STUDY FOR POSSIBLE STORAGE SOLUTIONS FOR KLOE

F. DONNO, E. PACE

INFN, Laboratory Nazionali di Frascati, Via Enrico Fermi 40, 00044 Frascati, Italy

in *collaboration with the KLOE DAQ and Off-Line Analysis Groups*

The requirements of KLOE represent a challenge in this field, since we have to deal with data rates (about 10^{11} events/year) considerably higher than those of typical collider experiments such as LEP or TEV-1, albeit a factor ten smaller then those foreseen at the LHC. It is estimated that the average event size is 5 kbytes, corresponding to a total bandwidth requirement of 50 Mbytes/s. The amount of data collected each year of running is then of the order of 500 Tbytes. The total storage requirements are even greater due to the MonteCarlo data that will be required and could only be managed with special robotics and a good computer organization. In this paper we discuss the characteristics of actual tape drive systems and robotic libraries that could be of interest for KLOE (DLT, IBM 359o, STK Redwood, AMPEX DST 310) and results of test are shown for DLT's (under HP UX and Digital UNIX). The software that will be used for reading/writing data is also described. Home made software will be employed to drive the robotic parts.

1 The KLOE Data Acquisition General Requirements.

The KLOE experiment [1] at DAΦNE (Frascati) will run in the middle of 1997. Its major aim is to perform CP violation studies at sensitivities of 0(10-4). The KLOE data output of $\mathcal{O}(10^{11}$ events/year) must be handled [2] maintaining biases to values smaller than the experimental sensitivity. The maximum expected data rate from the KLOE detector, at full DAΦNE luminosity, has been estimated as 10^4 events per second of size of 5 kbytes each in average, corresponding to a total bandwidth of 50 Mbytes/s. The characteristic of the KLOE DAQ system, that is extensively described elsewere[3], is to be completely scalable: raw data from detector are collected in parallel chains, sent using a FDDI switch to a farm of CPUs, and written on tape using several devices in parallel. Using a different number of chains, of FDDI ports, of CPUs, and of tape drives, the aggregate throughput substainable can vary. In particular, subdividing data to be written on tape in 10 streams, it is possible to use in parallel 10 drives that can substain a throughput of 5 Mbytes/s each, instead of a single device that should have to substain a throughput of 50 Mbytes/s.

Raw data, coming from the data acquisition system, will be processed by a very powerful production engine that will constantly work to provide data sets ready for final analysis. The production process not only will translate raw data in physics quantities, but it will perform a kind of offline-prescaling, reducing data to be analyzed to those interesting, plus a certain randomized fraction of other events. A splitting of the events in different streams could also be performed. The processed data will be kept on-line in a library for at least one month, and will be served via a catalog (FATMEN [6]) to users. A very efficient staging system will take care of spooling data on disks on demand, keeping disks clean, covering collisions, and disks unavailability.

This system can be easily integrated in the current analysis environment. KLOE decided to employ an analysis driver (the ANALYSIS-CONTROL [7] product developed at Fermilab and modified slightly to implement KLOE specific needs), which also offers a very nice interface to most of the delicate operations in the analysis.

2 Quick Overview of Storage Technologies of Interest for KLOE.

We performed a market research to see which of the commercial tape technologies is better suited for KLOE in terms of performance and price.

Today there are essentially two recording techniques: helical scan recording and linear recording.

.2.1 *Helical Scan Technology.*

Ampex DD2 recorders are able to transfer data at up to 15 Mbytes/s. Three cartridge dimensions exist, small , medium, and large, corresponding to different capacity (50, 150, and 330 Gbytes/s). The DST 600 serie has a rotating head with 4 pairs of heads mounted at 90 degree intervals along the drum circumference. The data recording format is DD2: data is recorded in a helical fashion with 3 longitudinal tracks along one edge of the tape. One of these tracks is for servo information and therefore unavailable to the user. The other two tracks are used for file structure, labels and similar information.

Redwood is the new StorageTek SD-3 helical scan tape cartridge subsystem, codeveloped with 3M. Its capacity can be 10 or 15 Gbytes (5 Gbytes in the near future). Its actual ESCON interface can sustain 11Mbytes/s, while a SCSI-2 interface will be available soon. StorageTek supports various computers (such as CRAY, CONVEX, Silicon Graphics, and RS/6000) and operating systems (Cray/UNICOS/COS, DEC/VMS/Ultrix, IBM/AIX, Sun/Solaris,...). An Internal Leader Header (ILH) is architected into RedWood's tape recording format. This keeps track of several usage factors including the number of read/write passes and a history of tape mounts. ILH also contains index search information such as logical block ID, corresponding to physical sector numbers, etc. This facilitates fast random-access to data at any point on the tape. A 3:1 compression is possible using ICRC code.

2.2 *Linear Tape Technology.*

Since in serpentine recording technique the heads do not rotate, they do not scratch the tape on STOP operations as it happens for the helical scan technology. Digital and IBM have adopted this recording mode on half-inch metal powdered tape cartridge. IBM stated that some of the reasons that brought them to choose the linear technology are mechanical simplicity which gives reliability advantages, smoother START/STOP operations which grant bigger flexibility, and longer head and tape life.

3590 is the new IBM high performance tape system, also known as NTP (New Tape Product). The Magstar drive has 16 heads and uses a 128 tracks serpentine longitudinal recording technique. It offers a 9 Mbytes/s data transfer rate, using

10 Gbytes cartridges. The drive has a SCSI-II controller and supports FW SCSI adapter (15 Mbytes/s). An ESCON interface will be available. A special error correction code and servo tracks written on tape ensure data integrity: 10^{14} is the mean bytes read error and 25 years the shelf life. A 1:3 compression algorithm (LZ-1) can be used.

Quantum produces today two models, DLT2000, and DLT4000, that are proposed also by many other vendors (DEC, SGI, Compaq, HP,..). These models have a SCSI-2 S/E or differential interface, 1.25 Mbytes/s nominal throughput, a 10 (for DLT2000) or 20 (for DLT4000) Gbytes native capacity. A modified Lempel-Ziv data compression technique can double the effective capacity and throughput. At the beginning of 1996, a new model will be available, called DLT6000. It will have a 30 Gbytes native capacity (60 Gbytes compressed) and a native throughput of 5 Mbytes/s. All the DLT models have the same physical dimension. After 20 years on shelf, less than 5% of data are left for demagnetization. Reliability specifications include head life of 10,000 hours, media usage averaging of 10,000 passes, and MTBF of 80,000 hours. The hard error rate is specified as 1 error in 10^{17} bits. The undetected error rate is 10^{30}.

3 Libraries.

3.1 Libraries for DST.

DST41O is a 7 cartridge loader for small, medium, and large tapes. DST800 is a larger library: able of holding only small tapes, it can house up to 4 drives and 256 tapes.

3.2 *Libraries for RedWood.*

Due to the same physical dimension of Redwood cartridges and IBM3490, STK old Powderhorn can be used for Redwood, just changing the tape drives. It can contain a maximum of 4 drives and up to 24 storage modules, each holding 6,000 tapes.

3.3 *Libraries for IBM 3590.*

A 10 cartridges loader, RACL (Random Access Library), is produced by IBM. It is also possible to assemble up to four RACLS all together.

Also IBM 3590 cartridges have the same form factor as IBM 3480/3490 cartridges and than could be placed in the same libraries already used for these older tapes. For example, the IBM 3494 Tape Library Dataserver that is composed of a control unit, up to **3 drive units, and** up to **7 storage units. A control unit has up to 4 drives and can host at maximum 210 cartridges. Up to 7 storage units can be attached to a control unit, each providing the capability to store an additional 400 tape cartridges. The maximum IBM 3494 configuration is a combination** of storage units (maximum of 3), but the total number of optional units cannot exceed 7. The maximum capacity of this library is 3040 tapes.

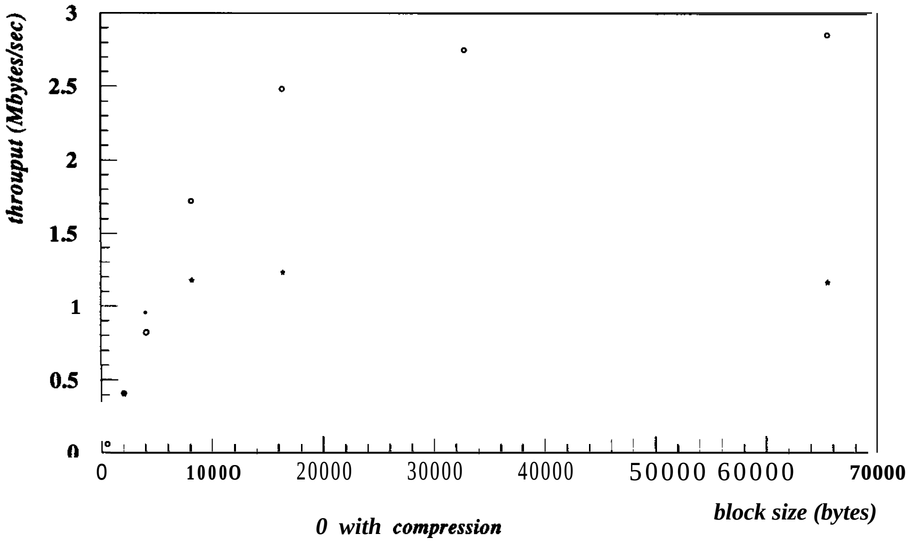

Figure 1: DLT tests made using a HP 735/125 running HP UX ver. 9.05 and copying a fixed pattern on tape.

3.4 *Libraries for DLT.*

Loaders for 5 or 7 cartridegs exist. ODETICS produces a library that can house up 52 cartridges and only one drive. More interesting for KLOE is a larger ODETICS library, ATL2640, that has a modular architecture. Each unit can house three DLT drives and can hold up to 264 cartridges. A single robotic controller can support up to **5 library units.**

4 DLT tests.

Some tests have been performed in Frascati using a DLT2000 drive by DEC (TZ87) on a DEC 2100 4/275 running Digital UNIX version 3.2 and on a HP 735/125 running HP-UX version 9.05.

The software used to perform these tests was written in FORTRAN and C. Using different block sizes, a piece of memory, or a given set of integer numbers, or a binary physics file was written on tape, in unlabeled and labeled format.

Results are shown in figure when copying a fixed pattern from memory to tape, varying the block size used. This case is particularly good for compression because the algorithm used create a new compression table for each logical block copied. In this case the translation table has only one entry. Test with MonteCarlo generated data show good results as well.

5 An Example of Data Organization: the Production of One Million KLOE Events Using a DLT Loader.

We tested our choices and the data flow organization generating and processing 1 million of KLOE events. All physics modules were linked together with Analysis-Control. YBOS[4] has been internally used as memory management system.

A complex data-base structure has been realized to store the apparatus constants and the calibration parameters: this KLOE database[5], based on the CERN utility HEPDB, has been extensively used to analyze real test data.

The Fermilab RBIO[s] package has been taken as a starting point to develop a tape manager facility, able to handle ANSI labeled tapes under UNIX. The package, officially supported only for SGI/IRIX and IBM/AIX, has been ported by us to HP-UX and Digital UNIX, made capable to deal with DAT and DLT drives, and new routines have been added for ANSI labelling tapes, volume handling and scanning.

1 million events have been Monte Carlo generated, staged on disk, and then stored on DLT tapes using a 7 layer loader (DEC877) attached to a DEC 2100 4/275, running Digital UNIX. RBIO modified library has been used to write tapes and a SCSI access library to mount and switch between tapes in a randomized way, talking to the loader via CAM (Common Access Method) control blocks[9].

With the adoption of the FATMEN package developed at CERN, all data stored in the DLT loader could be accessed by the user in a location independent way. A staging system is under development to serve files on disk. Also SHIFT is under examination.

6 Conclusions.

The DLT technology seems to be the most suitable for KLOE, due to its performance in terms of capacity and throughput, low error rate, high on-shelf life, and especially due to low cost of drives. Furthermore, it is proposed by many computer vendors: a good technical support and performance enhancement are espected. Our mini-production of one million KLOE events and its success has proven that the choices made up to this point work correctly for our goals. We still have to investigate more about the software adopted and under development for our tape management and staging system. Some more practice is also needed with the FATMEN package. Some concerns still hold us in declaring it the product of choice: in particular the support that CERN promises to give in the next years.

References.

1. The KLOE collaboration, *KLOE, a general purpose detector for DA@ NE,* LNF-92/019 *(1992)*
   The KLOE collaboration, *The KLOE detector, Technical Proposal,* LNF-93/002 (1993)
2. The KLOE collaboration, *The KLOE Data Acquisition System,* LNF-95/014 *(1995)*

3. **M.L. Ferrer,** *The KLOE DAQ system and DAQ issues for experiments in the year 2000,* plenary talk in this conference
4. D. Quarrie, B. Troemel, *YBOS programmers Reference Manual,* CDF Note 4.00, 18 Jan 1994
5. F. Pelucchi, *Database user and reference manual,* KLOE Note 99
6. Application Software Group, *FATMEN,* CERN Program Library Long Write-ups Q123
7. M. Shapiro, et al., *A beginner's* **guide** *to A_C and B-J,* CDF Note 384, Apr 1994
8. C. Debaun, S. Rojack, *RBIO User and Reference Manual,* FNAL Computing Division Document
9. program by ED Bailey, National Institue Angel Park, NC