are at a distance far away to network right. It terms out those kind of protocols will not work very well. So, they provide some other type, a protocol you have figure out that protocol. In Amazon it is called let us 3 and I will briefly talk about it, sometime later and Microsoft also has something similar, then user does not worry about, how this are done. It is the responsibility of the cloud provider, the cloud storage provider to ensure that the data are that, the storage at to promise, the use of right stays intact, they have to take care of things like application.

So, that whatever happens that data is stay there right. So, typically they agree only guarantee search for trialability access. They do not give any guarantees with respect to things like latency etcetera, that is there on. So, this is becoming more or more important and you will see lot of development in the future. So, that finally, most likely, your laptops will be without storage, most likely I think you already see it happening. If you have used an apples air or whatever right, most of the devices, storage devices to be take out, only you they a similar to you, have a network CD ROMs have been taken out DVDs out all are taken out they might give you a small SSD storage, but may not be fit, all the things what you want and apple himself, apple itself, provides you cloud storage, but; that means, I did not have high speed storage to network. So, for people without high speed storage it is not a good solution, but it is the trend. So, we will start seeing quite a bit of that is in future.

So, web scale storage, this let us look at secure storage, that is also of, this also is an important issue.

(Refer Slide Time: 27:03)



Storage Security

- Appl
  - Can be selective wrt data but need to know what is sensitive
  - Each appl has its own mgmt of security
- FS/OS
  - Can be selective but costly in time in sw
  - Changes in FS/OS?
- Device driver (HBA), netw interface
  - Only for data in flight (SAN)
- Network
  - Heterogenous devices OK (disk/tape)
- Storage Controller
- Device

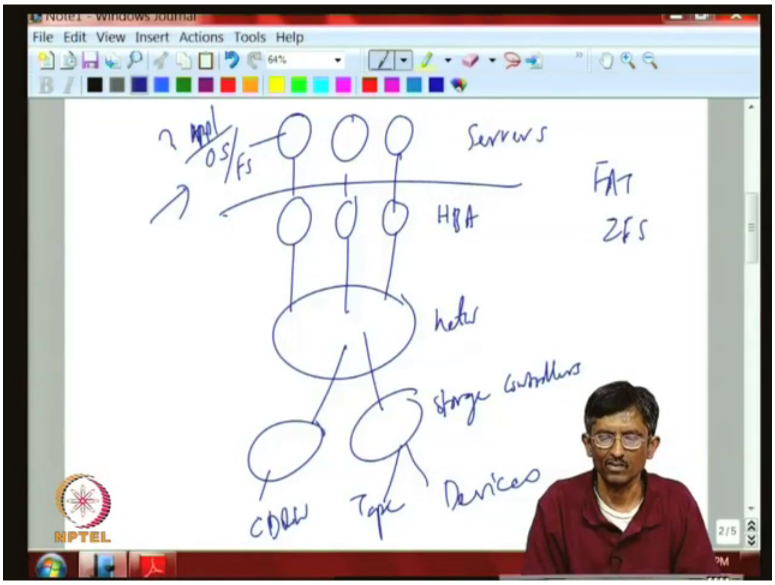
Now, storage security is a very complex subject, because I can do it at multiple levels, I can decide that the application should decide about the security, not anybody else, because then the applications complete control or it can be the file system can be interested to this or the operation system can be interested to it or there are intermediate devices, that come in the picture will talk about it. Now, already talked about little bit, but I will again go through it, it can be through some other intermediary devices like HBA or the network interface that can also protect these things, but usually only for there data, that is being transmitted or it could be for at the device level, where it is actually data at rest; that means, nothing is happening you.

Still want to protect it. So, there are some things at the level of data in flight, that is as the data is moving from one place, another place, I want to protect, it is contents, there is some, which is at rest for example, I have a memory stick, I am not accessing it, I just keep it a side, I want to ensure that particular piece of data stays there for the next 5 years, without anybody else, other than my be able to read it that is data interest. I am not transmitting them, I am just keeping does.

So, anybody else also has got it. It tries to you, use it, you cannot use it. It is not able to use it. So, you can do it an typically is an data in flight. There is also other kinds, a solutions, when do it in the network itself and then you can do it in the storage controller and device. So, I will go through this a bit. So, it is a, that we are clear about what this is.

So, as I mentioned before.

(Refer Slide Time: 29:02)



We have servers you have that agents before a red agents. Let us say that there is the file system, also on it the software can be operating systems or file systems, those kind of things it as got agents. Let us call it HBs right. Host bus adapter, I think we talked about a little bit, all right and then they could be attached through some network and then they can be attached through storage controllers and finally, devices.

So, there is application here on top again application OS FS etcetera, right so many things of there. So, what is it mean? To say that application handles it, application says that I do not trust any of these guys, I do not trust HPA networks storage controllers, etcetera. What is it say? It says I will do the key management, I will do the encryption and decryption myself; I will only ask you to storage my interfaces store. This bits, I will take this bits, it is not business to interpredicate, it is not business, take care of key management. These are example of I am responsible for at, I have for security become I do not have symbols of course; that means, that he has saw.

He has to trust the server code that is all his own code and that of the system that, it they are not coming in early that provided or it could be that the OS is told do it; that means, that for example, it may be that the pagings of system itself does the encryption and decryption or you might leave it to the file system, where it, when it tells the device driver, the HBA this side, which was, I will do the encryption decryption, it giving to you, now store it that is it.

Now, you can save different kinds of issues which each of these things I will go through each of them, but this they have to keep. So, what the advantage of going with application thing, you are completely in control, but all the headache is also with you with OS; that means, that it is now across all applications, there is one person and more typically application is likely to make a mess of things, because return by most of people like us whereas, OS is typical return by slightly more experience people.

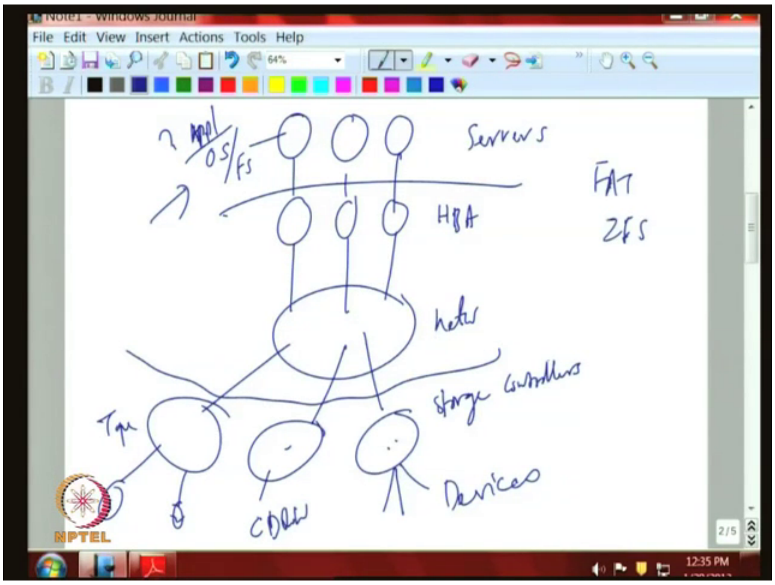
So, probably you might get it wrong, who knows OS probably, you can get it right most likely, because more experienced people again, but; that means, you need to get them to do it, but depend on them it may be there, a different versions of operating systems. You have to ensure that some of they are doing at a right thing across all these versions. Suppose, somebody says windows 98 and somebody says windows vista, who knows whatever that is, there in windows 98. It is not there in windows vista, I have to go and beg the guy. Please ensure that your prepared a solution, whatever solution have works from here also.

So, you are under the control of that guy, same thing with file systems, the good thing of file system, now is that it may be across multiple operating systems; example, fat file system right. Now, linux supports it, windows supports it, apple supports it, some other put encryption decryption whatever there a need. It is now neutral to which operating system, it is or EXT file system or if you take ZFS, which is some Microsystems developed it, but is available free BSD, also if somebody puts in encryption decryption. There it is available across multiple parties, know multiple operating systems, you know. So, that is the kind of stuff there are plus or minus things, same thing about HBS, if you are doing it across HBS.

Now, you do not have to worry about you can mix and match all these things devices for example, they can be tape device, here they could be CD ROMs, here CD not ROMs CD r write right become writing and reading. Now, this is guy is across any of this things, you can do it, because there are some other intermediary guys, who are taking care of how to handle, the differences in the system whether it is a tape or disk devices, etcetera.

So, same thing about for example, storage controllers, I made mistake here a (Refer Time:34:04) put it like this typical is storage controllers, they it like to talk to homogeneous sets of devices, the storage control will talk only to devices in my point um.

(Refer Slide Time: 34:25)

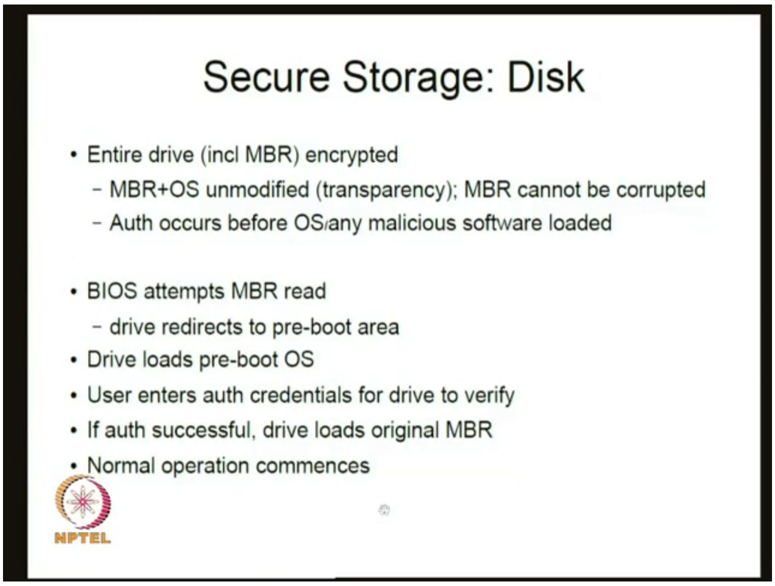


All the devices will be the same, it will be disks the regular storage controller, for tape, there will one tape, 2 etcetera; that means, that if you do any of these things right. From this side you do not care what type of devices. Here, you can actually not, do not have to worry about whether it is a tape here or whether it is a device, where is a start doing at a storage control in side

same as here, then what happens is that you have to or struck with did in only with that type of device; that means, you have to ensure that your solution right has to repeated both here and here. So, you should think careful about it, different solutions we have different requirements with respect to who does the managing, the security aspect and how much of it is good with respect to the heterogeneity. I think you already discuss about applications, whereas suppose right similarly, on this set also will have the institutions. So, yeah.

Now, we will just typically look at one example.

(Refer Slide Time: 36:22)



Secure Storage: Disk

- Entire drive (incl MBR) encrypted
  - MBR+OS unmodified (transparency); MBR cannot be corrupted
  - Auth occurs before OS/any malicious software loaded
- BIOS attempts MBR read
  - drive redirects to pre-boot area
- Drive loads pre-boot OS
- User enters auth credentials for drive to verify
- If auth successful, drive loads original MBR
- Normal operation commences

A disk suppose I am saying that I am interested in this model only device, I want everything to be in the device, why is it good? I have a disk, I will put some sensitive information, I update a new device, new machine, the information is still there in that old device. I might not think in price an I might give it to my brother sister are some friend. Let us see that I might not really think too much. We can, I might think is some file system is, there are they will take care of it, but who knows the part? I gave it to has a slightly devious bent of mind you wants to figure out what we as told. He can look into it, only solution is to completely put zeros or something out, it is not scares the whole disk, but people are discovered that even if you write zeros unless, you do it about some 32 on using all kinds of interesting patterns on it, there is a residue of what was there in the past.

So, with clever signal passing techniques, you can figure out what informations there. So; that means, that I am interested a solution which somehow I do nOt have a go through serious administrative procedures also the problem with newer disks, big disks, new are disks. Suppose, have a 4 terabyte disk a somebody wants write zeros on it you can this think about how long it will take right what is speed at which you can write a disk assume it is something like 750 megabytes, megabits per second like let us, we generals. Let us call it 1 gigabit per second; that means, that 2 right 1 terabyte, I will take 1000 seconds and highest speed possible, you know that you can get really 1 gigabyte, because the disk has got different densities right, different parts of a disk.

So, you can go all the wave form, pause all the about half then half or one third that capa is speed. So, we talking about even a first, say 1 gigabit per second. You are talking about 1000 seconds and here 4 terabyte, we have 4000 second that is have a 1 hour, this is the highest possible speed actually does not take 4 hours. It takes, sorry it does not take 1 hour it takes closed about 20 hours, is something like people at who do it and practice, define that it takes close to 15 to 17 hours.

So, if our a large organization and you want to decommission some disks and put a new one somebody access it and do all this things systemically, otherwise you cannot give it out anybody else, there is some ad missed over heads. So, some people of figured out, that you need, you can do slides something different and this is what is called on disk encryption, you do that disk in decrement to inside a description. See get us some solution of this kind, there are some standards know.

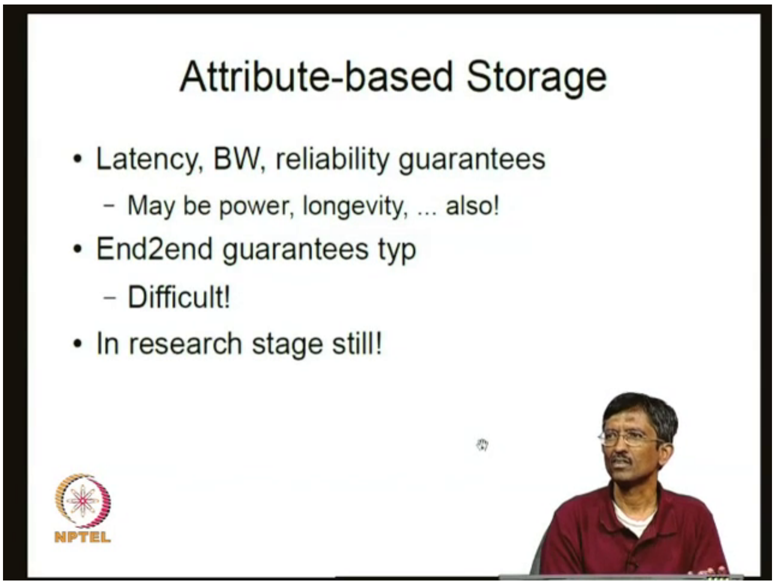
So, here the entire drive is encrypted, but what is interesting about this is that the MBR and OS unmodified, it is complete transparent and MBR also cannot be corrupted, because there is a shadow MBR, that is kept and authentication occurs before OS, any malicious software loaded maybe with just have it, happens first, what is the situation. Let us assume, you are a PC like architecture of course, this a problem with this, you have to think about, because you are talking about BIOs here; that means, it is not useful for apple kind of people, they you something else; that means, depending on this BIOs right different situations here. So, the BIOs is reading, you get us, say you now how do thus, storage system work what is a disk actually, is a very I as a mention, early it is an intelligent device, you are asking it at logical level. You are asking it, give me sector 0 or 1 depending are where are the MBRs that activentity on the disk right.

You can look at what you are asking for it no. So, you asking for MBR. I will not give you, the actually MBRs sitting, there I will fake it. Now, it you try reading it and because is in act discuss actual be, we has a processing in a, in it right as I mentioned one. Some time back a discuss got as much processing power as a finite megahertz processor are one give hertz processor, nowadays HM.

So, it is an activentity. So, you give it something, it can look at in and decide. This is your asking for MBR instead of giving you the MBR that we are expecting, it gives you some other piece of information, basically pre boot area, you might call it and this is what is actually sodden on to the system and that is the thing, which actually does the authentication, just like your authentication for BIOs right. You have seen the password mechanism there. Some is similar happens now, authentic. It is successful drive loads original MBR normal operation commences of course; this requires that BIOs or the boots that are there on the system, cannot be the security of that, system is assumed that part of it.

If hear in the, you put in some fake cards, etcetera, which can take control of this BIOs for example, then your security is gone. So, there is some the caller trusted computing model by which you can give guarantees with respect to what is there on the motherboards are. So, you are sure that you should doing is authentication here, you are actually talking to the legitimate BIOs there is some issues here, but as I will all this things of straightly deleted, you have to really get it right otherwise, nothing works there is always over to break the system. If you are careless, we looked at parallel storage, web scale storage, secure storage, it is also attribute based storage, this is one.

(Refer Slide Time: 42:43)



Attribute-based Storage

- Latency, BW, reliability guarantees
  - May be power, longevity, ... also!
- End2end guarantees typ
  - Difficult!
- In research stage still!

Which has not really taken off, it is a bit difficult, do this like real time systems. What is issue, but real time systems you guaranteeing that some operation has a certain clear bound on how long it takes, will guarantee how long it takes in some sense. It can take 7 seconds, but you have to guarantee exactly take 6 on 7, or it can take 300 milliseconds. You want to say that is has 300 milli seconds, it was giving very good guarantees about direct, takes in 300 mill seconds or 7 seconds or whatever. Now it terms out to be able to do such things, you need to do in the whole system, can actually have impact. I am what. You doing right. You have to analyze the whole system, ensure that the bound the 7 seconds are 300 milliseconds the bound is given right

So, in storage also same story I am use, you seen already that when you are accessing things, you can have the file system, the picture, the words in the picture, the application, the picture, the HBA in the picture network, in the picture control of storage control of picture in the disk, in the picture, all in the same work. If you want to do anything with guarantee gives latency are bandwidth. You have to control all this guys, just like in a real time system also, if you want to guarantee the latency you really have to take tight control about all the areas that say invite difficult to guarantee these things.

That is why this still not widely offered or rarely offered or not offered similar reliability. Can I say that this will last for 3 years or instead of that kind of reliability in terms of time, can I say that for example, I give you that a chance of losing data is 10 to the power of bit error rate. For example, 10 to the power of minus 20, can I give it. I have disc width trend to the power minus 15, but I do some other coding kind of things. I give it 10 to the power of minus 20 bit error it right.

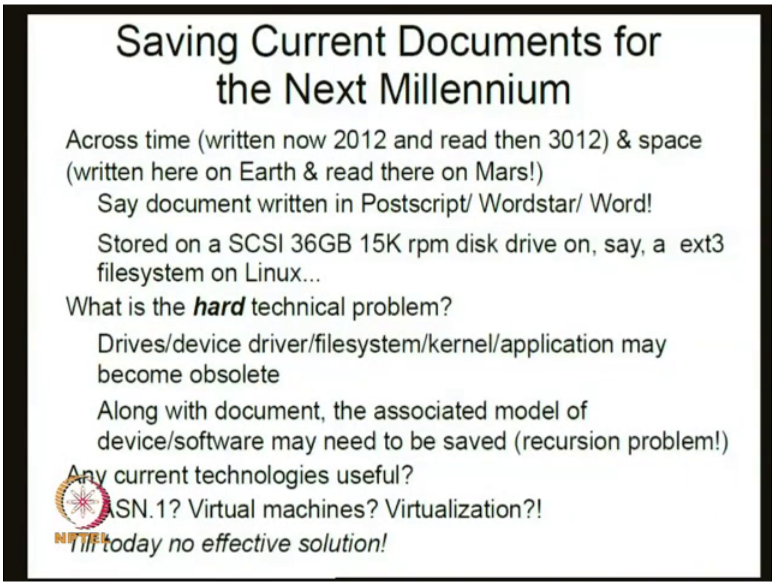
So, those can reliability guarantees. I think that is be comings slightly more easier. This part of it is being people have starting to give. Other kind or reliability is availability; I mentioned reliability, this also availability. I go to Gmail, I expect to the available Gmail, because Gmail is using some storage, Gmail will not work with the storage is not available.

So, it, Gmail wants to give you some guarantees saying that 1 second in one layer is all that my system can be down, I cannot see my Gmail for at the most one second, all most any other second in the year. I should, will be able to see it. Suppose I want to do that so that I think the commercial world understands. So, there is lot of effort at this point, and there are people who give you this kind of guarantees. You got any vendor, they will tell you my availabilities 99.99; that means, if you take one year as unit only 1 by what is the 10 to the power of minus 3 by 10 square 10 to the power of minus 5 into 1 year that is all the seconds, that is all the time I can break down.

Similarly, in the case of power and longevity, power means what say that, I can give, I can give you guarantees that it will not take more than for your access is, it will not burn more than this amount of power per second. This amount of energy it uses per second longevity. I can guarantee that is available 10 years from now. Whatever data you have put will be a value 10 years. Now this is becomes important for cloud storage guys, and I am putting something in a drop box. 10 years later I want to access it. How does it provide or Gmail for example, has been there for the at least 10 years now right.

As far as you know they still keep your made, no very beginning, unless you have your exciter a quota on. So, it is there headache, if they were to going from Linux to some other system in the future. Whatever they are doing, they have to take my data, and give it a new form keep on putting the new data format. So, whatever. So, end to end guarantees typically difficult that is why it is an rich stage.

(Refer Slide Time: 47:11)



Saving Current Documents for the Next Millennium

Across time (written now 2012 and read then 3012) & space (written here on Earth & read there on Mars!)

- Say document written in Postscript/ Wordstar/ Word!
- Stored on a SCSI 36GB 15K rpm disk drive on, say, a ext3 filesystem on Linux...

What is the ***hard*** technical problem?

- Drives/device driver/filesystem/kernel/application may become obsolete
- Along with document, the associated model of device/software may need to be saved (recursion problem!)

Any current technologies useful?

- ASN.1? Virtual machines? Virtualization?!

Till today no effective solution!

Now, let us look at this same longevity we talked about. Let us look at this.

So, one simple thing is, let us try to understand the problem saving current documents for the next million of. Let us say I am just artificially, let us think about this problem I want to store it, some data I want to access it thousand years from now. Of course, you will not be here, you will not be there at that time (Refer Time: 47:35), but let us say I want to do. You done some amazing research, you want people in. There is a 3000 equal to, still know about your research that is it right. So, your thesis (Refer Time: 47:48) you want people access in (Refer Time: 47:50).

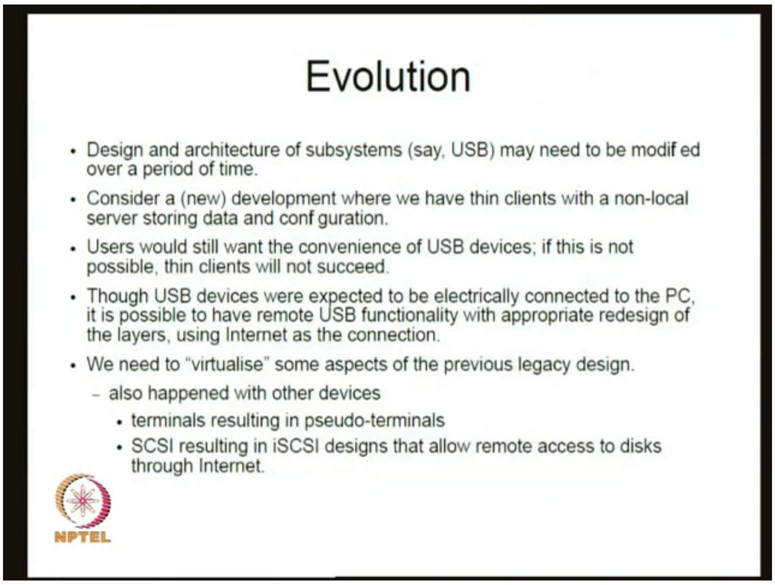
So, now our problem is that, we writing in some particular format, it could be open office, it could be, but how do you know that open office will be there, thousand is from now. You might think I will put on my memory stick, but you will notice that 5 years ago is to use a 128 megabyte memory stick. I do not know where it is sittings, somewhere I do not know where it is. So, the fact that I return onto it, this does not mean its accessible, but it could be on a particulars file system. The file system also could have changed, once I call, I will use it to, is it or fact 16. Now this fact 32 etcetera right, the file system can you change.

So basically all the things can change drives, device, driver, file system, kernel application, everything can change. Why is the application important? Because your postscript word stars word excla are basically user application. They have put information, particular wave in the file, and you know what that is ok.

So, in a sense along with a document, there is some metadata, that also have to be stored that is the saved recursion problem use. You wanted the data to be alive thousand years from now, but if to make it sensible, you have to store the metadata also, but that is same. Problem is stored in data itself. If I am able to store the metadata thousand year from now, I can also store the data, what is the big problem right. It is the same problem.

So, I am not solving the problem right. So, various methods have been attempted till today. There is no effective solution. There is no solution what say from solutions, coming in, the coming out, the picture you will just quickly look at it. So, I thought it may interesting to look at older solutions, will see how it works.

(Refer Slide Time: 49:35)



Evolution

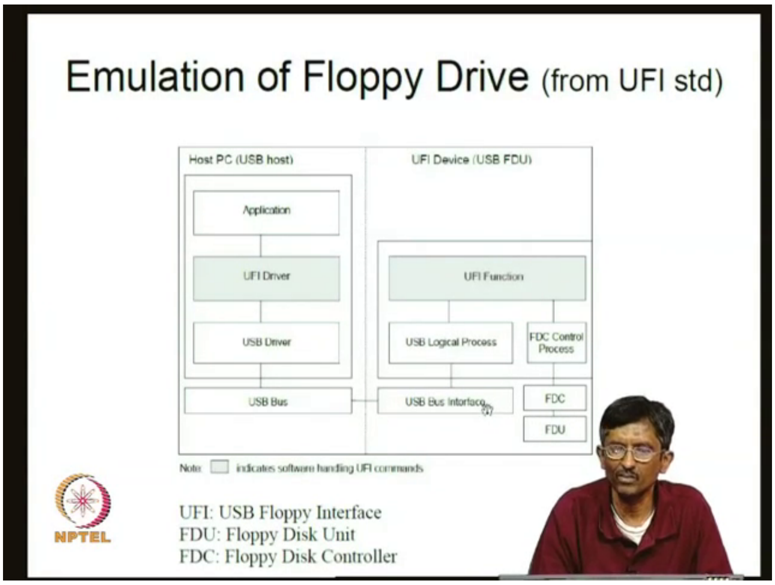
- Design and architecture of subsystems (say, USB) may need to be modified over a period of time.
- Consider a (new) development where we have thin clients with a non-local server storing data and configuration.
- Users would still want the convenience of USB devices; if this is not possible, thin clients will not succeed.
- Though USB devices were expected to be electrically connected to the PC, it is possible to have remote USB functionality with appropriate redesign of the layers, using Internet as the connection.
- We need to "virtualise" some aspects of the previous legacy design.
  - also happened with other devices
    - terminals resulting in pseudo-terminals
    - SCSI resulting in iSCSI designs that allow remote access to disks through Internet.

One solution is two emulation. Let us just take on storage systems theory again. Suppose you have USB. USB is you expected to put it into a somewhat, is this called some expected lots, some kind right. You put in it, it works electrical connections made, but now suppose I have thin clients, what is in client. I just have a display, I have probably some multimedia kind of things here in this.

. So, I can listen to music etcetera, but all the processing is happening when a server side. Now I am expecting that when I put my memory stick, the server is able to see it all right. In server, is not able to see it; that means, I need to have a another system on my side hear, able on the clients hear able, it does not make in sense. So, basically I want, if I put a stick here magically through network, you should be accessible or the server signet, it is able to mount it, and able to see the files, what is there here right. So, I want a convenience USB devices, but I want it through thin clients also. What is interesting, is USB devices. As far as I know they always been assumed to be a electrical connected, but I want now USB remote functionality, how were do it. I need do some kind of emulation or somewhere called virtualization of some kind I want to Virtualize a device so that I take this device and across a network, it looks to the server right, as if somebody had put this particular USB memory stick into its server memory stick slots. Somebody has to fake it, but its happen through network, and this is the quite a common thing. Let us happen, I think you must notice that something call terminals are there on Unix once. Once upon a time they were electrically connected, once upon a time, and later people are accessing it through network, but they do not want to change the application call terminal. So, now, we have something call pseudo terminals, which fake a terminal, completely fake the terminal.

So, that even though your connected, the network from the point of view of. Let us say you know typing back space all those things right. I do back space on this one, it looks to the CPU server as CPU doing a backspace on a terminal, which an attach to it electrically to that system. There it should give you exact same semantics. Something similar also has happened in SCSI, SCSI you was basically was not across internet, but now it is available on internet also. So, now, there is some faking going on.

(Refer Slide Time: 52:23)



Emulation of Floppy Drive (from UFI std)

UFI: USB Floppy Interface
FDU: Floppy Disk Unit
FDC: Floppy Disk Controller

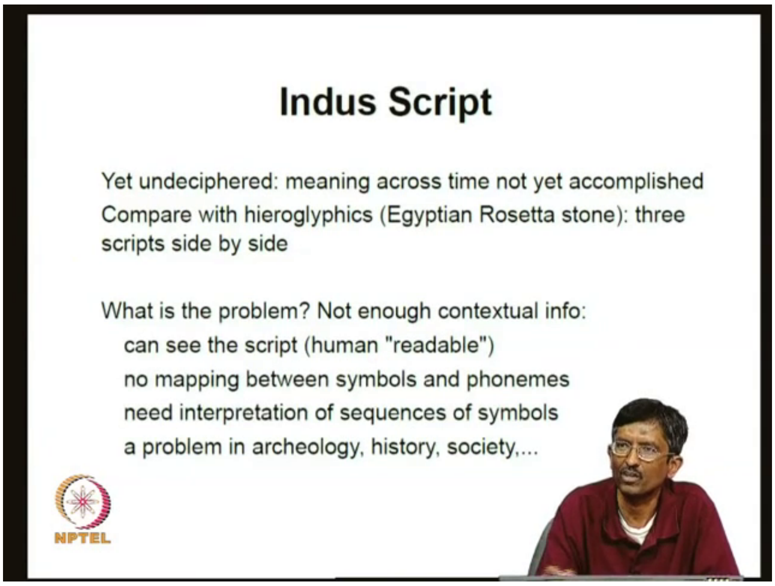
And you will see similar things happening in the case of. Once upon a time floppies were useful. Now suddenly some people have eliminated floppy drives right, this is not available, but then I want to have USB floppy; that is I have a USB slot, but I do not have a floppy drive attach to the system.

So, I will get a gadget call USB floppy which has only USB way to connect to the system, but I can put floppy into its. Let us say it is a floppy slot. Now what has to be done for that. This is an example of some standard of, some kind that occurred in the 98 99 2000. Even floppies are still being used. You will see that, what you are to do is, your host PC is running of application.

There is some specific piece of software then, has to be added and that actually essentially makes it. This guy makes it look as though. Sorry it makes, it look as though to the application if floppy is sitting there. So, that translations happening here, and sorry, and then from this part, this part actually, you will actually USB driver, USB bus, and then that is going through the network. It is going through the network across, and then it basically. Again you have to go through some additional piece of it, let us say software is actually does the complete faking its also.

So, these only across floppy etcetera, but if doing in a network, there will be network also. Here I am not assuming network. Here in this we are directory, assuming that we are attaching with it. So, there is, this is quite common. So, these businesses of devices keep probably.

(Refer Slide Time: 54:14)



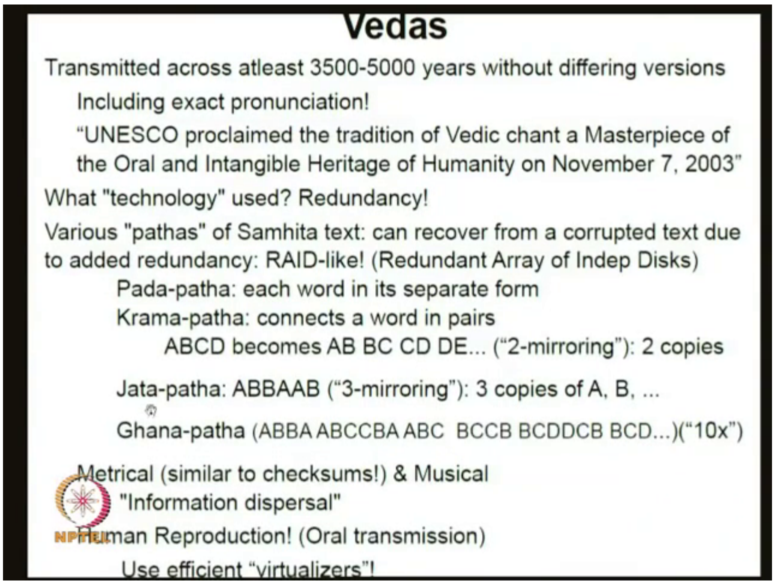
Indus Script

Yet undeciphered: meaning across time not yet accomplished
Compare with hieroglyphics (Egyptian Rosetta stone): three scripts side by side

What is the problem? Not enough contextual info:
- can see the script (human "readable")
- no mapping between symbols and phonemes
- need interpretation of sequences of symbols
- a problem in archeology, history, society,...

So, let me end with slightly more usual thing, might never come across. If you look at in the script right, nobody has disappeared (Refer Time: 54:20), nobody has done it. If you see the hieroglyphics, the reason why they were disappeared was, because they were able to find one, which had three scripts side by side. The hieroglyphics some the called demotic Greek, and something else, they are all in the same. Let us say the same stone, Rosetta stone. So, therefore, the question is, what is happening, but nobody has done it for in the script. Nobody knows what it, what does symbols means. There are some inspired guesses, but nobody knows.

(Refer Slide Time: 54:56)



Vedas

Transmitted across atleast 3500-5000 years without differing versions

- Including exact pronunciation!
- "UNESCO proclaimed the tradition of Vedic chant a Masterpiece of the Oral and Intangible Heritage of Humanity on November 7, 2003"

What "technology" used? Redundancy!

Various "pathas" of Samhita text: can recover from a corrupted text due to added redundancy: RAID-like! (Redundant Array of Indep Disks)

- Pada-patha: each word in its separate form
- Krama-patha: connects a word in pairs
  - ABCD becomes AB BC CD DE... ("2-mirroring"): 2 copies
- Jata-patha: ABBAAB ("3-mirroring"): 3 copies of A, B, ...
- Ghana-patha (ABBA ABCCBA ABC BCCB BCDDCB BCD...)("10x")

Metrical (similar to checksums!) & Musical

- "Information dispersal"

Human Reproduction! (Oral transmission)

- Use efficient "virtualizers"!

But one thing which you find very interesting, this you look at in over country. I think some of you all are familiar with some text call Vedas right. They say that it has been transmitted across anywhere from depending on who you talk to. It could be 5000 years ago or 3500 years ago, without differing versions, people have studied various Rigvedha. For example, across many places an India, some in north, some in the south. I think as far as I know, there has been only one or two minor changes that is been discovered.

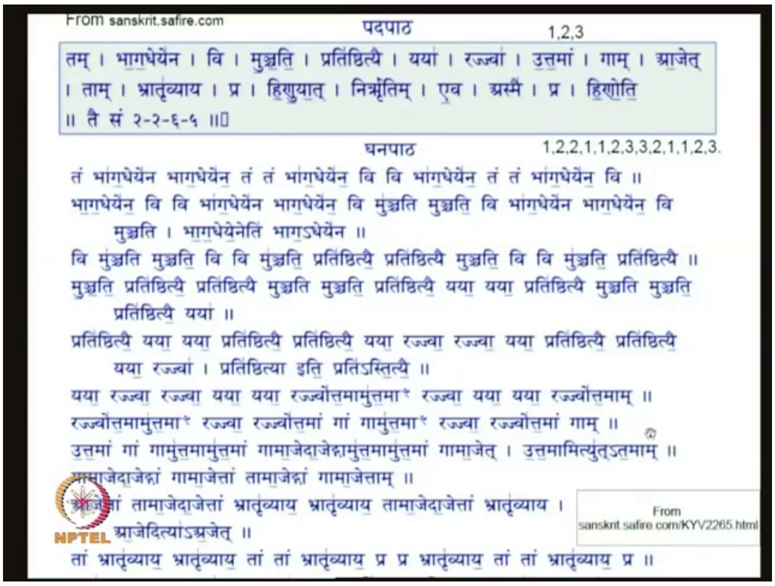
That intact, exact think of you, what is more. You amazing is that these Vedas also are chanted its [FL] for example, right, and they have some notions of pronunciation. Even this are supposed to be identical, because its no difference between Vedas versions except some one or two minor.

Once you people were discovered, and then this absolutely identity question is, how did they do it across so many years, because we are talking about 5,000 years now. So, here, it turns out, they are using multiple technologies here; one is a redundancy, and that redundancy have just talking about, because we talked a bit about it, it turns out if you take this text right. You are what is call as Samhita text, which is the, some of you might know that in the languages, have is notion of [FL] [FL] means you join things together right. For example, Rama plus Eshwar, it is like Rameshwara, it is not Rama. you do not say Rama and Eshwar, you say Rameshwar.

So is Samhita text is basically, the joint text. There is also what is call [FL] which means it is each, what in a separate form. Now, already you have two types of text now. So, there are separate bunches of people, who recite Samhita text, there is another bunch of people who do [FL] they recited and there are actually some 10 types of people. Here, one is called [FL] I just given in 4 or 5 and they these people [FL] and there are few, which are mentioned, they actually do it in more interesting ways, he in [FL] and the [FL] I know it is basically, you are basically, you words call [FL] you basically, spread over search for, did you, when you go from [FL] at a [FL], but [FL] what happens is that they do some kind of redundancy also.

So, basically it is something like what you disk what we called red one mirroring. So, is something is ABCD. Suppose, A is one word, B is another word, C is another word, D is, they will chant it as AB BC CD DE. So, if your right, if anything goes wrong right, if one go has. So, chants one way, somebody has chants some other way, this is some difference, they will say let us talk with this [FL] of person and you will see which one (Refer Time: 57:33) right, because now there are two copies ability. We can see of course, it does not, it is possible that both the guys get corrupted, that is possible, that is why they had other solutions [FL] [FL] is [FL] is very complicated. You have ABC for example, it will be encoded as ABBAABCBAABC, etcetera.

(Refer Slide Time: 57:57)



From sanskrit.safire.com

पदपाठ 1,2,3

तम् । भागधेयेन । वि । मुञ्चति । प्रतिष्ठित्यै । यया । रज्ज्वा । उत्तमां । गाम् । आजेत् । ताम् । भ्रातृव्याय । प्र । हिणुयात् । निर्ऋतिम् । एव । अस्मै । प्र । हिणोति
॥ तै सं २-२-६-५ ॥

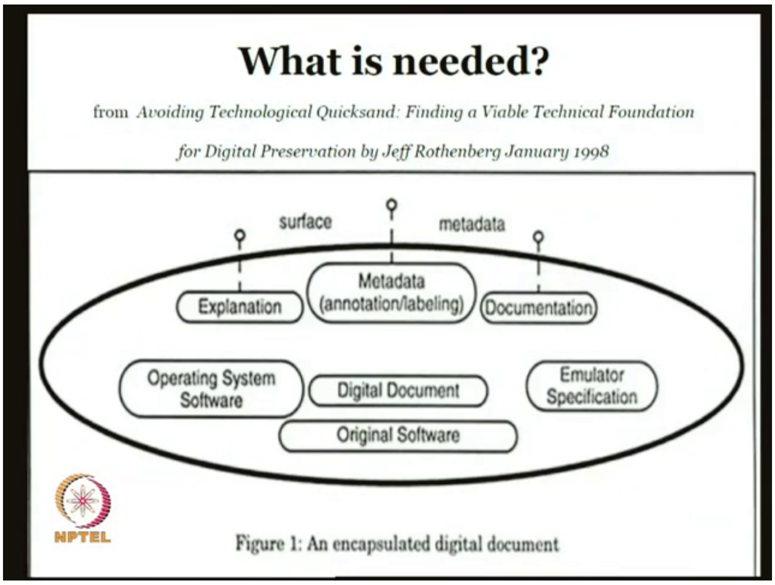
घनपाठ 1,2,2,1,1,2,3,3,2,1,1,2,3.

तं भागधेयेन भागधेयेन तं तं भागधेयेन वि वि भागधेयेन तं तं भागधेयेन वि ॥
भागधेयेन वि वि भागधेयेन भागधेयेन वि मुञ्चति मुञ्चति वि भागधेयेन भागधेयेन वि मुञ्चति । भागधेयेनेति भागऽधेयेन ॥
वि मुञ्चति मुञ्चति वि वि मुञ्चति प्रतिष्ठित्यै प्रतिष्ठित्यै मुञ्चति वि वि मुञ्चति प्रतिष्ठित्यै ॥
मुञ्चति प्रतिष्ठित्यै प्रतिष्ठित्यै मुञ्चति मुञ्चति प्रतिष्ठित्यै यया यया प्रतिष्ठित्यै मुञ्चति मुञ्चति प्रतिष्ठित्यै यया ॥
प्रतिष्ठित्यै यया यया प्रतिष्ठित्यै प्रतिष्ठित्यै यया रज्ज्वा रज्ज्वा यया प्रतिष्ठित्यै प्रतिष्ठित्यै यया रज्ज्वा । प्रतिष्ठित्या इति प्रतिऽस्तित्यै ॥
यया रज्ज्वा रज्ज्वा यया यया रज्ज्वोत्तमामुत्तमा॑ रज्ज्वा यया यया रज्ज्वोत्तमाम् ॥
रज्ज्वोत्तमामुत्तमा॑ रज्ज्वा रज्ज्वोत्तमां गां गामुत्तमा॑ रज्ज्वा रज्ज्वोत्तमां गाम् ॥
उत्तमां गां गामुत्तमामुत्तमां गामाजेदाजेद्गामुत्तमामुत्तमां गामाजेत् । उत्तमामित्युत्ऽतमाम् ॥
गामाजेदाजेद्गां गामाजेत्तां तामाजेद्गां गामाजेत्ताम् ॥
आजेत्तां तामाजेदाजेत्तां भ्रातृव्याय भ्रातृव्याय तामाजेदाजेत्तां भ्रातृव्याय । आजेदित्याऽअजेत् ॥
तां भ्रातृव्याय भ्रातृव्याय तां तां भ्रातृव्याय प्र प्र भ्रातृव्याय तां तां भ्रातृव्याय प्र ॥

From
sanskrit.safire.com/KYV2265.html

So, we will just look at one example of this. This is the actual text [FL] [FL] that is just like that actually is chanted from [FL] like this [FL] 1 2 3 becomes 1 2 2 1 1 2 3 3 2 1 1 2 3. So, you can see [FL] is here [FL] it is 2 2. So, [FL] has to repeated [FL] again becomes 1 1. So, has to be [FL] again one as come [FL] again it will be 2 again, [FL] again 3 3. So, [FL] is 3 [FL] right again, you have to get to [FL]. So, you can little bit chanted as [FL] something on it goes. So, this is the what call [FL] [FL]. So, with thing is, but they have done is they given enough redundancy. So, that even if something happen that able to recovered parts of it and this is done through human beings. So, the, what is called virtualization happening through human beings?

(Refer Slide Time: 59:00)



What is needed?

from *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation by Jeff Rothenberg January 1998*

Figure 1: An encapsulated digital document

I think, I running out of time. So, we will continue from here next time.