IBM Magstar 3494
Peer-to-Peer Virtual Tape Server

Performance White Paper

Version 1.0

IBM Corporation 28 July 2000

---

1. Introduction

This paper provides performance information on the IBM Magstar 3494 Peer-to-Peer Virtual Tape Server (PtP VTS). The PtP VTS provides automated dual copy tape data management and storage through a single storage system image, one copy on each of two VTS model B18s. The PtP VTS physically comprises two 3494 model B18 VTSs, each with four Extended Performance ESCON channels and the Performance Accelerator feature (PAF). The interconnection between the two VTSs is provided by four 3494 AX0 Virtual Tape Controllers which are unique to the PtP VTS. The performance related architecture of the 3494 model B18 and its performance is described in a separate performance white paper [1].

The two components of the PtP VTS (AX0s and B18s) can be physically adjacent or they can be separated by up to 26 km. The second copy of the data can be made immediately at volume close time (*immediate copy* mode), or its timing can be managed by the PtP VTS using customer set policies (*deferred copy* mode). The *deferred copy* mode is useful at times when having a write bandwidth approaching that of two VTSs for a period of time is an advantage.

2. Product Description (PtP VTS)

Figure 1 shows the physical configuration of the PtP VTS. The figure shows two VTS units connected via ESCON communication links. Each of the VTS units shown comprises (from left to right) a *IBM 3494 Tape Library*, extendable from the two unit model shown, the *VTS B18*, and a frame which houses the *Virtual Tape Controllers*. Having Virtual Tape Controllers at both VTSs is optional; they can all be housed with one VTS. The ESCON connections are illustrated in the PtP VTS schematic in Fig. 2.

Fig. 1. Peer-to-Peer VTS.

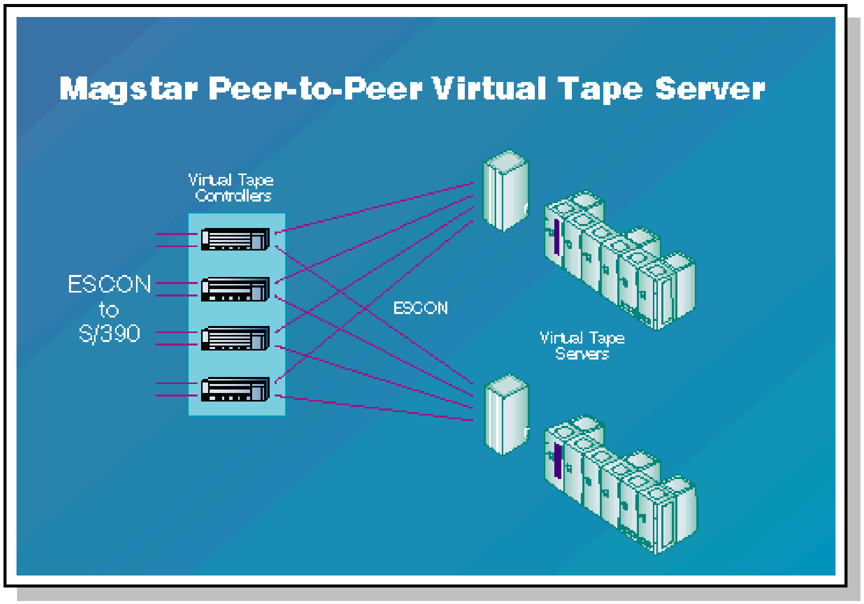

Fig. 2. Magstar Peer-to-Peer VTS showing ESCON interconnection wi the Virtual Tape Controllers.

Figure 2 shows the interconnection scheme between S/390, the Virtual Tape Controllers, and the VTSs. All the channels are ESCON; from the Virtual Tape Controllers there are eight channels to S/390 hosts, while there are four paths to each of the B18s. All data transmission between the two VTSs occurs via the four paths through the Virtual Tape Controllers.

3. Operational Modes

The operational modes refer to how data written to the PtP VTS are handled. Data is initially directed to one of the two VTSs. Balancing algorithms keep the loads on the two VTSs approximately equal. A copy is made to the other VTS when the tape volume on the first VTS is complete in tape cache. The copy is made prior to completion of the logical rewind of the original (*immediate copy* mode) or it can be deferred to a later time (*deferred copy* mode). In *immediate copy* mode, when a volume completes *close* processing, it means that the PtP VTS has completed performing the copy. The deferred copy mode is provided to balance the PtP VTS workload when very high input rates must be sustained for periods of time while periods of lower input rate allow the synchronization of the data on the two VTSs.

In addition to the PtP VTS operating modes just described, each VTS (B18) can operate in a "peak write manner," in which copying of data from tape cache to a stacked volume is not required to keep up with the input to the tape cache. To distinguish this VTS(B18) mode from the PtP VTS modes we refer to it as operating in the *peak write manner*. Similarly, the VTS *sustained write manner* refers to operation of the VTS with the write input rate balanced with the copy to stacked volumes.

Regardless of the mode/manner, the internal algorithms do as much of the background work (peer-to-peer copies and copies to physical tape) as possible with any excess bandwidth that is available. Thus, unless there is a strict requirement to keep the VTSs synchronized, the best performance can be observed with the PtP VTS in *deferred-copy/peak* manner, within the constraint of a maximum aggregate write input for a period defined by TVC capacity and availability of reduced input rate during which deferred copy volumes can be copied (details in Sec. 5). When the maximum write input is occurring, potentially almost all asynchronous background copy work is suspended in order to handle read/write traffic with the host. The other extreme is the *immediate-copy/sustained* manner in which all dual copies are being made. Performance in this manner can be sustained indefinitely.

4. Performance Metrics

In the following sections we present the performance of the PtP VTS as viewed from the host. The metric we use is megabyte per second (MB/s) in each of the possible combinations of PtP VTS modes and VTS(B18) operating manners. The data is presented as a function of data compressibility since data is compressed by the VTS. Data compressibility has a significant effect on the observed PtP VTS throughput.

All of the measurements and modeling of performance assume eight active host ESCON channels. The workload comprises sixty-four applications writing 800 MB tape volumes simultaneously. The block size used is 32 KB and the BUFNO parameter is set at 20. When there is a read component to the workload, it is assumed to be in volumes of 250 host MB. The workloads used in this paper are either "100% write" or "typical mix." The latter comprises 60% writes and 40% reads, with the

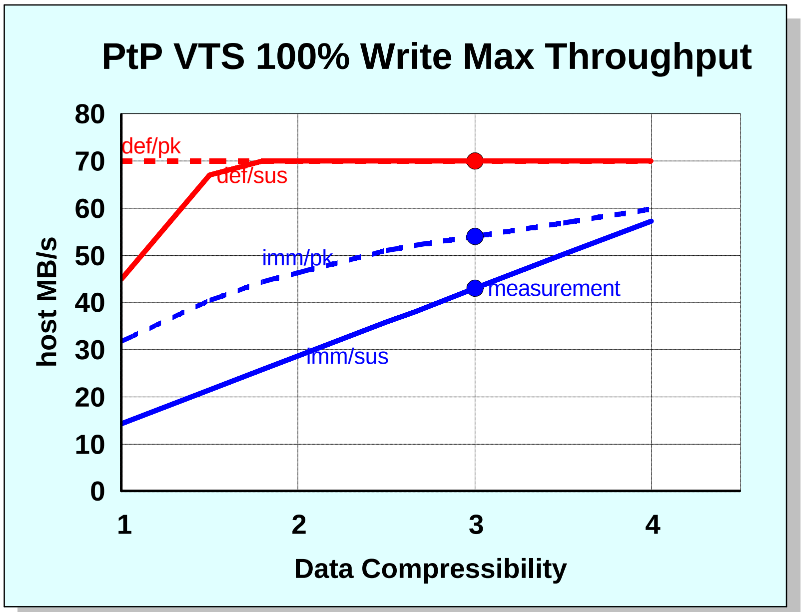

Fig. 3. PtP VTS maximum throughput in host MB/s for a 100% write workload as a function of data compressibility. These data assume that all AX0s and B18s are local.

reads having a 50% hit ratio in tape volume cache (TVC).

In this section we present data for the case when all PtP VTS components, AX0s and B18s, are local. Performance with any of the components remote (i.e., at greater than 1 km distance) is discussed in Sec. 6.

4.1 100% Write Performance; All AX0s and B18s Local

Figure 3 shows the expectec performance of the PtP VTS under a 100% write workload. The curves represent dat derived from a performance model calibrate to the measurements shown.

The upper curve marked def/pk is th performance in *deferred copy* mode; that is with copies from one VTS to the other bein deferred in favor of maximum host writ throughput.

The lower two curves show the performance in *immediate copy* mode. Ir this mode a copy is made on the second VT before the *rewind/unload* completion is give to the host. The second part of each of th performance curve designations, "pk" an "sus" refer to whether the VTSs (B18s) are operating in peak or sustained write manner, respectively. "Peak" write refers to the VTS operating manner in which copying of data from tape cache to physical tape is not being done, thereby allowing a higher host write input rate. In the "sustained" write manner, the rate of copying of data from tape cache to physical tape is the same as the host input rate. The maximum host throughput is reduced by the work required to make the second copies concurrently with host write input.

4.2 Write Copy Mode Algorithms

The *deferred* and *immediate copy* modes of operating performance represent the envelope of write throughput bandwidth. The default is the *deferred copy* mode. In this mode, copies make up the balance of the PtP VTS workload if the host input is not at peak bandwidth. Thus, there is no substantial loss of PtP VTS resource utilization in the *deferred copy* mode if the host input should lapse. In effect, any time the host I/O rate drops below the maximum *deferred copy* mode level some of the background peer-to-peer copies and copies to the stacked tape are done. If the write input drops to about 40 MB/s (at CF=3), then the background work in host MB/s approximately keeps up with the write input rate (i.e., the amount of deferred copy and copy-to-tape data in TVC stops increasing). This is different from the *immediate*

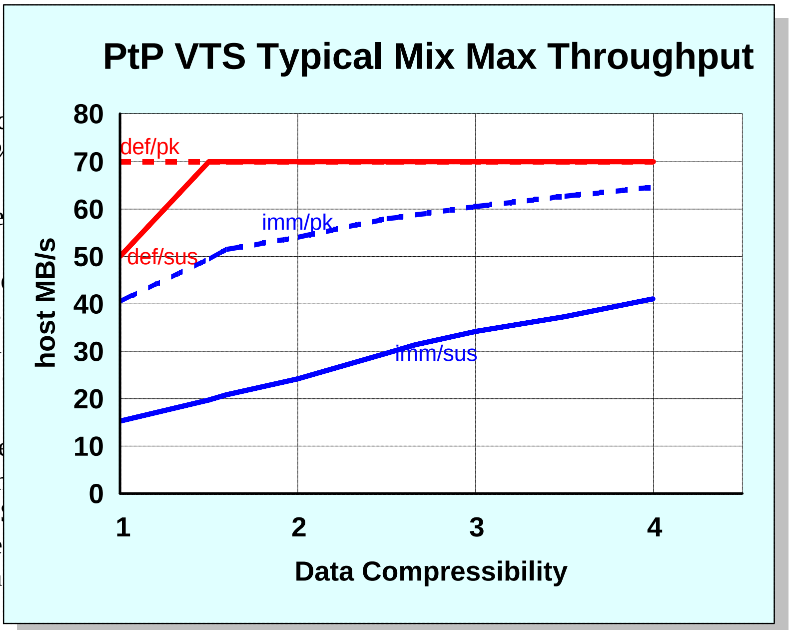

Fig. 4. Modeled PtP VTS maximum throughput in host MB/s for a "Typical Mix" workload as a function of data compressibility. These data assume that all AX0s and B18s are local.

copy mode in that response time to the end of tape *close* processing is quicker; but the copy process is executed at a lower priority. Thus, except for applications requiring the higher level of data availability and security against loss, the *deferred copy* mode makes efficient use of PtP VTS resources, offers the better response time, and the best response to host requirements for maximum throughput.

4.3 Typical Mix Performance; All AX0s and B18s Local

The *typical mix* workload is defined here as a mix of 60% writes and 40% reads (in terms of host write/read MB). In addition it is assumed that 50% of the reads are tape volume cache hits and that the remaining reads require staging of data from physical tape. This mix represents approximately what is found in a grand average of operating statistics in actual application environments. Individual applications or time periods during operation can vary widely from this average.

The typical mix workload performance projected from modeling is shown in Fig. 4. The *deferred copy* mode maximum throughput is approximately the same as for 100% writes in the typical operating range assuming a data compressibility of three. In the immediate copy mode the maximum throughput is somewhat higher than with 100% writes. That results primarily from the fact that read I/Os do not require a copy operation between the peer VTSs.

5. Tape Volume Cache (TVC) Effects on Performance

5.1 Host Write Performance

The length of time for which a *deferred copy* write mode can be sustained is dependent on the host input rate and th tape volume cache capacity, because th mode requires the buffering of data in th TVC for later copy operations.

Figure 5 shows, for four TVC sizes (physical GB per VTS), the length of tim that the PtP VTS can be operated in *deferred copy* write mode before it become necessary for the PtP VTS to utilize some o its bandwidth capability to make dual copies of logical volumes which have accumulated in the TVC. For example, the PtP VTS can be operated at the maximum write rate of 70 MB/s host input (assume a data compressibility of three), for over eleven hours if the TVC capacity is 864 GB per VTS.

At lower input rates the time to fill the TVC to the maximum allowed un-copied data level is extended. For example, if the average deferred copy write rate is 60 MB/s, then the capacity criterion for the 864 GB TVC allows the accumulation of the input for about 17 hours. Note that these curves include an estimate of how much background peer-to-peer copy work is getting done at any given input rate.

In estimating the hours at a 60 MB/s deferred copy write input rate from Fig. 5, we start at the 60 MB/s level on the vertical axis and proceed to the right until we intercept the TVC size of the PtP VTS configuration being assessed. Note that, if the TVC size is 864 GB, we first cross the line labeled "24 Hr Limit" at about 13.5 hours. This limit means that if we want to manage the PtP VTS installation in such a way that there is no un-copied data remaining more than twenty-four hours, then we have to stop host input entirely at 13.5 hours for there to be enough time for all the un-copied data to be copied to the peer VTS in the remaining time of that twenty-four hour cycle.

The corresponding "24 Hr Limit" at the maximum deferred copy write input rate of 70 MB/s is at about 11 hours, somewhat before the maximum

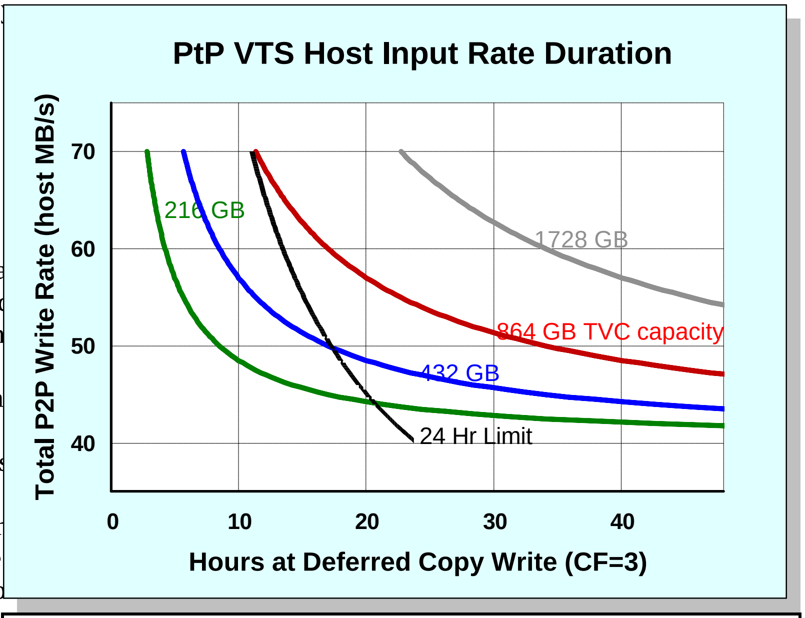

Fig. 5. The PtP VTS host deferred copy write input rate maximum duration as a function of VTS TVC capacity and the "24 hour limit," if it is required that all data be copied within a twenty-four hour period. A data compressibility of three is assumed. The TVC sizes are given in physical GB per VTS.

un-copied data limit of the 864 GB cache has been reached.

Another parameter affecting deferred-copy write performance is the *deferred copy priority threshold* that specifies the maximum age of un-copied data in the TVC (in integral 0 to 24 hours). Once this *hours* age has been reached by a tape volume its priority for being copied is increased. This parameter exists to minimize the time that the age of un-copied data will exceed the *hours* value. The way it can affect performance is that when the *hours* threshold is reached the copy load is added to the existing host I/O load. For example, if the PtP VTS is operating in deferred-copy mode and un-copied data in TVC reaches the specified *hours* threshold, the observed write rate will drop because of the additional copy load. Thus, for best performance, the *deferred copy priority threshold (hours)* parameter should be set to exceed the anticipated critical peak write requirement period duration. It also requires that there was no old un-copied data in the TVC at the beginning of the critical peak write period.

If there is no host I/O input, the PtP VTS can copy approximately 65 physical GB/hr between the PtP VTSs (32.5 GB/hr in each direction). For a PtP VTS with two 864 GB caches, only half of which is allowed to fill with un-copied data, the time to copy all the data then is about 13.3 hours.

All of the TVC size curves in Fig. 5 are asymptotic to the 40 MB/s level. At a deferred copy write rate of about 40 MB/s there is no net accumulation of un-copied data in the TVC.

If there is some write input during the off-peak time, the peak deferred copy write window under a twenty-four hour recovery constraint is reduced as shown in Fig. 6. For example, if the average write rate is expected to be about 20 MB/s during the off-peak time, then the maximum allowable peak period rate should be no more than about 7 hours. The formula for the curve in Fig. 6 is:

$$HrAtPeak \; ? \; \frac{24}{\left(\frac{m}{1?y}\right)?1},$$

where m = Pk/(CF*18.06) and y = 1- Wr/(CF*13.33), CF is the data compressibility, and Pk and Wr are the "net" deferred-copy peak and off-peak write rates, respectively. "Net" refers to the fact that only about 90% of the maximum deferred copy write input is accumulating in cache; about 10% of the write input is getting copied to the peer VTS at the maximum write input rate.

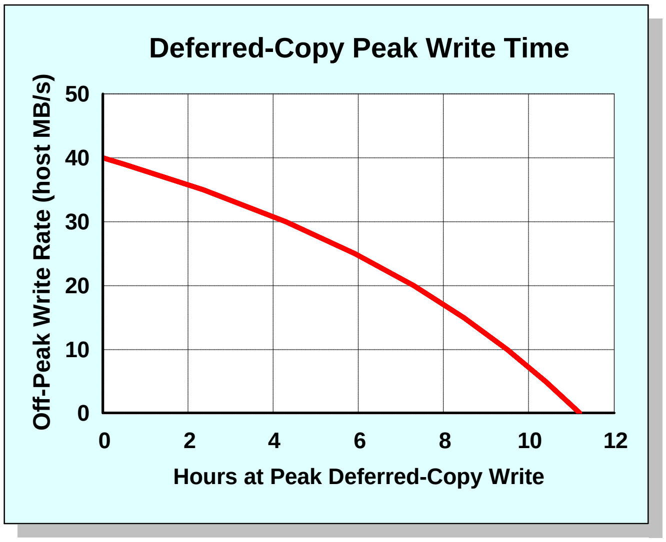

Fig. 6. An approximate empirical guideline for estimating the number of hours one can operate the PtP VTS at the maximum deferred copy write rate (70 MB/s) in a twenty-four hour period given an off-peak write input shown on the vertical axis. The TVC size must be large enough to hold the un-copied data for the peak period (see Fig. 5) and the deferred copy priority threshold (hours) *has to allow it. For this figure a data compressibility of three is assumed*

Based on the foregoing analysis shown in Figs. 5 and 6, we can conclude that in order to have access to the full range of deferred copy write duration in a twenty-four hour cycle, a 864 GB TVC per VTS is necessary, assuming a data compressibility of three. This allows for storage of the peak deferred copy write data as well as extra capacity, thus increasing the read hit ratio. Smaller TVC capacity is possible in environments with a more limited peak deferred-write requirement. Larger TVC capacity can extend peak deferred copy write time over twenty hours and/or increase performance through an increased read hit probability.

All of the data and recommendations in this section should be taken as planning guideline approximations. It is suggested that, in using the deferred copy periods given in Figs. 5 and 6, a safety factor be included as some variation from these numbers can be expected with usage patterns and the specific workload.

5.2 Host Read Performance

Read throughput and response time performance is significantly better if the I/O can be served from TVC (read hit), versus requiring a recall from physical tape. As a result performance planning should take into account an allowance of TVC capacity for read hits. In order to improve the TVC capacity available for read hits, the PtP VTS attempts to keep only one copy of a particular logical volume in cache at any given time. This makes the effective cache size for the purpose of read hits approximately equal to the combined size of the TVCs of the two VTSs.

6. PtP VTS Remote Dual Copy Performance

The performance cited in Secs. 4 and 5 is with all of the hardware "local," i.e., within a computer facility complex (< 1 km). With dual copy tape such as is provided by PtP VTS, having one of the copies at a remote location provides additional protection against data loss and availability. The throughput performance of the PtP VTS, however, is affected by the distance and the nature of the connection to the remote VTS. Fig. 7 shows the base "local" configuration and an expected most common remote configuration.

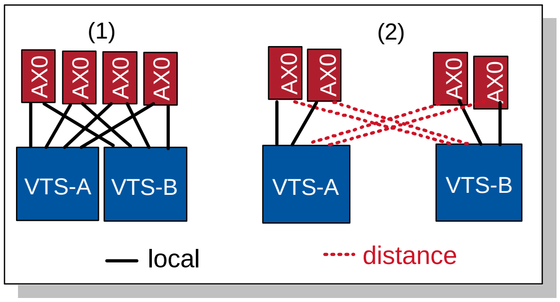

Fig. 7. Common PtP VTS configurations. The configuration (1) has all hardware local. Configuration (2) has the hardware "split" to two locations at a distance. It is assumed here that the host-AX0 connections are "local."

The underlying reason affecting performance at a distance is the data propagation time from the origin of the data to the remote location. The propagation rate is that of the speed of light reduced somewhat by the dielectric properties of the transmission fiber. Besides the protocol, data between the source and destination is transmitted in blocks. Before a transmission block is sent by the source, it waits for an acknowledgment from the destination that the previous block has been received without error. The resulting ESCON data rate is determined by a combination involving the distance, transmission and noise characteristics of the line (determines the number of re-transmissions required due to data error), buffer sizes at the source and destination, as well as the logical and transmission block sizes. None of these parameters are user selectable except the choice of distance and directors, when required.

6.1 PtP VTS Performance with a workload balanced on the AX0s

In the following remote VTS performance projections we use modeled data based on the at-distance transmission characteristics obtained from a measurement at a distance of 25 km. The configuration used is (2) in Fig. 7, with a symmetric half of the PtP VTS at each of two locations separated by a distance. It is assumed that the workload input to the four AX0s remains balanced. The maximum throughput is reduced if a workload skew is introduced by splitting the AX0s (cf., Sec 6.2).

In the split VTS configuration (2), the elapsed time (as viewed from an AX0) is generally shorter for I/Os executed on the local VTS compared to the remote VTS. This *response time* effect will be apparent for read and deferred copy mode write I/Os if they are directed preferentially to the local VTS. We examine below the *throughput* performance consequences of operating the AX0s in preferred VTS mode. (The choice of "no preference" or "preferred" VTS mode is made at the AX0. It is a static choice requiring a power-down of the AX0.)

The at-distance PtP VTS throughput performance is shown in Fig. 8. Let us first focus on the immediate copy mode write performance (Imm). The "no preference" mode is the normally recommended mode of operation for the PtP VTS. In

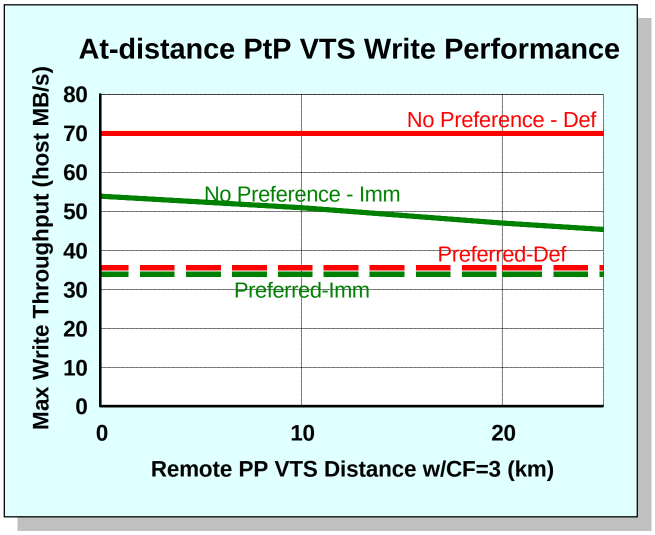

Fig. 8. The at-distance Immediate Copy *and* Deferred Copy *Mode write throughput performance of the PtP VTS. Data compressibility of three is assumed.*

this mode the AX0s attempt to balance the primary workload between the two VTSs. An AX0 can direct the incoming write to either of the two VTSs, regardless of whether the VTS is local or remote.

In the "preferred" mode all of the incoming writes are directed to the VTS specified. It is assumed in the discussions here that in preferred mode the preferred VTS for an AX0 is always the local one.

The expected maximum throughput is given in Fig. 8. The performance of the "no preference" mode is always expected to have a higher maximum throughput than the "preferred" mode. This is because the “no preference” mode has the ability to shift new work to balance the work at the two VTSs. This tends to make uniform use of the PtP VTS resources. Specifying a "preferred" VTS at an AX0 can leave one path to a VTS underutilized while the other is operating at maximum throughput, for example.

The “no preference” mode throughput is asymptotic to the “preferred” mode at large distances; for at very large distances the best work balance for performance is to have most of it done on the local VTS. The "preferred" mode lines are straight because the throughput, at the distances shown, is determined by the local ESCON paths, which remain constant in length.

Figure 8 also shows the expected at-distance throughput for writes in the deferred copy mode (Def**).** All the same considerations as for the immediate copy mode apply except that with deferred mode write there is no copy traffic (the copy process is deferred). The obvious result in Fig. 8 is that the "no preference" deferred copy mode offers significantly better throughput at all distances in the range shown. The “preferred” mode writes have about half the throughput of the “no preference” mode because each AX0 essentially has only a single ESCON channel to the “preferred” VTS.

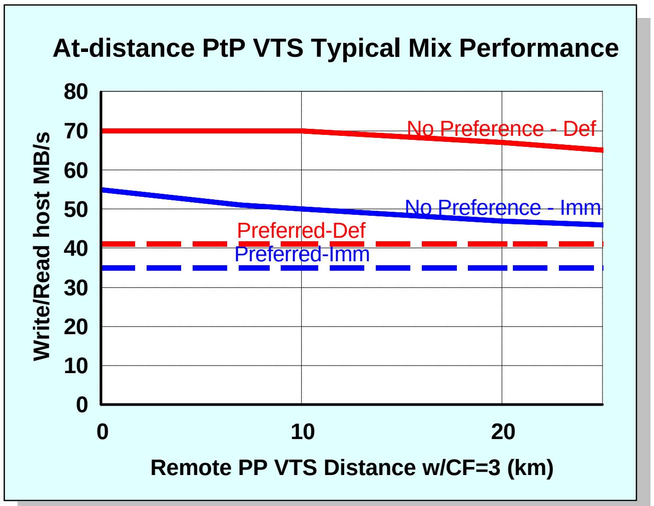

Fig. 9. The at-distance Typical Mix (60% Write, 50% Read Hit) throughput performance of the PtP VTS in Immediate Copy and Deferred Copy modes.

As a result of the performance characteristics exhibited in Fig. 8, **the general recommendation is to operate the PtP VTS in "no preference" mode.** Only in cases where the hosts, AX0s, and VTSs are split at two separated sites should one consider “preferred” mode operation. In that case one will clearly want the local input to be preferentially targeted first to the local VTS. Even then, there is no *throughput* advantage to “preferred” mode; there is a throughput penalty. The principal advantage is in *response time* performance; namely, by having all tape I/O served locally, the deferred copy mode writes and reads will be more likely to have a shorter open time.

For practical workloads that involve a mix of read and write I/Os we have modeled the throughput performance of a **Typical Mix workload** (60% writes, 50% read hits). These results are shown in Fig. 9. The similarity of these results to those in Fig. 8 arise from an approximate balancing of the effects of less write copy traffic and the fact that the ESCON bandwidth for reads is somewhat smaller than for writes. The results are also sensitive to the size of the recall volume, here assumed to be 250 MB. The “no preference” mode throughput is always greater than the corresponding “preferred” mode case.

(Note that in the "no preference/Imm" mode a fraction of the I/Os incur a double "distance hit." Namely, if a remote VTS is chosen as the primary target for a write by the AX0, then the peer-to-peer copy has to travel the distance back again. From a throughput point of view this is still better than making all primary writes local. The latter preferred mode leaves the extended distance ESCON channels underutilized at all distances within 26 km.)

6.2 Split Host/AX0 Configuration Performance Considerations

In Sec. 6.1, the performance discussion assumed that the workload was balanced at the AX0s (meaning that the number of MB/s passing through these controllers was identical and in the same write/read ratio). In this section we consider the case of a likely split PtP VTS configuration with host processors at each location (cf., Fig. 10). We estimate the effect on total throughput performance in the case when the workload input from the two hosts is unbalanced. This might occur, for example, if the two hosts sharing the PtP VTS each periodically require the maximum eight ESCON write throughput to the PtP VTS, but at different times. Because the ESCON channels to the distant AX0s have a reduced throughput (because of the distance), this amounts to an unbalance in the work arriving at the two pairs of AX0s.

We have estimated the bandwidth of the ESCON network shown in Fig. 10 for the case when all the host I/O is coming from one host up to distances of 25 km. We find that the total write bandwidth of the eight ESCON network between the active host and the four AX0s is always on the order of 20% greater than the maximum throughput of the PtP VTS in its highest throughput mode ("no preference" /deferred copy write or typical mix; cf., Figs. 8 and 9). This margin is greater than the 25 km distance reduction in PtP VTS bandwidth when its components are at distance. We thus believe that Host A/Host B I/O rate skew will not reduce the net maximum PtP VTS throughput by a significant amount when the interconnection is as shown in Fig. 10. This is also true for the lower throughput modes of operation.

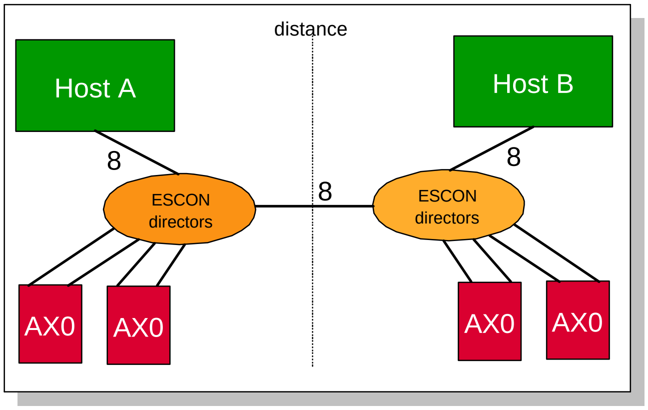

Fig. 10. A logical representation of a split PtP VTS configuration with host processors at each location. Each pair of AX0s is locally connected to a VTS. The detailed AX0/VTS connections are shown in Fig. 7. The configuration allows either host to have eight ESCON bandwidth access to the split PtP VTS. (The physical implementation of this configuration is best accomplished with four ESCON directors for data availability in case of director failure; resulting in eight ESCON channels interconnecting the two groups of directors, as shown.)

7. Performance Tools

7.1 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.

7.2 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.

7.3 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 to also, with assistance provided by IBM field personnel, to determine when the limits of the VTS subsystems are being reached.

8. Conclusions

The virtual tape server, beginning with the VTS model B16, 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 model B18 VTS, built on the B16 base offers significantly improved throughput performance using data compression (EHPO). The introduction of the *Performance Accelerator* feature provided significantly more power and efficiency in data handling. The Peer-to-Peer VTS builds on this continued improvement by implementing an automated dual copy capability together with a hardware configuration that has the ability to maintain access to data after component failures. The PtP VTS can be split among multiple locations to help ensure continuous data availability even in the event that a disaster at one site makes that hardware completely unavailable. This extension of the VTS tape storage solution technology reflects the IBM storage modular *Seascape* architecture in which technological improvements in components can be quickly incorporated and proven building blocks can be combined to offer new functionality.

9. References

1. IBM Corporation, "IBM Magstar 3494 Virtual Tape Server, Performance White Paper, Version 4.0," 28 July 2000.

DISCLAIMERS

TRADEMARKS and COPYRIGHTS

Table of Contents