IBM Magstar 3494 Virtual Tape Server Performance White Paper

Version 4.0

IBM Corporation

28 July 2000

1. Introduction

This paper provides performance information on the IBM Magstar Virtual Tape Server (VTS), the 3494 model B18, in a fourth general availability of the VTS with enhanced performance features. It is intended for use by IBM field personnel and their customers in designing tape server solutions for their applications. The VTS, model B18, has been available beginning August 1998. Since that time it has been enhanced with a number of optional features including the *Extended High Performance Option* (EHPO) which includes data compression, the *SCSI Host Attachment* feature, the data *Import/Export* feature, the *Extended Performance ESCON Channels*, and the *Performance Accelerator* feature (PAF). The current release of the VTS, model B18, includes extended host bandwidth and connectivity with up to eight ESCON channels and enhancements in tape volume cache (TVC) capacity and performance. With these features, the present VTS has a host data throughput bandwidth up to fourteen times that of the original model B16 in 1997.

Concurrent with this new release of the VTS, model B18, is a Peer-to-Peer VTS version which provides automatic dual copy of tape data, locally or at separate locations, together with higher availability of data in the event of the failure at one of the locations. Its performance is described in a separate manuscript: *IBM Magstar 3494 Peer-to-Peer Virtual Tape Server Performance White Paper*, Version 1.0, 28 July 2000.

2. Product Description (VTS, model B18)

Figure 1 shows the physical configuration of the VTS, model B18. The left hand frame includes the controller and tape cache (disk storage), while the right hand frames constitute the IBM 3494 library component of the system, including up to six 3590 tape drives.

Figure 2 shows a schematic of the principal functional units in the VTS, model B18. In this description we focus as a default on the fully configured B18, including the PAF and all the prerequisites for the PAF. The host connectivity is provided by two or four adapters, each providing two channels, either ESCON or SCSI. Up to two of the channel adapters can be SCSI Host Attachment Features (FC #3422), as shown in Fig. 2, to provide up to four SCSI host attachments to the B18. Additional Extended Performance ESCON channels

Fig. 1. The Virtual Tape Server is shown on the left together with the associated IBM 3494 Tape Library on the right.

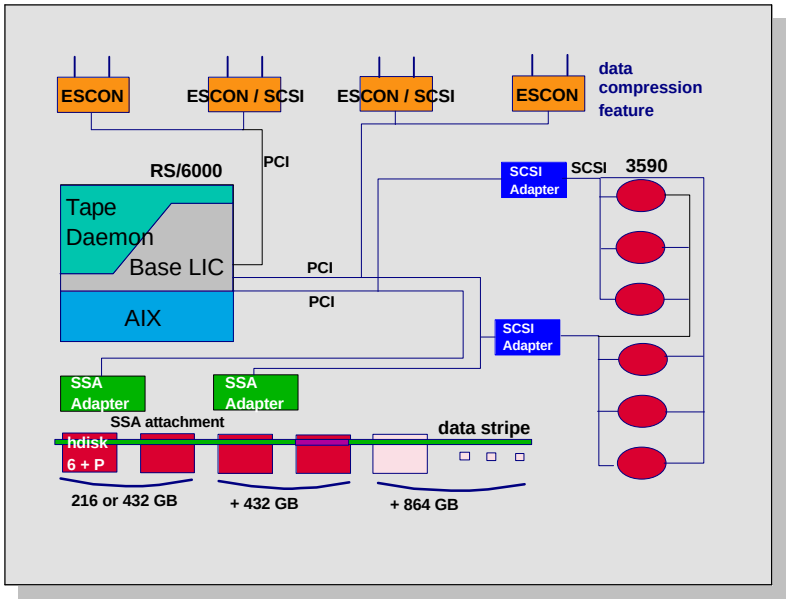

Fig. 2. A schematic diagram of the VTS architecture, shown here with the HDD configuration for the Performance Accelerator Feature (PAF).

are provided by FC #3412. The data compression feature allows data compression through the ESCON adapters and the SCSI host adapters. With EHPO all data within the system below the attachment adapters is in compressed form. All data passes through, and is managed, by the logical software components indicated in the RS/6000 logical block. Incoming data is then stored in TVC (HDDs) that is attached through two SSA adapters. The TVC is available in capacities of 216, 432, 864, and 1728 GB (physical). The HDDs are organized in RAID-5 units of 6 + parity + spare (2 + P + S for configurations without the PAF). Data is striped across all the RAID units as shown in Fig. 2. The primary unit of data handled is a full 6+P stride, comprising a 64 KB logical track from each of the seven HDDs in a parity group (incl. parity).

One of the significant features of the virtual tape server is that scratch mounts do not require the physical mounting of a tape cartridge. The tape mount request is handled at electronic speed and data transfer can begin immediately into the TVC. Once a complete tape volume is received in TVC, it is scheduled for copying to physical tape under the control of licensed internal code. The 3590 tape drives are part of the IBM 3494 tape library. Without the data compression feature, data is compressed at the 3590 and stacked as multiple compressed volumes per physical tape until the tape is full. With the data compression feature, data is compressed at the host adapter. With the current release of the B18, the customer has a choice of 3590 model B or the new model E tape drives. Relative to the model B, the model E drives provide up to a 50% increase in native data rate (14 MB/s) and a doubled cartridge capacity (20 GB, native). On the VTS, the model E characteristics significantly reduce tape mount activity associated with copying data to the 3590.

Host read requests are satisfied at either the tape volume cache (read hits), or by first recalling the data from 3590 physical tape to the tape volume cache (read misses).

The base model B18 VTS presents itself to the host as two fully configured 3490E tape subsystems with 16 tape drives each (32 virtual drives). With the EHPO feature and a minimum of 144 GB of cache, 64 virtual drives are seen as four control units with 16 tape drives each. The optional larger TVC capacities can improve throughput performance directly through providing read hits and also provide for increased time at peak performance during write intensive periods. They are also designed to improve the effective bandwidth of the VTS by increasing the bandwidth of the TVC.

The VTS configurations include a choice in the number of 3590 tape drives from three to six (four to six with the Performance Accelerator). A larger number of physical 3590 tape drives improves performance during periods of high activity, especially on specific mount requests (read misses).

Relative to the EHPO feature, the PAF (FC #5236), can provide an additional significant boost in data throughput rate as viewed from the host; up to 100% for sustained write throughput at a data compressibility of three, for example. The PAF performance gain is achieved through a combination of hardware (increasing the number of processors in the VTS's RS/6000 from two to four) and software (buffering logically contiguous data so that it can be written in full strides to the tape cache).

3. Performance Metrics

In the following sections we discuss the performance of the VTS. The primary metric we use is the *sustained write* rate in MB/s. The opposite of this rate is the *read miss* (or *read recall*) rate, which closely mirrors the resources used by the sustained write. We use these metrics, especially the

sustained write, as a way of succinctly characterizing VTS performance. The *sustained write* rate can only be obtained after the system has been run at least once on the order of several days to assure that a steady state has been reached between the tape cache and the I/O and copying activity. The *sustained write* rate measures the collaborative overlapped performance of the host channel input, the RS/6000 processor and its internal paths, the file system, the disk adapters, the HDDs, and the 3590 tape drives, and it does it under circumstances where the system is managing a large number of volumes, with at least sixteen of them active. This is the net performance value of the VTS to the customer.

All of the measurements in Sections 3.1 through 3.4 were made with four ESCON channels on a VTS (GA Sep 99) and are repeated here with updated cache capacities (see Appendix A) to account for the higher capacity HDDs. No other significant difference in base or EHPO performance is expected between that and the current VTS.

We also report *peak write/fill* rates and *read* hit rates because they are an integral part of VTS performance, and since they are higher than the sustained rates, can be used to enhance overall average system throughput when they are taken advantage of appropriately.

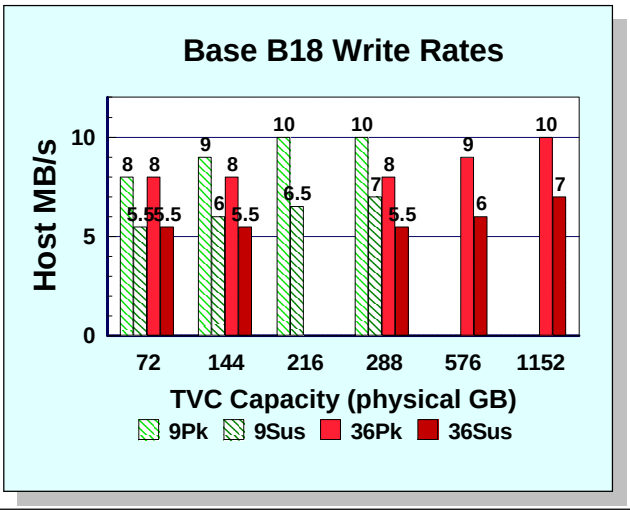

Fig. 3. The peak and sustained write throughput of the base VTS at the available TVC capacities. The configurations are with 9 or 36 GB HDDs. Pk *and* Sus *refer to* peak *or* sustained *write throughput, respectively.*

3.1 Base B18 Write Performance

Write performance quoted here has been derived from measurements that simulate typical user environments; namely sixteen applications writing 800 MB tape volumes simultaneously. The block size used is 32 KB and the BUFNO parameter is set at 20. The measurements quoted in these sections apply to a VTS without the EHPO or PA options. We describe the effect of the optional performance enhancement features in Sec. 3.5.

The tape volume cache is allowed to fill and the system is then operated for some time in a sustained mode with new data being added to the tape cache while older data in the tape cache are being copied to physical tape. We then characterize write performance in two ways: the "*peak write/fill*" rate and the "*sustained write*" rate as described below.

3.1.1 Peak Write/Fill Rate

The peak write/fill rate is defined for a tape server that has been in long term use, has most of its new data copied to tape, and now is subject to an incoming rate of writes that saturates its throughput. (Following the copying of data from tape cache to physical tape, the copy of the data in tape cache remains, providing for possible read hits in tape volume cache while making the space the logical volumes occupy immediately available for new writes, if necessary.) Under these conditions, the predominant activity of the virtual tape server is to place the incoming data into tape volume cache

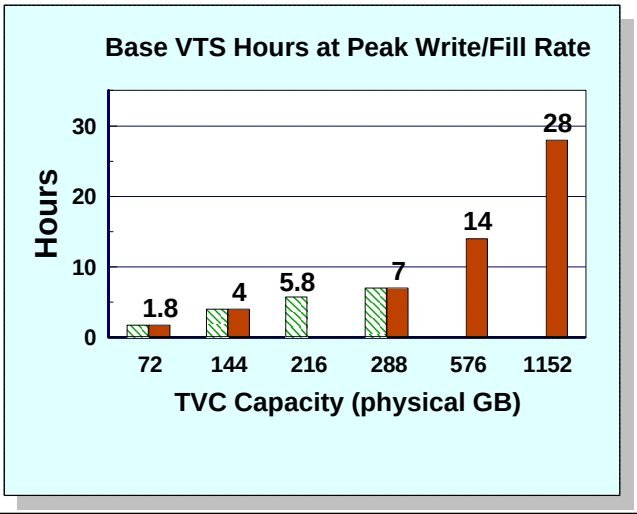

Fig. 4. The maximum time at peak write rate for various TVC capacities. It is assumed that the TVC has no un-migrated data at the time the write activity begins. The (striped) green bars represent TVC capacities available with 9 GB HDDs; the (solid) red bars are available 36 GB HDD TVC capacities.

(the HDDs). Until a fill threshold of the tape volume cache is reached there is little copy activity of the data from tape volume cache to physical tape. This peak write rate (Peak Write/Fill) for the virtual tape server is shown in Fig. 3 as a function of the capacity of the tape volume cache. The peak write rate ceiling is between 8 and 10 MB per second, varying with the size of the cache. The peak write rate performance can be used to ride through periods of heavy write activity if all of the TVC has been previously copied to physical tape. The approximate hours of peak write/fill rate possible with different tape cache sizes is shown in Fig. 4 for the base B18 model and with the EHPO feature. For the optional PAF feature, the hours at *peak write* are discussed in Sec. 4 (cf., Fig. 10).

3.1.2 Sustained Write Rate

The *sustained write* rate is defined for a tape server that has been in long term equilibrium operation with the rate of incoming data equaling the rate at which data is being copied from tape volume cache to physical tape. Thus for sustained write there are three separate I/O activities going on simultaneously. New data is being written to the tape volume cache, and aged data is being read from the tape cache and then written to the tape library. Because of this additional activity the sustained write rate is lower than the *peak write/fill* rate. The sustained write rate varies with tape volume cache size between 5.5 and 7 MB per second (cf., Fig 3).

3.2 Base B18 Read Performance

As with the write performance data above, we quote the projected read throughput performance for a particular operation operating exclusively. As on the write data, the data quoted here have been obtained with 800 MB volumes, 32 KB blocking of data, and BUFNO=20.

3.2.1 Read Hit (Read from Tape Volume Cache)

A *read hit* is a host read request which finds the requested volume in tape volume cache (the HDDs), whether the data has been copied to physical tape or not. Thus the data transfer to the host involves mostly locating the data on the HDDs, staging them to the VTS CPU memory, and then transferring them to the channel adapter(s). The read hit throughput in MB/s, assuming several concurrent applications issuing overlapping read requests, is shown in Fig. 5.

3.2.2 Read Miss (Read from Tape Library)

A *read miss* (also referred to as *read recall*) is a host generated request to read a tape volume which has been copied to physical tape, and is no longer available in tape volume cache. Thus a read miss requires recalling of the requested tape volume from the tape library to tape volume cache (*mount time*), the transfer from tape volume cache into VTS system memory, and then transfer from memory to the channel adapter. This process is approximately the reverse of the *sustained write* operation. The read miss throughput in the limit of several such overlapping operations operating exclusively is shown in Fig. 5.

With respect to read misses, it is necessary

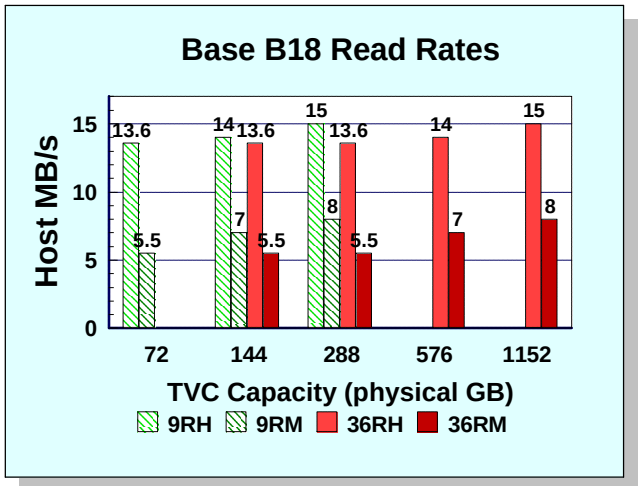

Fig. 5. Read rates for the base VTS at the available TVC cache capacities. The configurations refer to 9 and 36 GB HDDs. RH and RM refer to read hit and read miss rates, respectively.

to observe that if these are the result of an application requesting the mounting of a large number of volumes nearly simultaneously, there will be a delay before all volumes are mounted. For example, if the VTS is configured with the maximum of six tape drives, and two of them are assigned to receiving of data being copied from tape volume cache, then only four recall mounts can be handled simultaneously. Additional mounts will be queued on those tape drives and will be handled sequentially. Thus average volume mount time statistics on the VTS may appear larger than expected (see also Sec. 3.4).

3.3 Read Backwards

Reads backwards refers to a read request sequence which accesses the tape data in the opposite sequence order from the one in which they were written; for example, if the application

requests the previous block of data from the one it just read back from tape. This is a very cumbersome, low performance type of tape data access on physical tape drives. This kind of tape access method is much more efficient on the VTS because the block IOs are actually executed on HDDs.

3.4 Volume Mount Response Time

The VTS has two characteristic mount response times; the time required for the VTS to respond with a "ready" after an host request to mount a tape volume. A *fast ready* is associated with the transfer of data which is already in tape cache and occurs in typically two to three seconds. A *normal ready* requires the mounting of a physical tape and transfer of the tape volume data involved to tape cache.

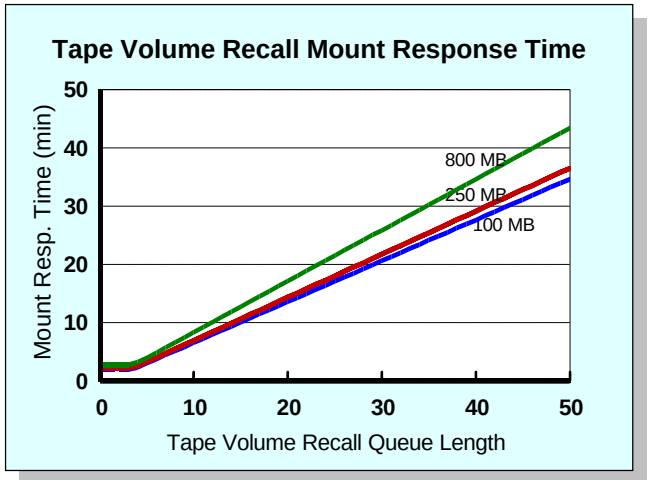

Fig. 6. Tape volume recall mount time as a function of the number of recalls in the mount queue.

*Specific mount write update*s to tape cache (the writing of an updated tape volume which is still in the VTS tape cache), *scratch mounts* (the writing of a volume which does not already exist on the VTS to a public scratch tape), and *read hits* have a characteristic response time of approximately two to three seconds (fast ready) unless the system is heavily occupied by other activity. These response times are much shorter than with conventional tape drives because they are done at electronic speeds.

The mount response time for *read misses* and *writes to specific mount* or *private scratch pool* volumes (if their data is not in tape cache) are longer because they require the recall of the requested tape volume to tape cache before a "ready" is presented to the host. The response times thus include the time to physically read an entire volume of the requested data from tape to tape volume cache. (The additional time to destage the data from VTS memory to tape cache (HDDs) mostly overlaps the tape data transfer.) A best case response time for such mounts is on the order of approximately two to three minutes, and would only happen when little else is active on the VTS. Figure 6 shows a theoretical estimate of tape volume recall mount response time as a function of the recall queue length at the time the recall request arrives. The estimate is based on the parameters exhibited in Eq. 3 (Sec. 7), and are shown for three average volume sizes (100, 250, and 800 MB), a data compressibility of three, and assumes four 3590B tape drives allocated to recall (the result with 3590E drives is approximately the same). The mount response times in Fig. 6 are meant as a guideline only. Actual response time can vary widely. It is affected by both other activity on the VTS and the performance enhancement features installed on the VTS.

3.5 Optional Performance Enhancement Features

In this section we describe the effects of the ESCON High Performance Option (EHPO) and Performance Accelerator (PAF) features on VTS performance. The PA feature has a prerequisite of the EHPO feature.

3.5.1 The EHPO Data Compression

The optional EHPO feature applies compression to data coming into the VTS and decompression to data leaving the VTS via the ESCON channels or SCSI ports. The principal benefits of compression are to increase the throughput performance of the VTS and increase the number of virtual volumes in tape cache available for cache hits. Depending on the nature of the workload, throughput performance is constrained by the ESCON adapters to the VTS, the RS/6000, the SSA device adapters (HDD attachment), the HDDs, or tape drives. For most realistic workloads, however, the throughput of the VTS with only the EHPO option (i.e., without the PAF) is determined by the size of the tape cache; i.e., the number of HDDs available. Then, if the tape cache I/Os involve compressed data, the number of I/Os required is reduced approximately by the compressibility of the data. This results in a significant improvement in the VTS throughput with data compressibility as is shown in Figs. 7 to 9. The throughput shown in the figures has been measured with eight ESCON channels. The curves

in the figures are based on selected measurements with a data compressibility factor, *CF*, up to four. These data have been extended to the curves shown using analytical performance modeling. Analytical expressions for the curves in Figs. 7 and 8 are given in Appendix B.

3.5.2 The Performance Accelerator Feature (PAF)

The Performance Accelerator feature is a direct solution to the tape cache throughput limitation described in the previous section. Although data compression increases the tape cache I/O bandwidth, it nevertheless remains the throughput limiting resource of the VTS in many workload situations. The PAF is a package of hardware and software enhancements which (in most instances) remove the tape cache bandwidth as a limiting factor in the throughput of the VTS. One principal element of the PAF is an increase of the RS/6000 processing capability from two to four processors. Another is the buffering of the incoming data to the RS/6000, whether from an host or tape, in logically contiguous blocks that can be written to the tape cache in a full stride (i.e., as a full parity group stripe which does not require parity data construction via read-before-write I/Os). This enhancement can increase the MB/s throughput of the tape cache to a level where it no longer is a throughput constraint for any practical VTS workload.

In addition to the above enhancements, the PAF performance benefits include those of the Extended High Performance Option (EHPO), which is a prerequisite for the PAF feature. The performance of the VTS with the PAF is also shown in Figs. 7,8, and 9.

3.5.3 Sustained Write Throughput with the Optional VTS Performance Features

As described above, sustained write performance is the VTS performance benchmark which exercises most of the VTS internal processes simultaneously. It is thus the single best performance metric by which one can judge the data throughput bandwidth capability of the VTS.

The EHPO curves in Fig. 7 are shown for the minimum and maximum available physical tape cache capacities and assume a configuration with a minimum of four ESCON channels. Their intercept at *CF* = 1 corresponds to the base VTS B18 throughput performance, in the range 5.5 to 7 host MB/s, depending on tape cache capacity. With the EHPO option this throughput is increased

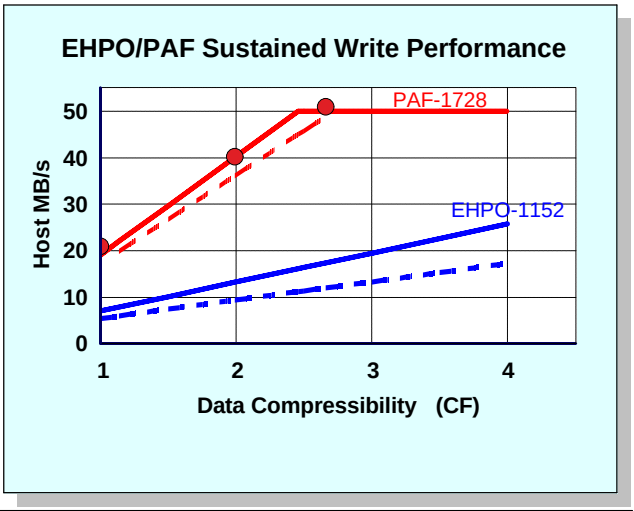

Fig. 7. Sustained Write throughput of the VTS with the PAF and EHPO as a function of data compressibility. The solid curves in each case show the throughput when configured with a 1728 GB TVC. Selected measurements are indicated by the solid circles. The dashed lower curve in each case represents the expected throughput when the VTS is configured with the smaller TVC capacities employing the 4(2+P) HDD configuration (see Appendix A).

approximately in proportion to the compressibility of the data being handled. For an average data compressibility of three the EHPO option with a 1728 GB TVC has a sustained write bandwidth of approximately 19.5 MB/s.

The sustained write throughput with the PAF feature can be more than doubled to 50 MB/s with respect to the EHPO feature alone at a working data compression ratio of three. At higher compression, the PAF throughput is not sensitive to compression factor. For smaller compression factors the throughput bandwidth is proportional to *CF*, starting from about 19 MB/s at *CF* = 1.

The PAF performance in Fig. 7 is shown for a tape cache capacity of 1728 GB. The performance with 216, 432, and 864 GB TVCs is approximately the same, but smaller TVC capacity appears to correlate with a somewhat reduced throughput. The dashed line for the PAF throughput represents an envelope which includes most measurements that have been made with TVC capacities less than 1728 GB. With PAF the principal advantage of the higher tape cache capacity is a higher read hit ratio which gives better response time performance and extends the tape drive limited read recall bandwidth. For data that compresses at less than a factor of three, the larger cache also permits

operation at peak write throughput for a longer period of time.

3.5.4 Peak Write Throughput with the Optional VTS Performance Features

Peak write throughput, as a benchmark, measures the bandwidth between the host adapters and the tape cache. Read hit bandwidth closely resembles this benchmark in function except that the data flow is in the opposite direction. In this manner of operating there is no data flow between the VTS and the physical tape drives. The principal observation of the peak manner of operating is that for limited times, depending on the TVC capacity and data compressibility, the VTS can accept data at a about 60% greater than the sustained write rate (cf., Figs. 7 and 8). With the PAF feature, the principal advantage occurs in being able to accept data at the maximum rate (about 80 MB/s) down to data compressibility as low as 2.0 and about 45 MB/s at a compressibility of one, for some time.

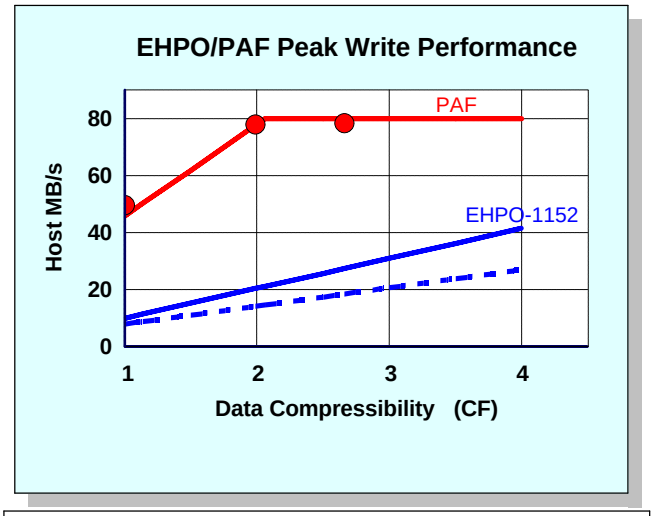

Fig. 8. Peak Write throughput of the VTS with PAF and EHPO as a function of data compressibility. The PAF curve is based on an average of measurements on all cache sizes. The 1728 GB TVC measurements, solid circles, are shown as an example. The EHPO expected performance is shown for a 1152 GB TVC, solid line, and for the smaller 4(2+P) HDD configurations (see Appendix A) by the dashed line.

Customers that require the maximum throughputs that are shown in Figs. 7 and 8 should configure the VTS with eight ESCON channels.

The read hit and read miss (read recall) rates with compressed data correspond approximately to the peak write and sustained write rates, respectively. The maximum read rate is usually lower than the corresponding write rate. This is due principally to the asymmetry in the details of the ESCON data transmission between the host and VTS for the two directions of data transfer.

3.5.5 Read Hit and Read Recall Performance with EHPO and PAF

Table 1. Independent Read Hit and Read Recall Rates for an 8-ESCON VTS with thePAF.

	CF	host MB/s
Read Hit	**1**	**56**
	2.66	**64**
Read Recall	**1**	**18**
	2.66	**38**

The maximum independent read hit and read recall rates are shown in Table 1. They are based on measurements with 64 active virtual tape drives reading 800 MB volumes. The TVC capacity does not appear to be a significant factor in determining read rates, as the measurement variability of the throughput was found to be of the same order as the variation with TVC capacity.

3.5.6 Typical Mix Workload Performance with EHPO and PAF

The sustained and peak write workloads are the easiest to measure and indeed constitute a comprehensive assessment of the VTS' throughput bandwidth when taken together with some auxiliary read hit and read miss (recall) measurements. However, most applications of the VTS have a workload that are a mix of writes and reads. It is useful, then, for addressing questions on what the expectations of a typical VTS application should be for performance to define a "typical mix" workload that will be representative of how a large fraction of the applications will use the VTS. Our definition of the typical mix workload is a 60/40 mix of writes/reads. The writes are all assumed to be scratch writes, and the reads are assumed to have a 50% hit ratio. It is assumed for modeling that this workload is in steady state operation on the VTS so that the actual host MB/s rates for the components reflect their proportions in the workload definition.

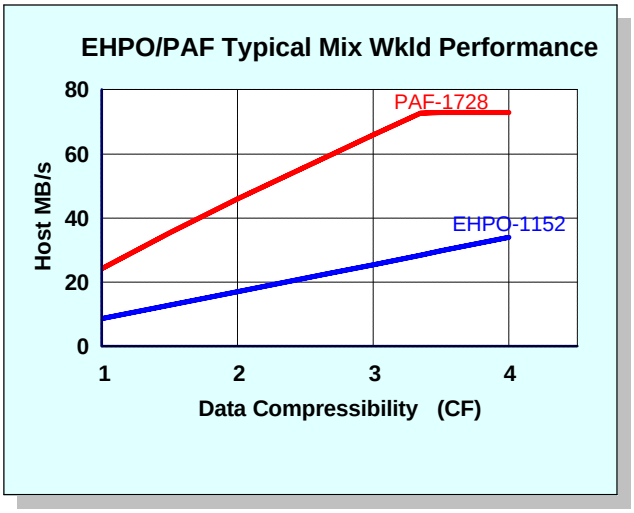

Fig. 9. The modeled Typical Mix (60W/40R, 50% read hit) throughput of the VTS with PAF and EHPO as a function of data compressibility. The TVC capacity is assumed to be 1728 GB; smaller caches will reduce the throughput somewhat in most cases (see Fig. 7, for example).

Figure 9 shows the total throughput of the modeled typical mix workload as viewed from the host.

3.5.7 Peak Write / Read Hit Maximum Throughput on Large Block Transfers

In this section we describe a measurement on the new VTS with eight ESCON channels and PAF using 64 KB block transfers. Most applications are limited to a block size of 32 KB which results from a current limit in the commonly used QSAM access method**.** This QSAM limit is assumed in the performance quoted elsewhere in this paper for OS/390 performance. Larger block sizes are available to applications that use some other tape drivers. Table 2 shows a comparison of peak write and read hit throughput with 64 KB blocking, compared with 32 KB blocking. The doubling of block size can improve the throughput for these functions by more than 25%.

Table 2. Maximum Peak Write and Read Hit (MB/s) throughput rates with 32 KB and 64 KB blocking for an 8-ESCON VTS with the PAF.*

	CF	32 KB	64 KB
Peak Write	2.66	**80**	**103**
Read Hit	2.66	**64**	**89**

*) BUFNO = 20

4. Time at Peak Write as a function of TVC Capacity

TVC capacity, above a minimum required to emulate sixty-four logical tape drives smoothly in a sustained mode, is principally a VTS performance enhancement. The performance improvements come in a number of ways:

1. With a larger TVC capacity the VTS can run in a *peak write* mode longer. This effectively allows the application to borrow performance from time periods when the VTS resources would be underutilized.
2. With a larger TVC, more recently written or accessed tape volumes are in tape cache. This increases the probability of read hits and write hits on specific mounts. This increases the number of tape I/Os that occur at disk speeds. It also decreases the number of physical tape mounts required (allowing other write and read misses to complete faster on the average) .

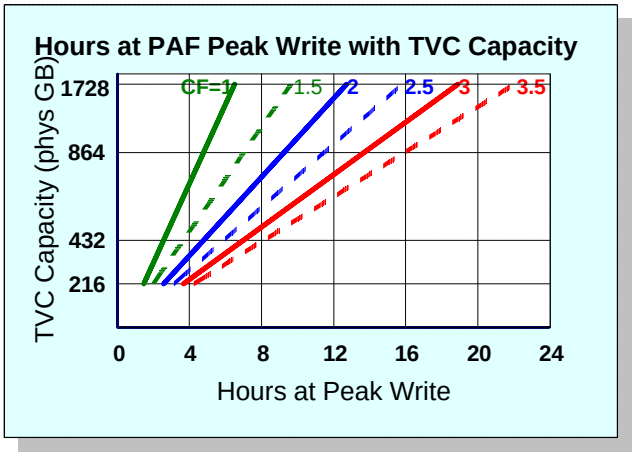

Fig. 10. The TVC capacity required to achieve a peak write throughput period given on the horizontal axis, as a function of the compressibility of the data (shown as labels on the curves). For TVC capacities greater than 864 GB the algorithms for managing the cache content are modified.

The bullet #2 above, depends on specific workload characteristics intimately and are thus hard to use in quantifying the general value of tape cache capacity except for the observation that more is better. Bullet #1, however, can be quantified approximately. Figure 10 shows the dependence of the time at PAF peak write throughput as a function of the TVC capacity. These curves have been determined from measurements. Note that in

Fig. 10 the trend in peak write duration with TVC capacity is modified between 864 GB and 1728 GB. This is because above a cache capacity of 864 GB the algorithms managing the TVC contents are modified. The peak write period duration read from Fig. 10 should be taken as an estimate, especially in the case of the 1200 GB effective capacity, since the specific workload characteristics can affect the duration.

5. VTS Input Throttling

VTS input throttling is implemented to allow background processes, those not involving data transfer over the ESCON channels, to obtain a sufficient share of the resources to maintain the system in balance. Such processes are, for example, recall of data that are on physical tape and copying of data in tape cache to physical tape.

If the VTS has been in a peak throughput mode for some time, i.e., with the copy process from tape volume cache to physical tape not keeping up with the rate of new incoming data, it is possible for the cache to reach a fill level with new data where no more can be accepted safely without first copying some of the data in the tape volume cache to physical tape. The same situation can also occur with recalls (i.e., read misses or specific mount write updates) since the data being recalled has to displace data in tape cache. To avoid such situations, the VTS throttles the input. A delay is added to each incoming write operation. This increases the available internal bandwidth for the background processes and allows it to stay up with new incoming data. If the VTS is observed to be often or persistently in a throttling mode (large input delays), the throughput of the VTS is being utilized at its maximum sustainable level and increased VTS capacity may be advisable. With PAF, input is not throttled for recalls and when copying to physical tape.

6. B18 Performance using the optional SCSI Host Attachment

A configuration option for the B18, available beginning in May 1999, is to replace the Additional Extended Performance ESCON Channels with one or two SCSI Host Attachment Features. Each of these SCSI attachment features, also referred to as *adapters* here, provides connectivity for up to two fast/wide or ultra-SCSI busses. This section presents the performance of the VTS as viewed from the SCSI interface. At the end of this section we also discuss the effect of SCSI attachment on the VTS performance as viewed at the ESCON interface. The data presented in this section have been obtained on a VTS (GA Sep 99) with the EHPO option; it is expected to be the same with the current VTS. This section will be updated later on SCSI performance with the PAF feature.

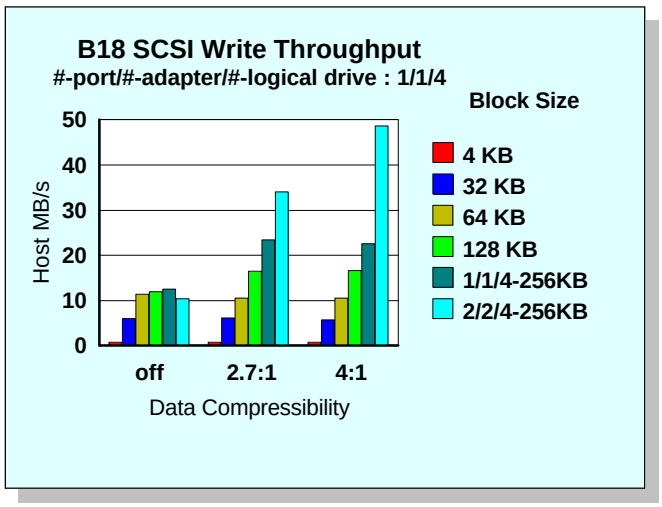

Fig. 11. The B18 SCSI write throughput under peak/fill *conditions for a variety of block sizes and values of data compressibility.*

A number of user controllable parameters determine the throughput performance of the SCSI attachment. Figure 11 shows the effect of block size on the write throughput of one adapter with data of various compressibility (a two adapter result at 256 KB is also shown for comparison). The measurements were done with no I/O load on the ESCON channels. The tape cache capacity was 288 GB and the experimental conditions are those for

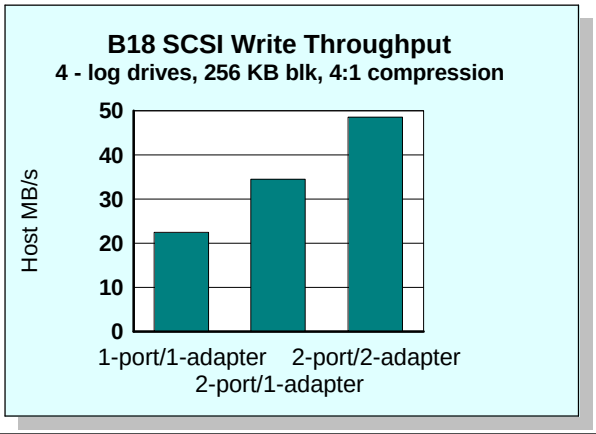

Fig. 12. The B18 SCSI write throughput under peak/fill conditions, varying the number of SCSI ports and adapters.

peak write/fill. For block sizes of 64 KB and below there is little effect of compressibility on throughput. At the larger block sizes, throughput improves with data compressibility and the improvement goes with block size. The write throughput goes from near one MB/s with 4 KB blocking to about 23 MB/s with 256 KB blocks (at compressibility of 2.7 and above). For the 256 KB block case we also show a bar for the case with two logical tape drives being driven through separate SCSI busses to two adapter cards (2-ports/4-log drives on two adapter cards).

The effect of multiple SCSI paths and adapter cards is further elaborated in Fig. 12 for the conditions indicated. The left and right bars correspond to the (1/1/4 and 2/2/4-256KB) cases in Fig. 11. For the 2-port/1-adapter case the number of logical drives was also varied. Four logical drives represents approximately the maximum throughput. Either decreasing or increasing (up to eight) the number of logical drives can reduce the throughput approximately 10%. **For maximum throughput one should maximize the number of data paths: both SCSI busses and the number of host adapters.**

The read throughput generally follows the write throughput dependence shown in Figs. 11 and 12, except that it tends to be somewhat lower (up to twenty percent) for the larger block sizes. Figure 13 shows the insensitivity of read throughput with data compressibility and the strong effect of block size for the same conditions as in Fig. 11 for writes.

The effect of SCSI input on ESCON throughput appears to be small, but the two throughputs share an aggregate maximum. For example, with a *peak write/fill* 2 x ESCON write input (32 KB block, 2.7:1 compression) saturated at about 16 MB/s, when a two-logical-drive/two-card SCSI write input (256 KB blocking, 4:1 compression) was added, the total VTS input rate , ESCON plus SCSI, was observed to reach a maximum of about 47 MB/s. This is about the same maximum rate as was observed with the SCSI input alone. Of the 47 MB/s about 15 MB/s was observed to be coming across the ESCON channels.

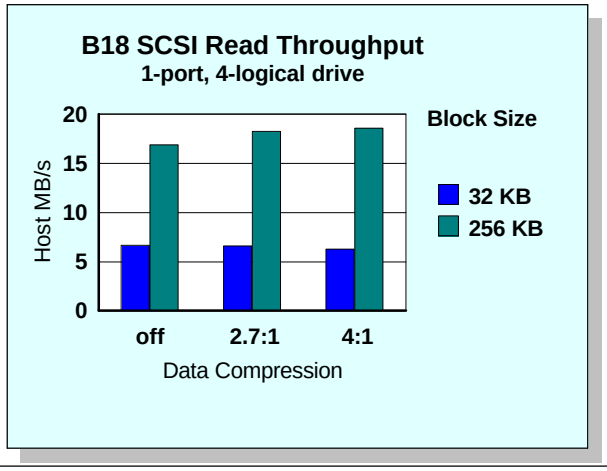

Fig. 13. The B18 SCSI read hit throughput as a function of block size and data compressibility.

7. B18 Performance using the optional Import/Export Feature

Beginning in May of 1999, the B18 offers an optional tape cartridge data Import/Export Feature (Advanced Function Feature Code 4000). Using this feature, user specified tape data logical volumes can be assembled on a tape cartridge which can be physically removed from the VTS (export). Similarly data thus exported on this or other VTSs can be physically "imported" (by introducing the cartridge into the convenience I/O station) to become part of the tape volume repository managed by this VTS.

Import/Export performance is measured by the elapsed time required to perform an export or import task consisting of a number of logical tape volumes on a number of physical cartridges. The considerations that determine the time required are:

1. Number of physical cartridge mounts/demounts in the aggregate task
2. Total amount of data transferred
3. If the tasks involve selected volume export/imports from the same physical volume, then tape drive repositioning time between volumes needs to be accounted for (avg. random repositioning time [search time] is about 21 seconds)
4. What the non-Export/Import workload is on the VTS

In Fig. 14 shows some measurements on a variety of import/export tasks. The measurements were done on a VTS (GA Sep 99); the results are not expected to differ on the current VTS. The principal determinant of task time is seen to be the aggregate number of physical tape cartridge mounts in the task. The trend lines shown (blue line [lower] for export, red line [upper] for import) are based on points all involving eighteen or less total mounts/demounts and amounting to a total data transfer of about 2.3 GB (data as compressed). They

describe well the outlying points, which involved about a 64% greater total data transfer.

At a minimum, the export or import of a single logical tape volume requires two physical tape cartridge mounts; to mount the source and destination cartridges.

The trend lines in Fig 14 for export and import respectively are given by:

$$t_e(\text{min}) = 0.00296 \times MB + 2.36 \times m + 0.35 \times s \quad [1]$$
$$t_i(\text{min}) = 0.00276 \times MB + 3.75 \times m + 0.35 \times s \quad [2]$$

where MB is the total number of megabytes (physical) transferred, m the number of mounts is the sum of source and destination drive mounts, and s is the number of searches along the length of the tape. Generally, the $m + s =$ the total number of logical volumes in the source of the export or import. However, in some cases the number can be larger than that. Although the data in Fig. 14 did not include search times, the formulas have been elaborated to also cover the cases of selective Export/Import which will involve search times. The MB are the net megabytes after VTS B18 EHPO feature compression. The present results were obtained from measurements on a three frame 3494 library with one dual gripper accessor and floating home cell operation. Larger libraries will incur longer mount times. A single gripper accessor and fixed home cell operation would also increase the observed mount/demount time. Furthermore, the exact export or import task time will also depend on cartridge placement in the library and specific distribution of logical volumes on the physical cartridges. Thus the illustrated results should be used as a guideline, and not a guarantee of performance.

There is an overhead associated with the beginning of a task involving the import or export of a number of logical volumes. This overhead is not significant on the scale of Fig. 14, and is not included in the task time formulas above. However, when the formulas are used to derive the time for the import or export of a single volume, for example, a discrepancy of up to two minutes can occur, both plus and minus. This discrepancy is attributed to the overhead and the fact that the concept of a statistical distribution of cartridge locations and accessor position breaks down.

As one might expect, the import/export task time will increase if there is another I/O load on the VTS. In a measurement of the time to complete a ten logical volume export request requiring *thirteen*

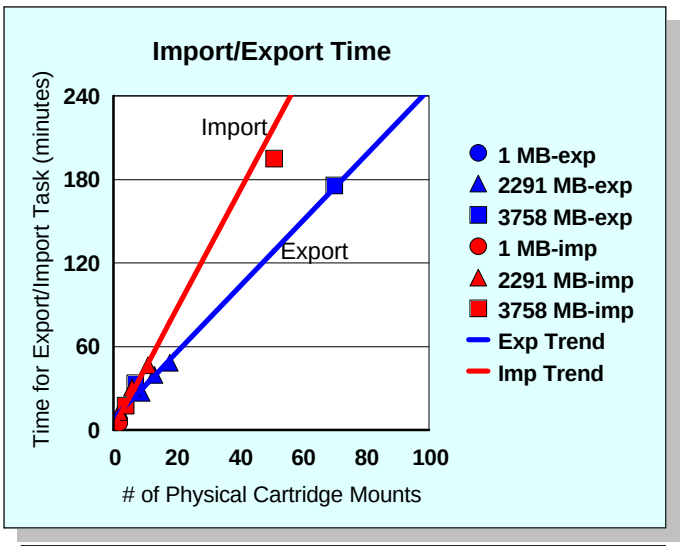

Fig. 14. The Import/Export task time as a function of the number of physical cartridge mounts.

mounts and a ten logical volume import request requiring *six* mounts (total 2.3 GB, each) at a time when the VTS was saturated in a "peak" write mode from the host, the former (export) time was elongated by about 25% versus no host load; the latter (import) by about 50%. Thus, the less mount activity an import/export task requires the more it is affected (relatively) by contention with other I/O activity on the VTS. Also, import/export with simultaneous (saturated) recall activity increases the import/export task time by about 50%.

Conversely, neither "peak" or "sustained" write throughput nor read hit throughput as measured at the ESCON host were greatly affected by simultaneous import/export activity on the VTS B18. However, read miss (recall) maximum rate was reduced to approximately half.

Import/Export Recommendations

1. It is recommended that import/export tasks be run during non-peak I/O activity times to minimize contention.
2. In order to reduce physical mounts required for export, the export operation should be performed as close to virtual volume creation time as possible.
3. Virtual volumes with similar creation dates should be exported together. This increases the probability that multiple volumes will be found on the same physical source cartridge.

8. Performance Considerations and Tools

8.1 Workload Considerations

One of the VTS objectives is to reduce the number of physical tape drives required and to achieve a more efficient use of tape by "stacking" multiple volumes on a single physical tape cartridge. The VTS design point has been to satisfy the typical workload of an installation, including peak activity which can be buffered in the tape cache, but excluding some workload components that can not achieve a performance benefit from VTS; namely, excluding HSM volumes, DUMPs, Tivoli Storage Manager volumes and others that already tend to write full physical tapes. These latter do not achieve any net efficiency in going from host to tape though the VTS. And they will not achieve any benefit from residence in tape volume cache. On the contrary, they will consume a significant fraction of the VTS internal bandwidth and will monopolize a fraction of the VTS tape drives.

Our performance recommendation is that, in considering whether to commit workload components such as HSM and DUMPs to the VTS, the tools described below should be used to assess the performance impact. There will be cases in which allocating such work to native tape drives or native tape drives within the 3494 tape library can significantly enhance the performance of the VTS.

It is recognized, however, that customers will want to assign their HSM and DUMP volumes to the VTS for simplicity and uniformity in data management. The current release of the B18 with the enhanced host attachment bandwidth and PAF significantly assist in this objective.

8.2 Effect of ESCON Distance

The performance results quoted above in this paper have been obtained with the ESCON connection between host and VTS within small distances compared to one kilometer (km). Within about 25 km distances the following rule of thumb can be used to estimate the effect of ESCON distance using IBM ESCON directors:

The VTS host write and read throughput is reduced approximately 0.6% and 2.3%, respectively, per kilometer of ESCON distance between the host and VTS for 32 KB block transfers.

This write rate at extended ESCON distance is an improvement introduced with the current VTS. It results from a modification of how VTS ESCON input buffering is managed to handle larger transfer block sizes.

8.3 Tape Magic

Tape Magic is a high-level tape subsystem configurator available to IBM customer representatives that is intended to give an initial prediction of a tape configuration that would satisfy a customer's tape processing needs. Tape Magic predicts both native and volume-stacking configurations. Input to Tape Magic is answers to a half-dozen or so simple questions about basic customer tape workload characteristics, typically entered via a Thinkpad on a visit to the customer's location. Because Tape Magic does not directly process any host-processor statistical data, such as MVS SMF records, it is also useful for host platforms that do not provide data that can be input to IBM's more detailed configuration tools.

8.4 Workload Analysis and Configuration Estimation Tools

A more accurate assessment of a VTS configuration than possible with Tape Magic can be made by a detailed analysis of the customer's workload as represented in SMF records, RMF data, and tape management system data. The current tool available to IBM representatives is called Consul Batch Magic and provides a detailed analysis of existing customer tape workload characteristics and projects the required VTS configurations for a subset of that workload. CBM uses as input, selected raw SMF records (14,15,21,30) to provide basic tape workload characteristics such as mount and drive allocation activity as well as input and output tape data transfer activity by hour. To project a VTS configuration, the user first uses the extensive filtering capabilities of CBM to identify certain tape activity, such as output files destined for trucking to a remote vault and tape activity that already efficiently utilizes native tape, that will not be volume-stacked. CBM then projects required VTS and native drive configurations based on the current workload. CBM also provides numerous statistics on expected VTS cache performance.

8.5 Performance Monitoring Tools

VTS generates data that is transmitted each hour to the host processor, where the data is embodied in an SMF type 94 record. This SMF record also contains information on library performance associated with native tape drives. Information provided in the SMF type 94 record

includes logical and physical drive usage, number of fast-ready (virtual scratch), read-hit, and recall mounts, channel and tape input and output data transfer activity, and cache usage statistics. IBM provides routines that give hourly and daily reports on these VTS statistics. This allows the customers to understand the level of activity of their VTS subsystems, and allows customers, with assistance provided by IBM field personnel, to determine when the limits of the VTS subsystems are being reached.

9. Conclusions

The virtual tape server, beginning with the VTS model B16 (GA May 97), has demonstrated a clear customer requirement for consolidated tape data management and automation, while taking advantage of technological advances that reduce hardware and floor-space requirements. The VTS, model B18, builds on the original VTS base to offer new function and significantly improved throughput performance. These offerings include SCSI host attachment, Import/Export of VTS format cartridges, enhanced ESCON attachment bandwidth and data compression (EHPO), and the *Performance Accelerator* feature (PAF). The new VTS extends VTS performance with the availability of up to eight ESCON channel host connectivity and up to a six-fold increase in tape volume cache (TVC) capacity. The result is a peak write throughput of up to 80 MB/s together with potentially additional significant performance possible via read/write TVC hits resulting from the larger cache capacity (the latter depends on workload characteristics). This continuous improvement reflects the IBM storage modular *Seascape* architecture in which technological improvements in components can be quickly incorporated in the product to provide customer solutions, and in many cases protects customer investment through upgrade compatibility.

Appendix A

Table 4. Linear Coefficients for the EHPO *host MB/s* Curves in Figs. 7 and 8.

	Tape Cache (*):	≤288 /72 GB	576 /144 GB	1152 /288 GB
Sustained Write	*m(GB)*	3.91	5.34	6.23
	b(GB)	1.58	0.71	0.82
Peak Write	*m(GB)*	6.33	9.04	10.54
	b(GB)	1.67	-0.03	-0.54

*) The first TVC capacity is for 9 GB HDDs, the second for 36 GB HDDs (see Table 3)

Table 3 shows the available HDD configurations for the VTS considering the capacity of the HDDs chosen and whether the VTS is a base configuration, has only the EPHO option, or has the PAF feature installed. The configuration is shown in terms of RAID-5 parity groups. For example, 8(2+P) describes an HDD configuration comprising eight 2+P RAID-5 parity groups. Spare drives are not shown; there is one spare drive with each parity group.

Table 3. VTS Tape Volume Cache Capacities and Configurations.

TVC Cap (GB)	9GB HDDs Base/ EHPO	36GB HDDs Base/ EHPO	36GB HDDs PAF
72	4(2+P)	4(2+P)	
144	8(2+P)*	4(2+P)	
216	12(2+P)		2(6+P)
288	16(2+P)*	4(2+P)	
432			2(6+P)
576		8(2+P)	
864			4(6+P)
1,152		16(2+P)	
1,728			8(6+P)

*) In earlier shipments these configurations were available with PAF

Appendix B

Analytical data on VTS/EHPO throughput as a function of tape cache capacity and data compressibility.

The B18 w/EHPO sustained and peak throughput rates shown in Figs. 7 and 8 can be reproduced from linear relationships of the form:

$$(hostMB/s) = [m(GB) \times CF] + b(GB), \quad [3]$$

where the parameters m and b are functions of the size of the Tape Cache in GB, and CF is the data compressibility. The values of m and b are given in Table 4.

Care should be taken to make sure that the input, CF, and output, (MB/s), of this formula is maintained within the bounds shown in Figs. 7 and 8; namely $1 \le CF \le 4$. Performance claims at larger compression factors, but still with $(hostMB/s) \le 50$, should be considered non-validated extrapolations.

DISCLAIMERS

The performance information contained in this document was derived under specific operating and environmental conditions. While the information has been reviewed by IBM for accuracy under the given conditions, the results obtained in specific operating environments may vary significantly. Accordingly, IBM does not provide any representations, assurances, guarantees or warranties regarding performance. Please contact your IBM marketing representative for assistance in assessing the performance implications of the product in your specific environment.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

TRADEMARKS and COPYRIGHTS

The following terms are trademarks of the IBM Corporation in the United States or other countries or both:

IBM	RS/6000	Magstar
ESCON	DFSMS	SMF
Thinkpad		

Table of Contents