Experience with Redwood/D3 Tapes

Hartmut Fichtel, Deutsches Klimarechenzentrum GmbH, Hamburg, Germany

ABSTRACT: *DKRZ currently operates its data management system with the UniTree product running in a Convex hardware environment. Archival technology in parallel use comprise Metrum/VHS and STK robotic systems with 3480, 3490E, and Redwood/D3 tape devices. Since installation, early in 1996, the new D3 tape drives have been systematically tested for functionality, performance, and reliable operation by continous stress tests both under the native operating system and the UniTree production environment. These test results will be presented together with first operational experiences in production use.*

Introduction

DKRZ is a mainly federally funded institution to provide the German climate research community with the necessary computing resources to perform numerically highly complex climate simulations. This task implies a virtually unlimited requirement for installed compute power in the first place, but directly correlated with available compute power are the corresponding data services requirements, both in terms of static storage capacity and dynamic data access.

The following diagram shows the development of installed compute power at DKRZ from 1985 to 1996.

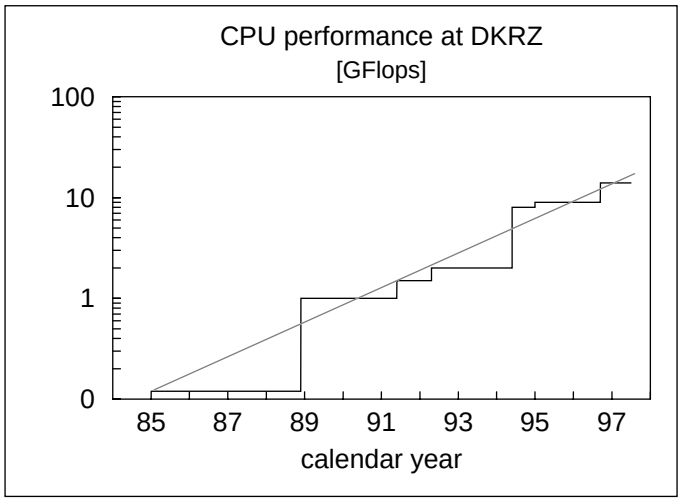

Figure 1: Installed CPU power at DKRZ.

This local development seems to be well synchronized with the overall technological development of supercomputer technology showing an average increase of almost a factor of 2 every 2 years.

Tightly coupled to the compute power dedicated to climate simulations is the data rate in terms of long term storage. As a general rule it turns out that **1 Flops compute power** (sustained) generates **1 KByte of data** per year which is of sufficient interest to justify long term storage. So the above diagram can also be interpreted as the development of actual annual long term storage rates if the y-axis is replaced by [TByte].

There are still more interesting rules of thumb defining the overall correlation between available compute cycles and the resulting requirements for the data management system which have been remarkably constant in the past. The second rule refers to data generation rates as opposed to long term storage: roughly 25 % of the total data generated will be stored for less than a year and thus does not contribute to long term storage requirements. This observed relation refers to climate model output only which is the dominating sort of data at DKRZ. Other types (e.g. observational data for model validation, general system backups etc.) display different characteristics. The third rule defines the resulting overall data access requirements: for every byte generated there will be 3 bytes accessed, so model data turns out to be relatively active.

With these 3 rules applied and the expectation that continous funding will be available it is rather easy to estimate the future requirements for the data management system which are summarized in the following table.

Table 1. Estimated future data management requirements

	96	97	2000
CPU performance [GFlops]	10	30	100
data generation rate [GByte/day]	40	120	400
data archival rate [TByte/year]	10	30	100
archival capacity [TByte]	25	55	215
average data moved [GByte/day]	160	500	1600
average transfer rate [MByte/s]	2	6	20
peak transfer rate [MByte/s]	20	60	200

Development of Storage Capacity

After a rather long test period starting in 1991 the decision was made to use the UniTree product in a Convex hardware environment as the central hierarchical storage management system with start of production use in April 1992. The following diagram shows the temporal development of long term archival data which closely reflects the computational power available over time.

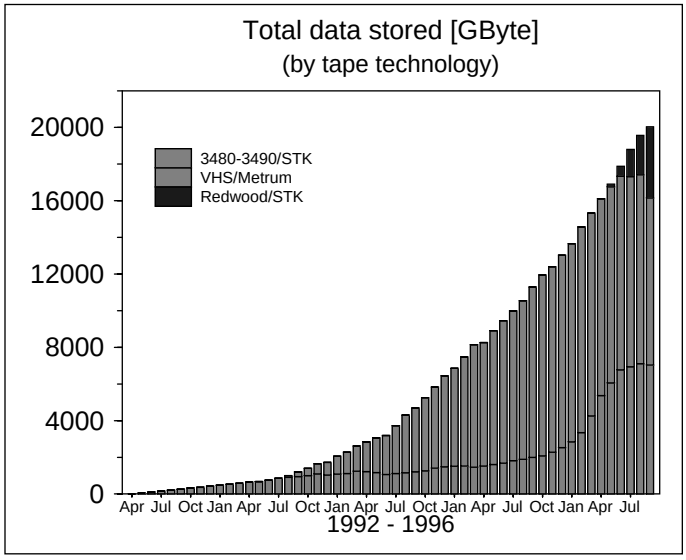

Figure 2: Temporal development of data stored.

Due to the (stepwise) exponential nature of the data growth function and the limited storage capacities of robotic environments until only recently the data management operations were mostly constrained by that limited capacity. Production use in April 1992 started in a limited operating mode since only 2 STK silos with 3480 cartridges were available resulting in a nominal storage capacity of 2.5 TByte. General production use could not be started before September 1993 when a Metrum RSS-600 robotic system with a (nominal) capacity of 8.7 TByte based on T-120 VHS cartridges was turned into production use. One year later VHS capacity could be increased to 13 TByte by utilizing T-180 cartridges. This required an internal repacking operation of roughly 4 TByte of data.

It was easy to forecast, however, that some time in 1995 the combined capacity based on the existing 3480 and VHS media in the robotic environment (15 TByte) would be filled up again. Actually, it is in practice not possible to use the full nominal capacity. Due to the increasing overhead in repacking fragmented tape media as a function of the filling factor this would imply an ever increasing amount of internal repacking. The number of tapes to be copied is $n = \lceil 1 / (1 - f) \rceil$ with filling factor $0 < f < 1$ and the total data to be copied thus becomes $c_t = (f * c_0) / (1 - f)$ for media with capacity c_0. The following table shows the number of tapes to copy, the total data moved and the necessary time for the repack operation for media with 50 GByte capacity with an effective read/write rate of 10 MByte/s to regain one single free medium of that capacity.

Table 2. Repack overhead as a function of filling factor.

f	n	c_t [GByte]	time
0.7	4	140	~ 4 h
0.8	5	200	~ 6 h
0.9	10	450	~ 12 h
0.95	20	950	~ 1 d
0.99	100	4950	~ 6 d

In a practical situation the filling factor of course will not be constant, and tapes with small filling factor will be compacted first. In general, however, it cannot be expected to utilize more than 70 to 90% of the nominal capacity depending on the file attributes and the removal behaviour of the users.

Consequently DKRZ had enough interest in the announced D3 helical scan technology with its inherent high specific storage density to place an order for initially 4 drives for installation in mid 1995. The actual delivery was delayed into the year 1996 which generated a difficult operational situation in 1995 as the robotic capacity was completely filled up as expected. So a short term decision had to be made to upgrade the 3480 tape drives to 3490 and replace roughly 10.000 3480 media by their 3490E counterparts with a nominal capacity of 800 MByte per cartridge. This added enough robotic capacity to survive the D3 delivery delay, but it also implied another repack operation of the 10.000 media involved which could be completed by the end of 1995.

Data Management Environment

Next to capacity the major load parameters are the data access requirements which also evolve as a function of compute power

used. The next diagram shows the development of aggregate access rates between the HSM and the client systems.

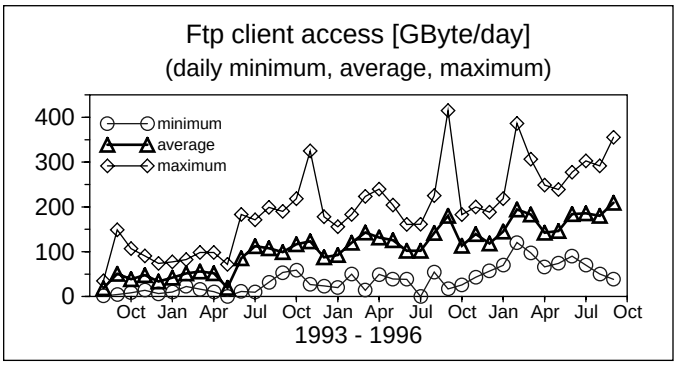

Figure 3: Temporal development of aggregate access rates.

The average daily data exchange currently is about 200 GByte/day with busy days of about twice that amount. This translates to an average sustained data rate of slightly more than 2 MByte/s which requires according to our experience a system enabling 20 MByte/s sustainable peak rate in busy intervals to guarantee acceptable turnaround times for client requests.

The next diagram shows the current hardware environment of the data management system which seems to be rather well balanced with the required performance both in terms of capacity and aggregate access rates.

Redwood/D3 Technology

The D3 tape technology is rather different from the well known longitudinal recording technologies used in computing environments for decades, like 9-track tapes or 3480/3490 cartridges. It is a helical recording technology (like VHS) which has been developed for the video industry to record movie data in high density format. The relatively slowly forwarding tape is moved with a fixed angle along a rapidly spinning recording cylinder (scanner) to produce a high relative velocity between tape and recording heads generating helical tracks on the tape as shown in the next diagram.

The physical recording is rather complicated so that the actual bit stream recorded on the tape is totally different from the stream sent to the device from the user buffer and is depicted in the next diagram in a simplified form. The complete details can be taken from the ECMA standard 210 “12,65 mm Wide Magnetic Tape Cartridge for Information Interchange - Helical Scan Recording - DATA - D3 - 1 Format”, published in 1994. The principal recording steps are the following:

1. the user data received via the peripheral channel are passed to a packet generator - possibly after compression - which adds both a packet header and trailer,
2. the packet is passed to a scan group generator which is the basic unit for formatting recordings on the tape. A scan group consists of 6 consecutive helical tracks. The data bytes from the packet are transferred into 2 interleave buffers - depending on even/odd byte addresses - with inclusion of Scan

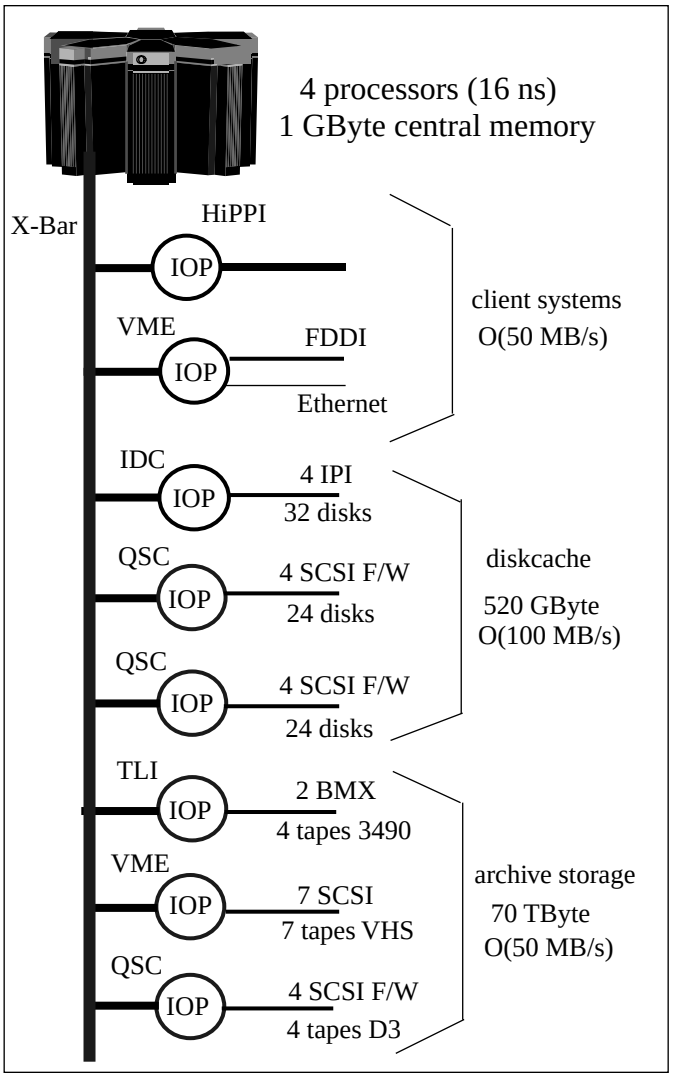

Figure 4: Current data management hardware.

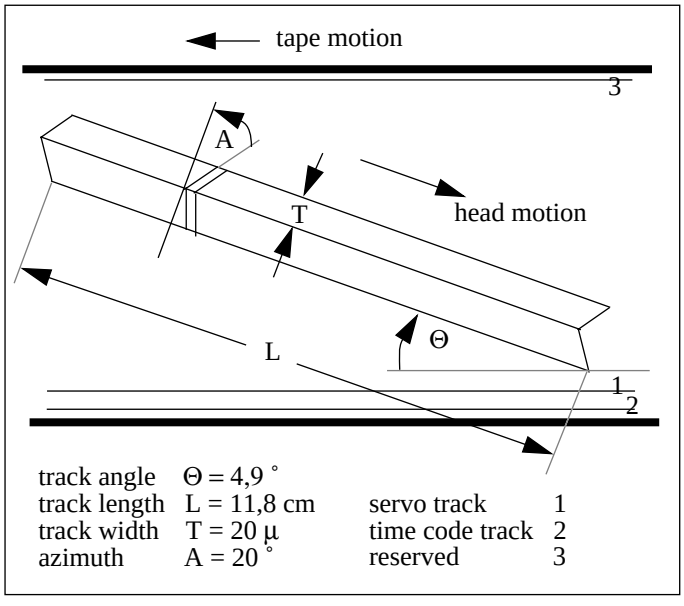

Figure 5: D3 tape geometry.

Group Start Data (SGSD), Helical Time Code (HTC) for block addressing, and Outer ECC codes. Each buffer section contains data to fill 3 helical tracks.

3. Data from the interleave buffer is formatted into sync blocks with inclusion of 8 inner ECC bytes.
4. Each byte is recorded to tape after applying an 8:14 code conversion generating the actual channel bits to restrict the number of consecutive identical bits to seven.

As a third level of protection a separate ECC-3 scan group is generated after at most 24 data scan groups. This ECC scan group and the corresponding data form a Super Group.

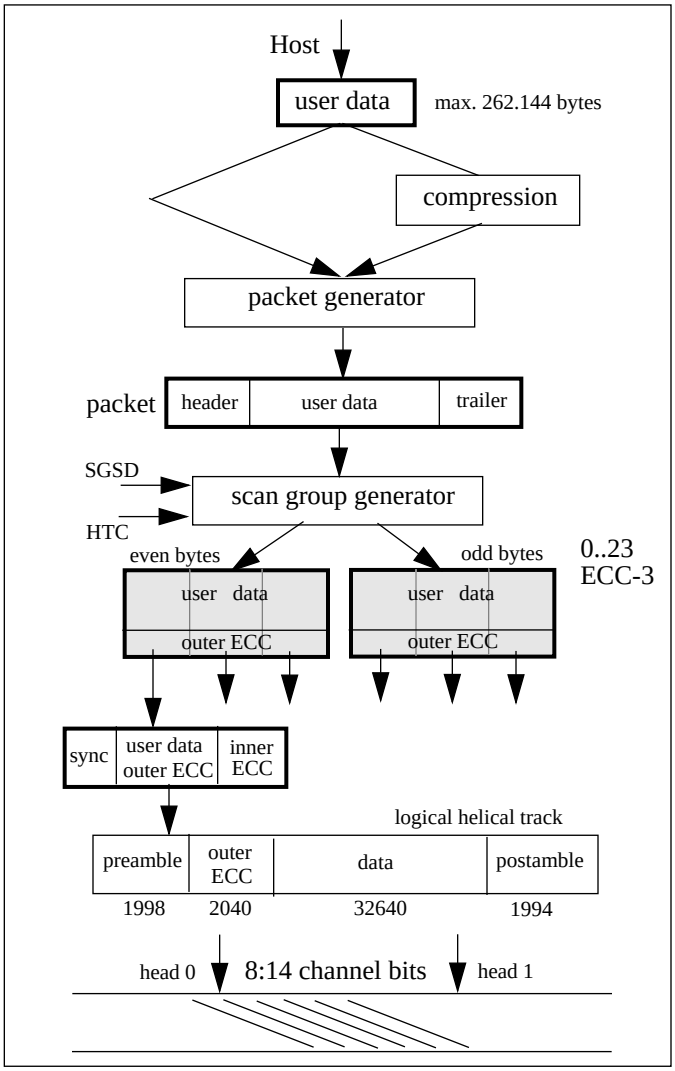

Figure 6: D3 physical recording.

Another major difference to the traditional tape technology is the D3 tape format defined by the ECMA standard. In particular there is a reservation of a considerable storage area at the beginning of the tape used for control information. This general layout of the tape is depicted in the next diagram.

A DID (density identification) scan group repeated 256 times starts the tape followed by an ILH (Internal Leader Header) consisting of 2 scan groups which contain the major administration information for the volume. This area will be read each time the cartridge is mounted and updated each time it is unloaded, provided writing is not physically inhibited. This area is protected by an ECC-3 scan group. If due to heavy usage this area degrades in quality the data can be rewritten into the SEP (Separator) area which provides 125 scan groups of buffer space. PAD scan groups can be written anywhere on the tape, but before the EOD (End of Data) scan group the PAD is required. The remaining area between SEP and PAD/EOD constitutes the data area of the volume.

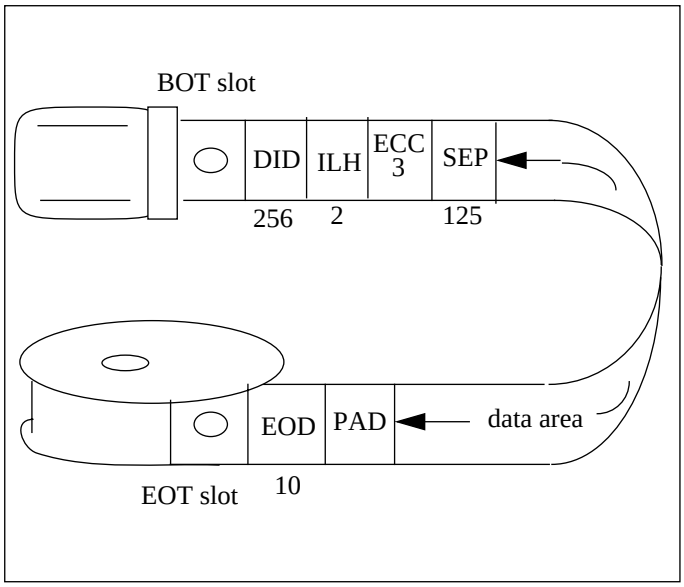

Figure 7: Partitioning of D3 cartridges.

Redwood/D3 Experience

As mentioned earlier the contracted delivery of the D3 equipment on order was delayed several times until the delivery finally took place on February 1, 1996. At that point in time only type A cartridges (10 GByte/medium) could be delivered which was not considered a problem for the initial testing period.

The test cycle was structured into 3 phases:

1. functionality and stress test under the native operating system to both verify the new equipment and the new drivers,
2. performance tests under the native operating system,
3. performance tests in the final UniTree production environment.

The test environment consisted of another C3800 (similar to the production system) with a peripheral configuration that should not constitute a performance bottleneck, as depicted in figure (8).

The first phase was performed according to the following pseudo script:

```
for (;;;) {
   load 4 media type A (10 GByte) into a drive each;
   write 20 data sets of 500 MByte each;
   rewind;
   fsf 1; fsf 2; fsf 3;
   read/verify data set;
   fsf 12;
   read/verify data set;
   bsf 18;
   read/verify data set;
   unload all media;
}
```

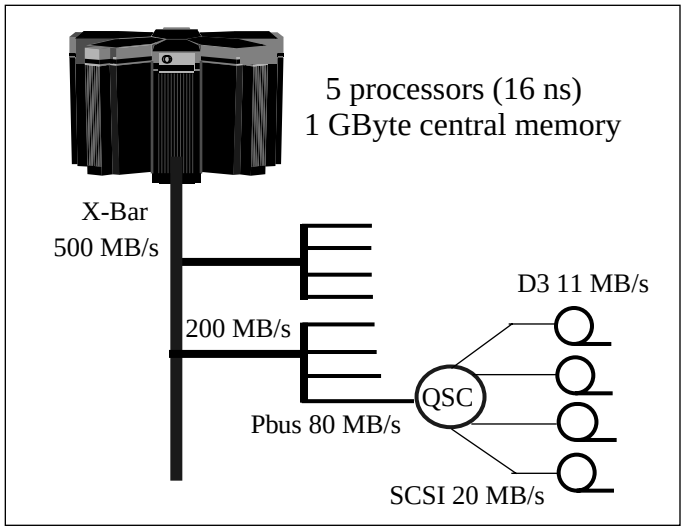

Figure 8: Hardware test configuration.

This type of test script was run for 24 hours a day over a period of 6 weeks generating some confidence that the equipment would be working under conditions of heavy duty cycles.

After the passing of these first confidence tests the next step was to proceed to basic single stream performance tests under the native operating system. These tests were performed with a small test program addressing the drives as a raw device with unlabeled tapes, using 256K buffer size and including the end of file processing into the timing. The results for the case “single stream write” are shown in the next diagram

The measured curve is very close to the expected one if the streaming rate of the device of just under 11 MByte/s is assumed for the transfer and a fixed overhead per file of approximately 30 seconds is needed independent of the size of the file. This constant time is mainly used for the end of file (tape mark) processing at the end of the transfer.

A similar curve is to be expected for the case “single stream read” except that the fixed overhead should be much shorter since only a tape mark has to be read and none written at all. The processing of one tape mark needs approximately 8 seconds but it typically takes no extra time since

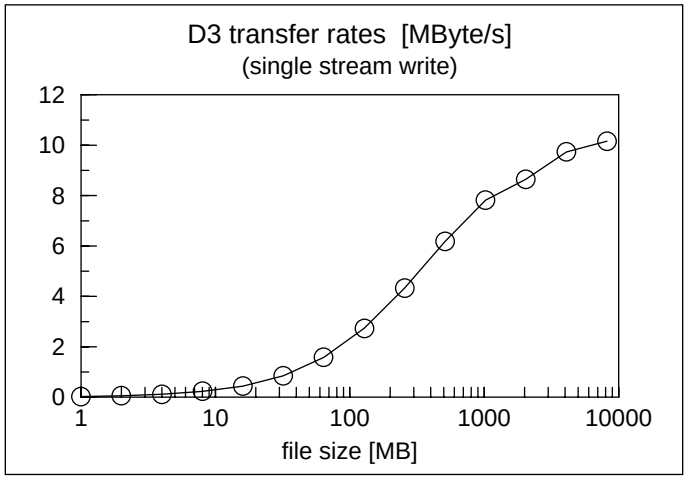

Figure 9: D3 rates “single stream write”.

the tape mark scan group including its preceding PAD scan group can be expected to be fetched from the 64 MB device buffer without further tape access. The measurements are depicted in the following diagram.

The same measurements have also been made with all 4 drives in parallel use. The results have not been different from the single stream case proving that the quad SCSI controller attached to the PBus is being able to drive all 4 SCSI buses and does not saturate at 40 MByte/s aggregate transfer rates, a fact which has been known before from disk performance tests.

The last diagram shows the performance of D3 technology again as a function of file size in the context of the UniTree production system in comparison with the existing technologies 3490 and VHS.

The results are rather disappointing with respect to D3 since the maximum performance seems to saturate at about 4 MByte/s which is far below the device streaming rate. File mark processing overhead is no explanation since the Unitree tape movers do not write tape marks after each file but only after a configurable amount of data which at DKRZ for the

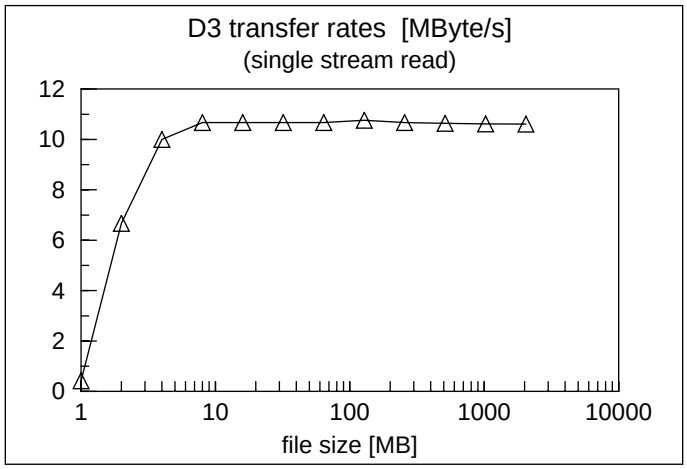

Figure 10: D3 rates “single stream read”.

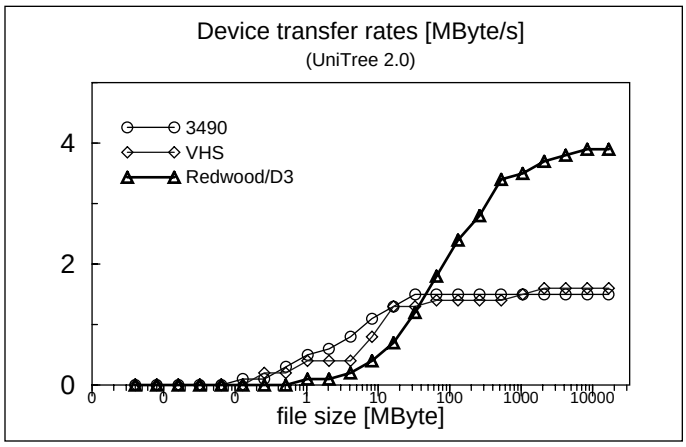

Figure 11: Relative performance of tape technologies.

case of D3 media is about once every GByte of data. Currently there are no provable explanations for this behaviour but there is some speculation that the rather inefficient server-server communication scheme within UniTree keeps the device from streaming.

Remarkable is the relative performance of the VHS drives compared to 3490 which achieve about 75 % of their peak transfer rate already for moderately sized files.

Summary

Following an initial test period of approximately 3 months the D3 tape drives have been used in the normal production environment for about 5 months by now. The devices perform as specified and appear also to be reliable in the production environment with tape duty cycles approaching 30 to 40% according to current experience. Longer production experience is needed to assess the long term behaviour of both tape drives and media in a heavy use environment. The drive performance in the specific application environment of DKRZ is unsufficient and has to be addresses as a special software issue.