The Next Generation of Science Archive Storage

White Paper

A.J.Wicenec, B. Pirenne

27-June-2000

Abstract:

The problem of exponentially growing data flow into the Archive needs to be calculated and addressed at several different points in the current schema. The most apparent place is the the storage technology used, but there are other points, like additional hardware needed to support the chosen storage technology and operational costs of the handling/production of the media. And there is a non-negligible amount of money/media needed to bring the media on-line and maintain them.

Other points which have to be discussed are the persistence/security of the data, and access times and the infrastructure needed to keep in pace with the data inflow. This white paper gives some of the numbers and some other facts for the CD/DVD media system currently in use and for a new system using EIDE hard disk drives. It will be shown that magnetic disks now are a real option, because their price/capacity ratio is currently even better than for optical technologies.

Prices

Assumptions:

Current data flow rate into the Archive: 700 GB/month

Compression factor: 0.5

Number of copies: 2

Total data inflow: 750 x 0.5 x 2 750 GB/month (including WFI!!)

Media Costs (DM):

Medium	*GB/medium*	*Price/medium*	*Price/GB*	*Total/month*	*#med/month*
DVD	4,7	60	12,77	9574,47	159,57
EIDE (IBM)	60	1000	16,67	12500	12,5
EIDE (Maxtor)	60	650	10,83	8125	12,5
SCSI	75	3000	40	30000	10

Total Costs:

The following table contains all applicable costs for each of the media, i.e. these are the only comparable numbers for the different media. The numbers do not contain any estimate about the operational costs, there is a discussion about this issue below.

Media	$/GB in JB	DM/GB in JB
CD-R	70.6	144.64
DVD-R (3.95)	17.6	36.13
DVD-R (4.7)	16.1	33.04
DVD-R (2x4.7)	10.2	20.97
MO	38.1	78.07
SCSI disks farn	22.9	46.91
Sony 12" OD	215.7	442.1
Beo IBM IDE	10.7	21.88
Beo Maxtor IDE	7.8	16.04

For the following calculations of EIDE prices the IBM disks have been chosen. The following sections explain the different additional costs for DVDs and EIDE disks only, which are the cheapest ones in the table above.

Additional hardware

Additional hardware costs in the case of DVDs are the burners 7500 DM each and the computer needed to support the quite high sustained IO rate during the production of multiple media.

Additional hardware costs in the case of disks is comparably low, since no additional drive is needed and the requirements for the computer are relaxed. The only additional costs are special disk mounting cases to allow for fast exchange of the disk drives in the ASTO machine. These cases are cheap (~30 to 50 DM) and one may also invent a operational scheme which does not require these cases.

Additional global hardware costs are the juke-boxes in the case of DVDs and disk cabinets plus racks in the case of hard-disks. A disk cabinet for 8 disks costs about 1000 to 2000 DM, i.e. in the current situation where we assume 60 GB disks it is 2000 DM/480 GB or 4266.7 DM/TB. The cost for a 700 slot jukebox is approximately 30000 DM. With the current DVD technology this is able to hold 2940 GB, i.e. 9337 DM/TB.

Operational costs

In the case of DVDs the operational costs are high, because it takes rather long to write and verify a single media and also requires some manual interaction. In addition there are some operational costs to put the data on-line and maintain it. The operational costs are also pretty much dependent on the number of media which have to be produced. Currently it takes approximately 0.7 hours to produce a single DVD and another 0.3 hours to verify it, i.e. 1 hour per 4.5 GB. There is also the issue of the per-slot licensing scheme of the juke-box software (Tracer).

In the case of hard disks the operational costs are probably much lower, but this has to be verified since there is not a lot of experience around at ESO with the handling of about ten disks per month. But since the number of disks is already much lower (factor of >7) than the number of DVDs it can be assumed that the operational costs will in fact be lower. The time to write and verify the media is also much lower and once a bunch of disks is mounted there is no interaction needed anymore.

Costs to bring data on-line

The time needed to bring one magnetic disk on-line is almost negligible, because it just takes some seconds to slide it into a slot and mount it. However it should be mentioned that the comparison here is not straight forward without having any experience with hard disks, but also here the factor 7 in the number of media will certainly help a lot to make hard disks cheaper in this respect.

Random Access times

Here it's really clear who the winner is: Hard disks have a random access seek time of some milliseconds and the sustained weighted mean transfer rate is of the order of at least 12 MB/s. In our case where we are dealing mostly with big files the latter number is probably even much higher. In the case of DVDs the random access time is mostly dominated by getting the media into the drive and mounting it. The transfer rate with current drives is of the order of ((6 MB/s)). This discussion might seem to be not very relevant, but one has to remember that we are processing all the service mode data and all the calibration data which is about two thirds of all the data. In addition one has to take into account that we probably have to migrate all of the data to the next version of DVDs in about two years, i.e. burning/verifying a lot of new media and in particular reading all of the data off the DVDs.

Data persistency/security

Long term persistency

On a first glance one would probably say that DVDs are a lot more secure in terms of persistency. But one has to take into account the expected usage duration of the current technology. Since the current capacity of DVDs is already too small for our needs we have to switch over to the latest technology as soon as it becomes available. To be precise the next generation of DVDs will be really a different technology, because it's a multiple layer and/or double sided media. It is certainly not just an increase in capacity like we saw saw recently with the 4.7 GB DVDs. Whenever we switch over to the new DVDs we will have startup problems, not to speak about having to buy new burners for all the sites and drives for the juke-boxes. We experienced these kind of problems with the introduction of DVDs and CD-Rs already. Thus even if DVDs offer a better long-term storage persistency, we are just not able to profit from that fact. As the experience with the first CDs, the first CD-Rs and the first DVDs showed that data has been corrupted during writing and/or lost after a short period of time. Since we have to follow the latest technology pretty closely we will be kind of beta testers and certainly face these kind of problems. DVD capacity has not been increased a lot during the last year and at least for DVD-R the situation is not very clear (single vendor). At least the market pressure on the increase of DVD capacity is not as high as for magnetic disks, since the main driver for this technology is not the data storage area, but the entertainment sector.

Short term persistency

Another point is the short term persistency. Everyone is afraid of disk crashes, but in fact this happens very rarely and usually it is not happening without any prior sign, i.e. one can do preventive maintenance. The MTBF of hard disks already is very high (of the order of 20000 hours MTBF) and in any case we will have a second clone disk on the mountain which in my view should just be stored on a shelve and thus it is not even spinning. The MTBF of a non spinning disk drive should be really high! Also in the case of disk drives we will have to copy the data to new drives, simply because of storage efficiency and physical space constraints of the whole archive. The disk capacity is doubling (at the same price!) within about 18 month. One fact which I found just recently is that the quoted bit failure rate of DVDs in reading is slightly worse than for hard disks, i.e. the data on the media might still be perfect, but drives are not able to read it correctly.

Future development

The most important point in my view is the following: In about 3 to 5 years (latest) we will find holographic storage devices on the market which will behave exactly like disks, but have the advantage of optical media and very high capacity/volume. Whenever they are an established product we may easily switch over to them, probably even with exactly the same kind of infrastructure as we would have for hard disks.

Infrastructure

DVD

Because the increase of the capacity of DVDs does grow slower than the data rate, we have to foresee a growing rate of parallelism in the production of the media, because else the daytime operations will simply be unable to keep pace with the data created during the previous night. Truly parallel operations, even if very well automatized, is error prone and there is always some hand-work to be done with the creation of DVDs. Shipping and packing is an almost linear function of the number of media one has to handle. The picture for the future ASTO machine with all those drives and

robots as shown during the DMD review with the DG looks a bit scary to me. On the other hand we do have a lot of experience with this kind of infrastructure.

As for magnetic disks I would propose the following infrastructure:

EIDE

Mountaintop

Industrial grade PC with 4 EIDE channels and 16 disk slots running Linux. This could be a rack mounted PC with rack mount slot boxes. One such a system per UT plus one for WFI and one for the rest of La Silla. The price for such a setup should be substantially less than 10000 DM. I expect a price like for one DVD burner. The PC could be equipped like one of our Beowulf masters. If the price allows we could even buy two PCs for redundancy and performance reasons. Another option for Paranal would be to buy just two of these systems (with two PCs each) and centralize them like the current ASTO setup. The configuration of the system itself is kind of a custom installation because every two disks have to be mirrored. The exact procedure has to be tested and iterated to reach an optimum, but in principle I would assume that during the filling up of one disk pair we could have four copies of the data. One operational night disk pair plus one disk pair which is currently filling up. The operational night disk pair can be cleared during daytime operations. A detailed operations plan has to be prepared and tested before operations can start.

Garching

Initially the same as one of the mountaintop systems. After that just additional 16 slot rack mount boxes as need arises. To control all that (remember that theses are EIDE disks) I would propose to procure one PC for each of the disk boxes and an additional master PC and a network switch after the procurement of 8 disk boxes (+8 PCs). This is still part of the supporting infrastructure, but I leave it to you to think about the implications (this should be discussed in a separate document). To build up the full picture we should buy a separate switch/router for each four masters. The last level is not part of the supporting infrastructure anymore, but it I think it would be worth to invest this relatively small amount of money. With the current situation in mind (see the table) we would have a new Beowulf system with 1 master/8 nodes configuration after about 10.9 month. With the expected data rate of the VLT instruments we can just keep the same speed, given the expected development of disk capacity, i.e. we just ramp up capacity but not the slots/PCs. Another option would be to keep the number of PCs/TB constant at about 1 PC/1 TB. If the disk capacity is growing faster than the data rate we could even think about migrating some of the older data to newer disks. For the VST and maybe temporarily also for VIMOS/VLTI the growth rate will be higher. For VST and VISTA again I would like to leave the implications and calculations to you.

General

The disks should be bought on-the-spot, i.e. as needed with very few on stock, because of the price/capacity development. The market and the data flow have to be constantly monitored to get the optimum price/media and media/month ratios. To guarantee on-spot delivery we should identify a hardware vendor giving good conditions for such a contract maybe including the PCs. The disks have to come system ready, i.e. with a file system. For our main purpose we do not want to have any RAID level but just plain disks. For the ease of use we still can concatenate the disks via the Linux kernel to be able to view them through one mount point. The PCs have to come fully configured, including the setup for the disks and the network. Also for the PCs the market has to be monitored to take advantage of new processors and other developments.

Operations

As for operations the picture will change of course. Instead of creating optical media, magnetic disks have to be exchanged, but the implementation should not be too problematic.

Physical volume

Since we have rather limited physical space at ESO headquarters in Garching this point has to be discussed as well.

Extrapolation with expected data rate

From the discussions given above we do not see any point to stay with DVD technology even with the current data rate. As pointed out in the assumptions the current data rate includes the WFI. Since this data is currently shipped on tapes it even has a much higher impact on the operational costs, while the media costs are substantially lower. The expected data rate from the VLT will soon catch up with the WFI data rate and with the advent of VST (not to speak about VISTA!) we will certainly hit a point where the operational costs due to the necessary parallelism for the production of the DVD media will become unacceptable.

Conclusion

I think this is the right time to discuss this, because we are about to install a new system at La Silla end of this year. We could take advantage and switch over to a scheme which is capable of dealing with the expected data rate for the whole of ESO. Even for VST and VISTA this system is feasible assuming that we will have something like 280 GB disks by mid 2002 and 560 GB disks 18 month later. With this picture in mind, I tend to be much less worried about the coming data. There are some other points which I did not mention, like the availability of the data (some of the

NTT data is not on-line since a long time) and the fact that we would not need to extract headers. Not to speak about the fact that we would add computing power in perfect sync with the volume of data. In this way you make absolutely sure that the system is able to not *only store* the data but also to *process* it. In addition the PCs could also be made really responsible for their data, i.e. they will not run idle but in fact will always check the data and the disks and report the status back to a DB. In this way we could be sure at any given time that our data holdings are readable and still o.k. Currently we only do spot checks when the data is read, because a request has been submitted and full checks are only done when we move to a new media type.

Side remarks

Archive Scans

A simple calculation shows another incredible advantage of the proposed hard disk solution:

Assumptions: Disk read rate 15 MB/s (sustained). One PC per 16 disks á 60 GB.

Conclusion: It takes about 16 x 60 x 1024 MB/ 15 MB/s / 3600 s/h = 19.5 hours to scan the whole Archive, no matter how many Terabytes it has in total!!

Pipeline processing

Along this line I am also thinking in the direction of the new pipeline infrastructure. Think about a pipeline which sends requests to a very generalized data object, which has not only data and methods combined, but also computing power! You do not ask for data anymore, but only for results!

Since queries are sent to a DB controlling all of the resources including the data, the available methods and the computing resources, it can report back whether it is able to do the requested processing by itself or not.

Queries upon the available recipes and computational resources should also be possible. Upon these there could eventually a system be build that allows to configure remotely the processing steps to be carried out in a graphical manner.

Virtual Observatory

Expanding these thoughts to a Virtual Observatory in a really global sense is what I believe George Djorgowski had in mind in trying to force us to be visionary. I think with the huge dynamical range of data volume affected by hypothetical queries/scientific programs (one astronomical object to whole sky in tiny pixels) we have to let the data object know or decide where the methods associated with it can be or have to be executed. The range is also huge and spans from a palm size computer in your hand to a supercomputer somewhere in a supercomputing center. Of course the resulting data object has to be associated with this knowledge as well, because the user does initially not know about the size and properties of the resulting data object.

References

U2WSCSI/LVD to UDMA/66 RAID controller	http://www.synetic.net/AccuSys/acs-8900.htm
U2WSCSI/LVD to UDMA/66 RAID subsystem	http://www.synetic.net/New_Folder/raidkit460.ht
Mini EIDE RAID-1 subsystems	http://www.synetic.net/Poware/praid-1.htm
	http://www.synetic.net/Accordance/ARAID99-300.htm