DØ note 3435
CDF/DOC/COMP_UPG/PUBLIC/4560

Serial Media Working Group First Run II Report

Jon Bakken, Keith Chadwick, Mike Diesburg, Don Holmgren, Stephan Lammel, Gene Oleynik, Don Petravick, Robert Schamberger, Dane Skow, and Rick Thies

Mar 31st, 1998

This note presents the first report of the serial media working group for Run II. The working group has researched the tape market including current and next generation products. Several currently available tape technologies have been tested. Next generation technologies that are expected to become available before the start of Run II will offer significant cost savings (up to a factor of two). The group thus recommends to defer the selection of the Run II mass storage technology at this time. The acquisition of first tape library systems is scheduled for this fiscal year. The tape libraries should be flexible and handle next generation tape technologies.

1 Introduction

The next collider run of the Tevatron, Run II, is scheduled for spring of 2000. Both experiments, the Collider Detector at Fermilab (CDF) and the DØ experiment are being upgraded. The experiments plan to record between 500 TByte and 1 PByte of data during the 2 year duration of the run.

In Run I both experiments used Exabyte 8200 and 8500 tape technologies for data storage. Each experiment recorded over 50 TByte of data on these 8mm tapes. While the technology was not the most reliable one, both drive and media costs were rather low. A solution based on Exabyte 8500 technology is considered to be not robust enough and not desirable for Run II. The experiments are interested to use a more reliable, higher performance, and more modern tape technology.

2 Report of the Working Group

A serial media working group was formed in autumn of 1997 to research and evaluate tape technologies for bulk data storage in Fermilab Run II. Start date for Run II is beginning of 2000. Two collider experiments, CDF and DØ, expect to record between 500 TBytes and 1 PByte of data in the two year duration. The charge to the group includes both data storage in the Feynman Computer Center at Fermilab but also data transport off site. The Computing Division ask for a first recommendation by April of 1998.

We have researched the tape market including current and next generation products. We have evaluated different technologies during the past months (and are still engaged in long term reliability tests). There are several solutions based on current tape technologies that we believe could satisfy the estimated Run II needs. Those solutions are in a competitive price range. Several next generation tape technologies are scheduled to become available between 9 month to 2 years from now. A solution based on any of these next generation products would result in a significant cost savings of more than a factor of two, i.e. several million dollars.

We recommend to continue the stress test on the current generation of such products, to acquire evaluation units of the next generation products as soon as they become available, and to defer the decision on the Run II serial media technology for one year. We recommend Fermilab to acquire robotic tape libraries for CDF and D0 Run II this year which will allow us the greatest flexibility in the tape technology selection next year. That is the tape libraries should handle as many media formats as possible that have attractive next generation technologies scheduled for availability before Run II. A small number of tape drives and media (for the most promising tape technology) should be acquired for prototyping work and the vertical slice test of the experiments.

Total system cost including media, drive, library, and maintenance cost is a dominant factor in the serial media decision. In our recommendation we assume that none of the current tape technologies can be acquired for a price that is competitive with next generation technology (and standard discount). In the case of equivalent price, our preference would be for the technology with higher capacity media and a multimedia capable tape library system.

3 Technology Report

Table 1 and 2 show a comparison of some key parameters for each tape technology. The tables also list the costs for a full Run II system based on this technology. For this estimate we assumed a total data volume of 500 TBytes for DØ and 1 PByte for CDF. The required number of tape drives is calculated from the DØ estimate of 100 MB/s streaming plus 60 MB/s random file pick and CDF estimate of 100 MB/s streaming plus 40 MB/s random. A file size of 1 GByte is assumed. All Bytes are in powers of 2 (not 10), i.e. 1 kByte is 1024 Bytes. All tape technologies in table 1 and 2 are currently available. The tables are broken up into data center type and commodity technologies.

technology vendor	Redwood StorageTek	DD-2 Ampex	3590 IBM	DTF Sony
media capacity	46.6 GB	46.6 GB	10 GB	~42 GB
bandwidth (streaming)	10.4 MB/s	14.2 MB/s	8.2 MB/s	~12 MB/s
bandwidth (random)	4.3 MB/s	7.9 MB/s	4.8 MB/s	~5.8 MB/s
cost (1 media)	$78	$72	$54	~$80
cost (1 drive)	$80,000	$80,000	$30,000	$30,000
media for 1.5 PB	33,752	33,752	157,286	37,449
media cost	$2,633k	$2,430k	$8,493k	$2,996k
drives for 200 MB/s (streaming)	20	15	25	17
drives for 100 MB/s (random)	24	13	21	18
drive cost	$3,520k	$2,240k	$1,380k	$1,050
total cost	$6,153k	$4,670k	$9,873k	$4,046k

Table 1: Table of currently available data center type tape technologies. Performance numbers are based on vendor specifications and for a 1 GByte filesize.

The tables send several clear messages:

- For most technology the total costs are dominated by the media costs.
- For a 1.5 PByte system the total costs for a solution based on data center type equipment and commodity tape technology is in the same range.

technology	Eliant	DLT7000	EXB-8900	AIT-1
vendor	Exabyte	Quantum	Exabyte	Sony
media capacity	6.5 GB	32.6 GB	20 GB	25 GB
bandwidth (streaming)	1.0 MB/s	4.9 MB/s	3.0 MB/s	3.0 MB/s
bandwidth (random)	0.9 MB/s	2.0 MB/s	2.1 MB/s	2.7 MB/s
cost (1 medium)	$5.4	$80	$72	$72
cost (1 drive)	$1,700	$5,500	$3,500	$3,000
media for 1.5 PB	241,979	48,247	78,643	62,914
media cost	$1,307k	$3,860k	$5,662k	$4,530k
drives for 200 MB/s (streaming)	200	41	67	67
drives for 100 MB/s (random)	112	50	48	38
drive cost	$530k	$501k	$403k	$315k
total cost	$1,837k	$4,361k	$6,065k	$4,845k

Table 2: Table of currently available commodity type tape technologies. Performance numbers are based on vendor specifications and for a 1 GByte filesize.

- Without folding in any maintenance and tape library costs Exabyte Eliant 8500 technology would still be the most cost effective solution.

Data center technology solutions have tape drives that are more than one order of magnitude more expensive than commodity drives. The price per media unit is about the same. The competitive edge of data center type technologies comes from the larger media capacity and the cost savings of requiring 20,000 to 30,000 fewer tapes. Conclusion: Commodity technology with larger medium capacity would result in significant savings and be the desired solution for Run II.

The media capacity and costs are the key issue for Run II serial media. Smaller capacity medium results not only in higher media costs but also in higher tape library costs. Smaller capacity medium requires more slots or shelves and thus larger tape libraries (to hold the same amount of data). The total data size of the experiments was estimated carefully. The I/O bandwidth requirements, however, have large uncertainties. The well known data acquisition and production rates make only about 25% of the total I/O requirements. A solution with high costs to increase I/O bandwidth is then not desirable.

StorageTek Redwood is a helical scan half inch technology. The technology is available since about 1995. Due to the cartridge style random file reads result in a very low aggregate throughput of the drive. The tape drives are one of the most expensive ones. We expect no major upgrades to the Redwood product line but consider it a dead-end technology.

Ampex DD-2 is a helical scan 19mm technology. The cassettes are rather large. Ampex

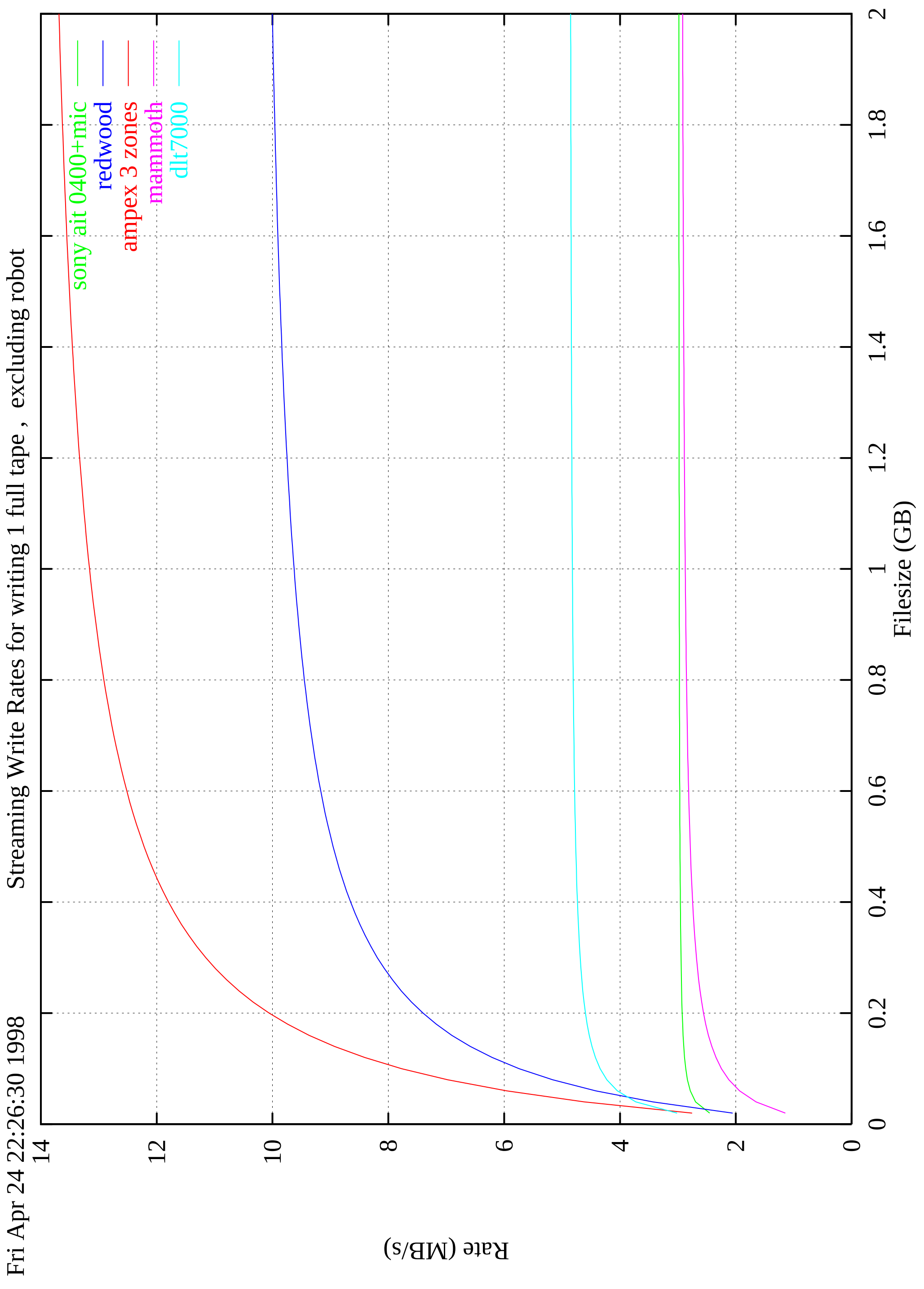

Figure 1: Plot of the effective I/O bandwith when writing a full tape of files as function of filesize (based on vendor specifications).

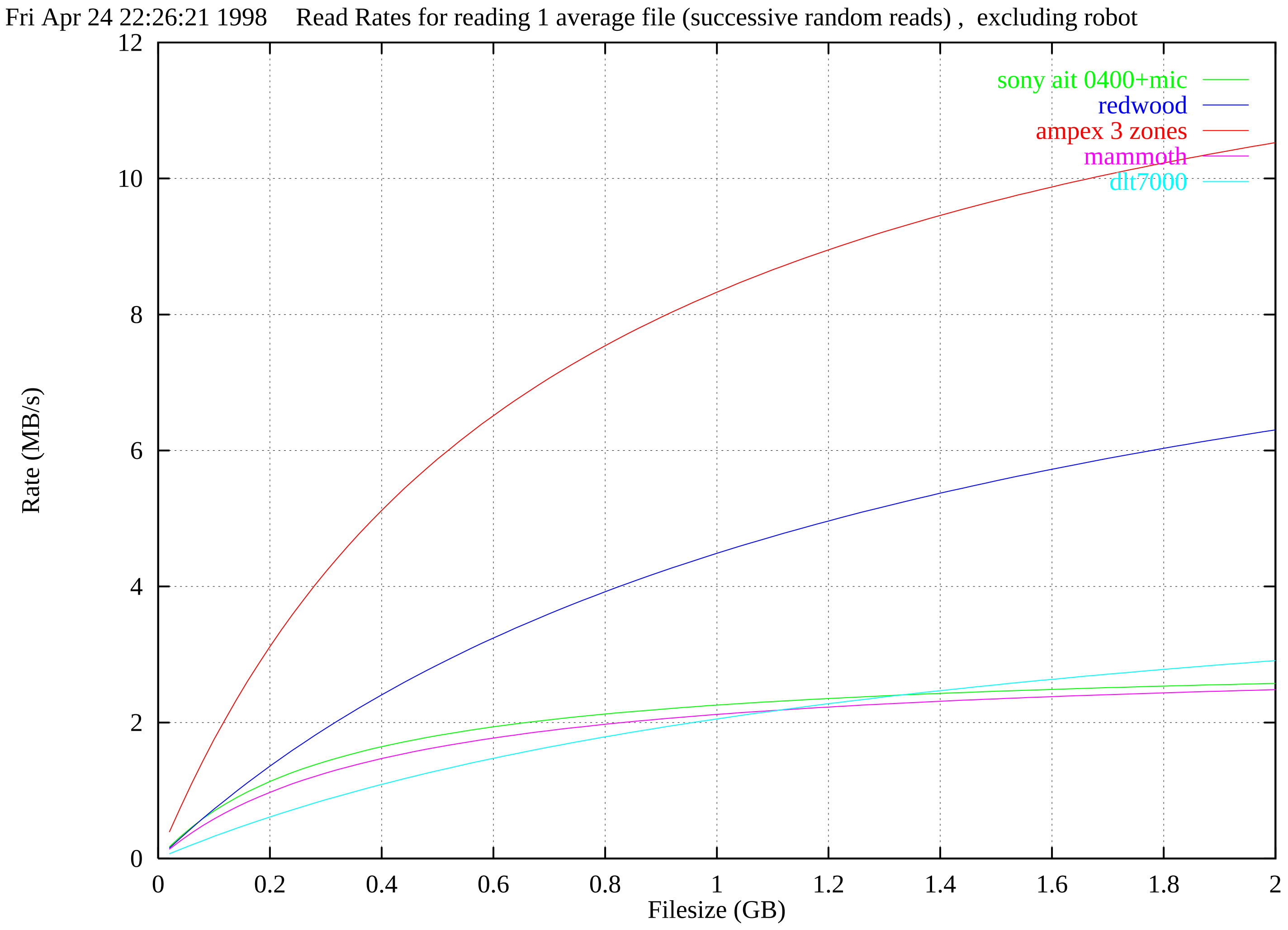

Figure 2: Plot of the effective I/O bandwith when reading randomly one file from a tape as function of filesize (based on vendor specifications).

plans to continue the product line with capacity and bandwidth upgrades. Today 330 GByte cassettes are available. However, those cassettes are huge and difficult to handle in a tape library. We have thus excluded them here. Tape drives are rather expensive but have the highest throughput. The media can be partitioned and unloaded/loaded at each partition boundary in addition to begin-of-tape (BOT), resulting in a high bandwidth for random file reads.

IBM 3590 is a linear half inch technology. The cartridges have rather small capacity of 10 GBytes. IBM plans to upgrade the line with 20 GByte cartridges in the near future. Even for a 20 GByte cartridge the media costs would still be very large.

Sony DTF is a helical scan half inch technology. The cassettes are rather large. Smaller cassettes are available but limited to 12 GBytes. Sony has a tape library that can handle the large cassettes. We did not test DTF drives at Fermilab.

Exabyte Eliant-820 is a helical scan 8mm technology. It is equivalent to Exabyte 8500 technology but with a doubled throughput of 1 MB/s. The media has very low capacity, 6.5 GB, but is extremely cheap. We did not make any large scale stress tests but expect reliability to be similar to EXB-8505 drives.

Quantum DLT7000 is a linear serpentine half inch technology. DLT cartridges are handled in most tape libraries that handle 3480 style half inch cartridges. The technology is available since 1997. It is the third generation of the DLTx000 product line. We expect Quantum to continue the product line with capacity and bandwidth improvements. Drives did not well in our stress test, failing with both mechanical problems and degraded capacity [2].

Exabyte 8900, mammoth, is a helical scan 8mm technology. The drives are backward read compatible with 8500 tapes. The technology is available since 1996. Upgrades to the product line have been delayed several times. Exabyte plans now an upgrade in middle of 1999 to 70 GB capacity tapes and 12 MB/s drives. In our stress tests drives performed significantly better than 8500 type drives [1]. According to Exabyte the failure mode that we encountered has been resolved with an engineering modification. We have not re-tested newer version drives.

Sony AIT-1 is a helical scan 8mm technology. The technology is available since 1997. It is a rather young product. The cassettes contain a small, 16k, chip that holds tape history and inventory information. The drive has a very short load time and can unload/load at positions other then BOT. This results in very quick file access for random file reads. Stress test of the technology are in progress [3].

Table 3 shows the costs of a tape library holding 500 TByte for each of the current tape technologies. The 500 TByte library capacity is the minimum estimate of the CDF and DØexperiment for Run II. The cost of robotic storage ranges from half a million to three million dollars (excluding the Exabyte Eliant 6.5 GB cartridge scenario). This is in agreement with initial cost estimates of the experiments and the Computing Division of about two million dollars.

technology	vendor	library vendor	capacity one library unit	cost one library unit	cost for 0.5 PB
Redwood	StorageTek	StorageTek	279.6 TB	$260k	$526k
DD-2	Ampex	EMASS	62.6 TB	$350k	$2,913k
3590	IBM	IBM	4.0 TB	$23k	$3,394k
DTF	Sony	EMASS	75.3 TB	$350k	$2,430k
Eliant	Exabyte	EMASS	33.7 TB	$350k	$5,368k
DLT7000	Quantum	EMASS	133.5 TB	$350k	$1,392k
EXB-8900	Exabyte	EMASS	103.7 TB	$350k	$1,778k
AIT-1	Sony	EMASS	129.6 TB	$350k	$1,433k

Table 3: Table of estimated tape library costs for current tape technology. Library costs are not rounded to integral number of library units. A fixed cost of $50k is included for all libraries but the IBM library where it is applied for each 15 library units.

In the table we show only one library option for each tape technology. We have choosen the most natural combination (same library and tape drive vendor where possible). The EMASS libraries are multimedia libraries and can also handle IBM 3590 and StorageTek Redwood media and drives. The costs in the table are systematically underestimated as it is based on a non-integral number of libraries. We have not included extra libraries to provide sufficient tape mount capacity per tape drive. This is especially important and thus a significant underestimate for technology with small media capacity like the Exabyte Eliant and IBM 3590.

4 Future Technologies

Most of our information on future technologies and products are confidential and obtained in non-disclosure agreements. In this section we want to broaden the tape view a bit to exchangable media. We also show the roadmap for a few tape technologies where it is publically available.

Writable CD-Roms are at about 70 cents for a 650 MByte disk. To store 1.5 PByte 2.5 million CD-Roms would be required. The media cost, however, with about $1.7M is equivalent to tape media costs. DVDs hold several Gbyte of data. Writable DVDs were expected to be available end of last year. A 20 to 30 GByte DVD would be most interesting for storing reconstructed data of Run II.

TeraStor promises to revolutionize the data storage industry with a new category of re-writable mass storage technology: Near Field Recording. The idea is to combine tape-like capacities with hard disk-like performance at a price of about a $ per GByte. Although no products are available to date, 20 GByte capacity per surface media and drives are scheduled for this year, 1998. This technology could be even more interesting for Run II than DVDs.

Exabyte plans to boost the capacity of mammoth to 70 GByte by mid 1999 and to 200 GByte per cassette by the year 2000. The transfer rates will increase to 12 and 20 MB/s.

Quantum's DLT7000 utilizes four parallel read/write channels to achive a transfer rate of 5 MB/s. Quantum considers eight, 12, even 16 parallel channels feasible, with a potential corresponding improvement in throughput. During the last years Quantum has brought a new generation of DLT technology to market every 12 to 18 months. According to Quantum 100 GB capacity cartridges are on the horizon.

Hewlett-Packard, IBM, and Seagate have formed an alliance to develop and market Linear Tape-Open (LTO) technology. Two formats, Ultrium and Accelis, are designed for highest capacity and fastest access. The Accelis roadmap is expected to result in products with access times of less than 7 sec, capacity of up to 200 GByte, and transfer rates ranging from 80 to 160 MB/s. Ultrium's four generation roadmap shows products with up to 800 GByte capacity per cartridge.

availability	technology	vendor	media capacity	band-width	cost (1 media)	cost (1 drive)
H2.1998	NTP-2	IBM	20 GB	>10 MB/s	$54	$30,000
Q4.1998	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$x,xxx
Q4.1998	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$xx,xxx
H1.1999	superDLT	Quantum	100 GB	>10 MB/s	$80	$5,500
mid 1999	mammoth-2	Exabyte	70 GB	12 MB/s	$72	$3,500
1999 (guess)	Ultrium-1	LTO	100 GB	>10 MB/s	$80	$30,000
Q4.1999	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$xx,xxx
Q1.2000	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$x,xxx
Q1.2000	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$xx,xxx
2000	mammoth-3	Exabyte	200 GB	20 MB/s	$72	$3,500
Q4.2000	XXXXX	XXXXXXXXX	xx GB	xx MB/s	$xx	$xx,xxx

Table 4: Table of future tape technologies. Prices are estimates based on information form the vendors where available.

If we acquire tape libraries in this fiscal year, then the capacity of the library system will effectively increase (in terms of GBytes—not in terms of media slots, of course) for next generation tape technology. Column six in table 6 we can then translate into the amount of tape library storage available for a fixed budget. If we take the budget at the level estimated in the joined CDF/DØ/CD hardware needs assesment document then for most of the next

availability	technology	vendor	media cost	drive cost	total cost
H2.1998	NTP-2	IBM	$4,147k	$1,590k	$5,737k
Q4.1998	XXXXX	XXXXXXXXX	$x,xxxk	$xxxk	$2,407k
Q4.1998	XXXXX	XXXXXXXXX	$x,xxxk	$xxxk	$6,749k
H1.1999	superDLT	Quantum	$1,228k	$248k	$1,476k
mid 1999	mammoth-2	Exabyte	$1,580k	$130k	$1,710k
1999 (guess)	Ultrium-1	LTO	$1,229k	$1,200k	$2,429k
Q4.1999	XXXXX	XXXXXXXXX	$x,xxxk	$x,xxxk	$4,196k
Q1.2000	XXXXX	XXXXXXXXX	$x,xxxk	$xxxk	$1,211k
Q1.2000	XXXXX	XXXXXXXXX	$x,xxxk	$xxxk	$3,301k
2000	mammoth-3	Exabyte	$553k	$70k	$623k
Q4.2000	XXXXX	XXXXXXXXX	$x,xxxk	$x,xxxk	$2,548k

Table 5: Cost estimate of future tape technologies.

technology	vendor	library vendor	capacity one library unit	cost one library unit	cost for 0.5 PB
NTP-2	IBM	IBM	8.0 TB	$23k	$1,722k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$741k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$937k
superDLT	Quantum	EMASS	409.6 TB	$350k	$488k
mammoth-2	Exabyte	EMASS	414.7 TB	$350k	$482k
Ultrium-1	LTO	IBM	40.0 TB	$23k	$344k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$494k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$396k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$494k
mammoth-3	Exabyte	EMASS	1,037 TB	$350k	$223k
XXXXX	XXXXXXXXX	XXXXXXXXX	xxx.x TB	$xxxk	$272k

Table 6: Table of estimated tape library costs for future tape technology. Library costs are not rounded to integral number of library units. A fixed cost of $50k is included for all libraries but the IBM library where it is applied for each 15 library units.

generation technologies the majority of the Run II data could be kept inside a tape library. This would eliminate the need for operator assisted tape mounts in the computer center and makes defering the serial media selection even more attractive.

A Charge to the Working Group

```
To investigate new data recording technologies and to identify options that
are suitable for use in various parts of our data acquisition, processing,
and analysis (including onsite and offsite analysis and desktop and
workgroup-based analysis) for Run 2.

More specifically:

    1.To think through the data recording requirements of CDF and D0 for
       Run 2 from the data acquisition stage to the point where data is
       available to physicists for analysis. This will necessarily include
       forming a model or models for the use of possibly different
       technologies at different stages of data processing - from data
       acquisition through to analysis, on-site and off-site.
       Consideration should be given to the potential for using robotics
       to reduce the number of manual tape mounts and perhaps the number
       of independent processing stages. Your considerations should include:
              Recording modes
              Software development and support on platforms of interest
              Medium cost per byte
              Cost of hardware (drives, interfaces, etc.)
              Reading and writing speed
              Physical volume of storage medium
              Medium reliability and long term stability -- including
              susceptibility to environmental damage
              Drive reliability and longevity -- cleaning and maintenance costs
    2.To make a broad survey of existing equipment and future plans of
       vendors in order to identify potentially interesting technologies
       and approaches. This will require examining technical specifications
       and reliability data and collection of information from users of the
       systems and testing organizations if possible.
    3.For equipment that is available (either in production or in development)
       to set up pilot projects to test the systems under conditions that are
       representative of the use patterns that we expect. For systems that you
       determine are likely to meet our performance requirements, to perform
```

tests that really stress the reliability of the devices over reasonably long periods of time. The results of these studies should be summarized in reports and/or technical notes.

4. To periodically report on the status and progress of your work and to make recommendations to the Division and the Run 2 committee on the suitability of various technologies to Run 2.

As a 'guideline', I suggest the following schedule:

Pilot projects: As aggressively as you deem necessary

First report on survey and initial pilots January, 1998

First draft of requirements Mar, 1998 and Recommendations for FY98 HSM procurements

Summary of activities and May 1998 findings

Additional Recommendations Summer 1998 (phased)

It should be clear that pilot projects are to be implemented through our CD departments who will be instructed to help and, where possible, in collaboration with interested users.

Since decisions in this area can affect the way non-Fermilab people do their work at the lab and at their home institutions, we are asking physicists from CDF and D0 who are outside of FNAL to participate in the working group. I would suggest that you go farther than that and make an effort to keep the 'public' well-informed concerning your work. This can be accomplished by having good notes from each meeting and making them readily available, perhaps by putting them on the web and by occasionally summarizing your status in Newsletter articles. the Run II committee will ask for periodic status reports.

Joel Butler

References

[1] Exabyte mammoth test report, in preparation.

[2] Quantum DLT7000 test report, to be written...

[3] Sony AIT-1 test report, to be written...