18 Oct 2001

IBM TotalStorage™
Peer-to-Peer Virtual Tape Server Performance

By Aare Onton Senior Engineer/Scientist
Jesse L Thrall Senior Programmer

Contents

Introduction

This paper provides performance information on the new IBM TotalStorage™ Peer-to-Peer Virtual Tape Server (PtP VTS), Models B10 (PtP VTS B10) and B20 (PtP VTS B20). The PtP VTS provides automated dual copy tape data management and storage through a single storage system image, one copy on each of two Virtual Tape Servers (VTS). The new models feature a new hardware architecture, with more powerful processors and expanded I/O capability when compared with the previous Model B18 PtP VTS (PtP VTS B18) [Ref. 1]. The PtP VTS is physically comprised of two VTSs, which for the standard offering are identical. The VTSs are interconnected with Virtual Tape Controllers (VTCs, Model AX0) which are designed for the PtP VTS function. The VTCs also serve as the interface to the ESCON channels from the host(s). The performance related architecture of the VTS Models B10 and B20 and their performance is described in a separate performance white paper [Ref. 2]. The stand-alone VTS, Model B18 is also described in a separate white paper [Ref. 3].

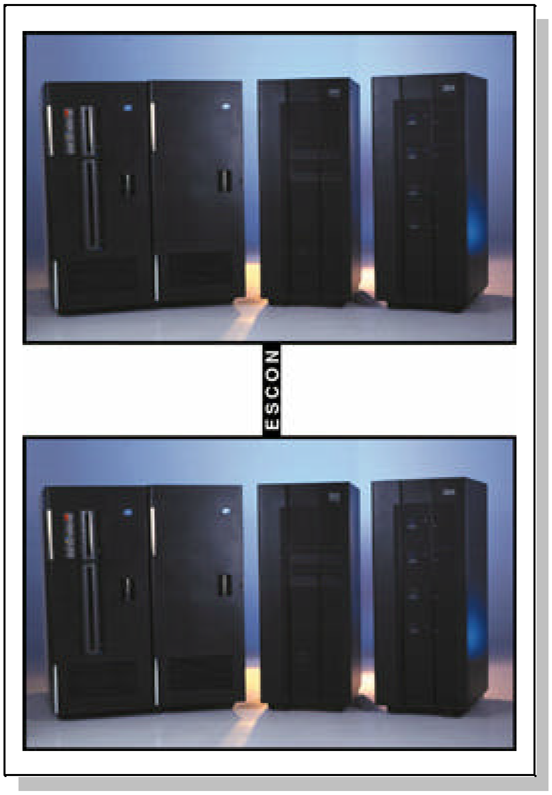

***Fig. 1**. Peer-to-Peer VTS.*

Highlights

The two components of the PtP VTS, VTCs and VTSs can be physically adjacent or they can be separated by extended distances (see the section on *Support of Remote Operation*). The second copy of the virtual volume can be made immediately at rewind/unload complete time (*immediate copy* mode), or its timing can be managed by the PtP VTS using customer set policies (*deferred copy* mode).

Product Description (PtP VTS)

Figure 1 shows the physical configuration of the PtP VTS. It shows two VTS units connected via ESCON communication links. The PtP VTS is comprised of (from left to right) an *IBM 3494 Tape Library*, extendable from the two unit model shown, the *VTS*, and a frame which houses the *VTCs.* Having VTCs at both VTSs is optional; they can all be located with one VTS, up to four VTCs per frame.

PtP VTS with sixteen host ESCON channels

The ESCON connections are illustrated in the PtP VTS schematic in Fig. 2. This figure shows the interconnection scheme between ESCON host(s), the VTCs, and the VTSs. All the channels are ESCON; for the PtP VTS B20 there are sixteen channels between the VTCs and zSeries host(s), while there are eight paths to each of the VTSs. All data transfer between the two VTSs occurs via the eight paths through the VTCs. A complete set of PtP VTS configuration options is given in Table 1.

IBM TotalStorage Peer-to-Peer Virtual Tape Server

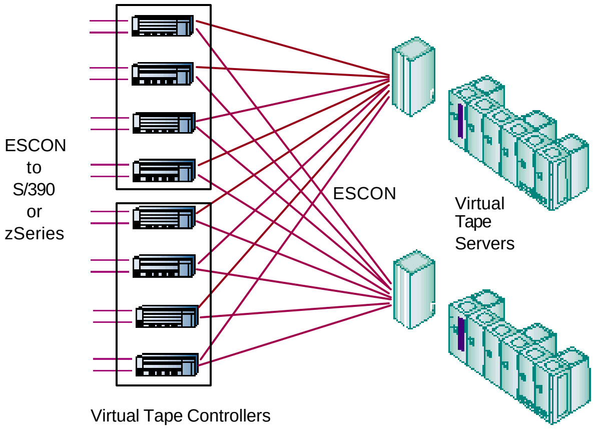

Fig. 2 *The PtP VTS B20 with a full VTC configuration, showing ESCON interconnection with the VTCs.*

Table 1. *PtP VTS Configurations**

	max				#
VTS	#	# host	# CU	# virtual	3590 per
model	AX0	ESCON	images	drives	VTS
B10	**4**	**8**	**4**	**64**	**4 to 6**
B18	**4** **8**	**8** **16**	**4** **8**	**64** **128**	**4 to 12**
B20	**8** **8**	**16** **16**	**8** **16**	**128** **256**	**6 to 12**

(*) Only specific fixed configurations are generally available

Operational Modes

The operational modes refer to how data written to the PtP VTS are handled. Data are 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 rewind/unload 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 later synchronization of the data on the two VTSs.

Immediate-Copy and Deferred-Copy Modes defined

The *immediate copy* and *deferred copy* modes are the only user selectable operating modes available on the PtP VTS. In addition to this mode selection, the observed throughput performance can depend on the initial state of the tape volume cache (TVC), the write content and compressibility of the workload, and how long the operation has been sustained. For each of the operating modes we define a *peak* throughput, observed at the beginning with either an empty TVC or one in which all new or updated volumes have been copied to the other VTS and physical tape. We define as a *sustained* throughput one that is observed after sufficient operation with a high workload; after which it can be verified that the content of the TVC is in dynamic equilibrium with the rate of copying to tape equal to the rate at which data is being written to the VTS. The sustained throughput is approximately the same for immediate copy and deferred copy operation. There can also be periods of other, intermediate, throughput in the transition from peak to sustained throughput.

Regardless of the operating mode, 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 typically be observed with the PtP VTS in *deferred copy* mode, within the constraints detailed in the section on *Peak Write Time and TVC Capacity Planning.* When the maximum write input is occurring, most of the asynchronous background copy work is suspended in order to handle read/write traffic with the host. The other extreme is the *immediate copy* mode in which both copies are made before rewind/unload complete is presented to the host. The sustained throughput performance is approximately the same in either mode.

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 operating states (i.e., *peak* or *sustained*). The data are presented as a function of compression factor since data is compressed by the VTS. Data compression has a significant effect on the observed maximum PtP VTS throughput.

All of the measurements and modeling of performance assume the maximum configuration of the PtP VTS in number of ESCON channels and number of tape drives (IBM 3590E's with 20 GB native cartridges assumed). The workload comprises up to 64 or 256 jobs (B10 or B20, respectively), writing 800 MB tape volumes simultaneously. Unless specified, the block size used is 32 KB and the BUFNO parameter in the job control 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 "mixed workloads," the latter defined below in Table 2.

A typical read/write/hit-ratio workloads defined

Table 2. *Mixed Workload Definitions*

Mixed Workload:	Fraction Writes	Fraction Read Hits	Fraction Recalls	Mounts/ Recall
Mix I	0.60	0.20	0.20	0.55
Mix II	0.55	0.405	0.045	0.3

Performance in Local Operation

In this section we present data for the case when all PtP VTS components, VTCs and VTSs, are local. Performance with any of the components remote (i.e., at greater than 1 km distance) is discussed in the section on *Remote Dual Copy Performance*.

B20

Write throughput as a function of:

- ***Data compression factor***
  - ***Mode of Operation***

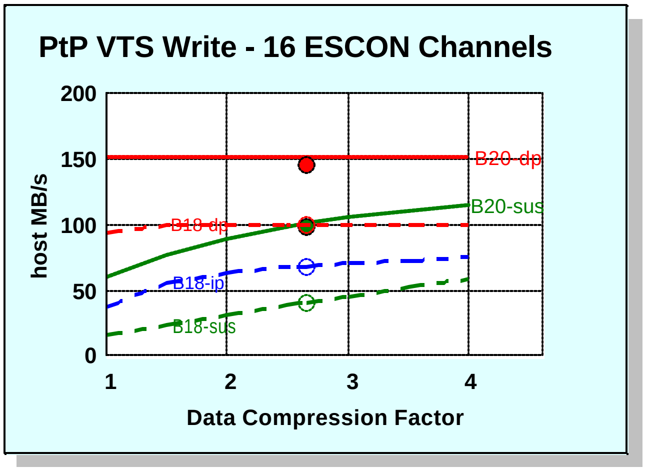

Fig. 3. *PtP VTS maximum* **sixteen channel** *throughput in host MB/s for a 100% write workload as a function of data compression factor. These are model projected data for the PtP VTS B20 and the new B18 with 16 ESCON and 12 tape drives. The points represent measurements. The labels are:* dp, *for* deferred copy *mode* peak*;* ip *for* immediate copy *mode* peak*; and* sus *for* sustained *operation. The* sus *and* ip *mode throughput of the B20, although not measured, is expected to be similar.*

Figure 3 shows the modeled performance of the PtP VTS under a 100% write workload for VTS models with a maximum of sixteen ESCON channels. The PtP VTS B20 is the newly available model, the B18 is the new release of that model with sixteen ESCON channels and twelve tape drives.

The B20 write throughput is significantly higher in all modes of operation. At a compression factor of three the *deferred copy mode* peak write rate of the B20 is about 150 MB/s while that of the sixteen channel B18 is about 100 MB/s. In *immediate copy* mode *peak* the rates are about 100 MB/s and 68 MB/s, respectively. The corresponding expected *sustained* write rate is more than double on the B20, compared with the B18.

The upper curve, marked *dp* for *deferred-copy/peak* is the performance with copies from one VTS to the other being deferred in favor of maximum host write throughput.

The next level of throughput performance is obtained in the *immediate copy* mode with the copying of data to tape not necessarily keeping up with the rate at which new data is being written to the VTS. This is the *immediate-copy/peak* mode designated as *ip* in the figures.

In the dp mode, data can build up in the TVC that need to be copied to the peer VTS, as well as needing to be copied to tape in the first VTS and then copied to tape in the peer VTS. In the ip mode the peer-to-peer copies are immediate, but there can still be buildup of data in both the TVCs that need to be copied to tape. When such data reach a fixed threshold in the TVC, the VTS begins to enforce a policy of not accepting new data until a corresponding amount of space has been released after copies have been made. This state of the VTS is called the *sustained* throughput state (designated by *sus*) in that it could be maintained indefinitely. With the PtP VTS B20, it is expected, although not measured, that the *immediate copy* mode can be maintained at its *peak* level indefinitely and is thus approximately equivalent to the *sustained* throughput state.

B10 Performance

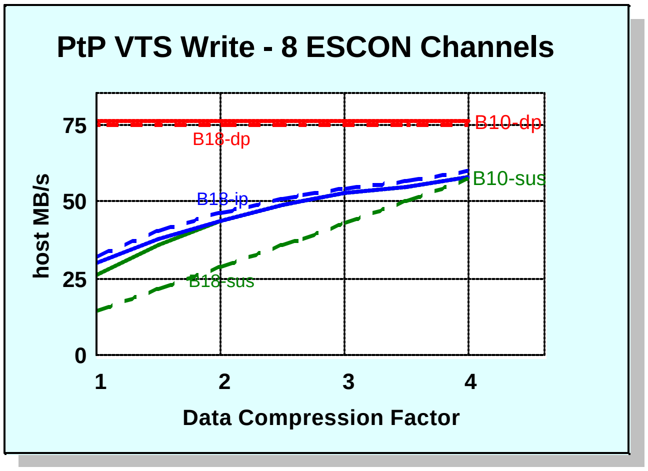

Fig. 4. *PtP VTS maximum* ***eight channel*** *throughput in host MB/s for a 100% write workload as a function of data compression factor. These are model projected data for the PtP VTS B10 and the older B18 with 8 ESCON and 6 tape drives. The labels are defined in Fig. 3. The* sus *and* ip *mode throughput of the B10 is similar except that at compression factors close to one the ip mode throughput is expected to be slightly higher.*

A similar performance comparison is shown in Fig. 4 for the newly available eight ESCON channel PtP VTS B10 and the existing PtP VTS B18 with a maximum of eight channels and six tape drives. These models have similar throughput performance except that the *sustained* throughput with the B10 is expected to be about 20% greater than with the B18 at a compression factor of three. As with the PtP VTS B20, the B10 is projected to be able to maintain the *immediate copy/peak* (= *sustained*) mode throughput indefinitely for most workloads (except for marginally compressible data).

The *deferred* and *immediate copy* modes of operating performance represent the envelope of write throughput bandwidth. In the *deferred copy* 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. For example, if the write input drops to about 105 MB/s on the PtP VTS B20 (at a data compression factor, CF=3), then the background work in host MB/s approximately keeps up with the write input rate. This is the *sustained* rate for the deferred copy mode. This is different from the *immediate 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. **Except for systems 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.**

PtP VTS B10/B20 Mixed Workload Performance

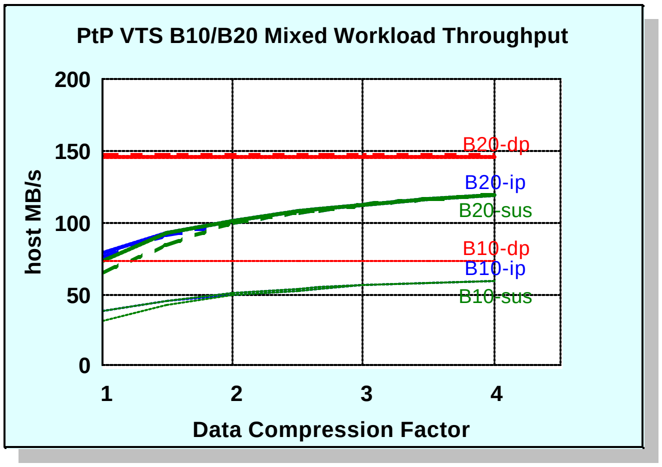

Fig. 5. *Modeled PtP VTS B10/B20 maximum throughput in host MB/s for the mixed workloads defined in Table 2. Mix II is given as solid lines while the Mix I throughput is shown as dashed lines; mostly they are very close or overlapped. These data assume that all VTCs and VTSs are local. The labels are defined in Fig. 3.*

The mixed workload performance projected from modeling is shown in Fig. 5 for the PtP VTS B10 and B20. The workloads are defined in Table 2. For applications whose data compression factor is about three, the observed throughput of the PtP VTS B20 for any given mode of operation is expected to be about twice that of the B10. Except for data with a small compression factor, the performance for Mix I and Mix II is essentially identical. Similarly, the *immediate copy* mode throughput is essentially the same as the *sustained* throughput.

PtP VTS B18 Mixed Workload Performance

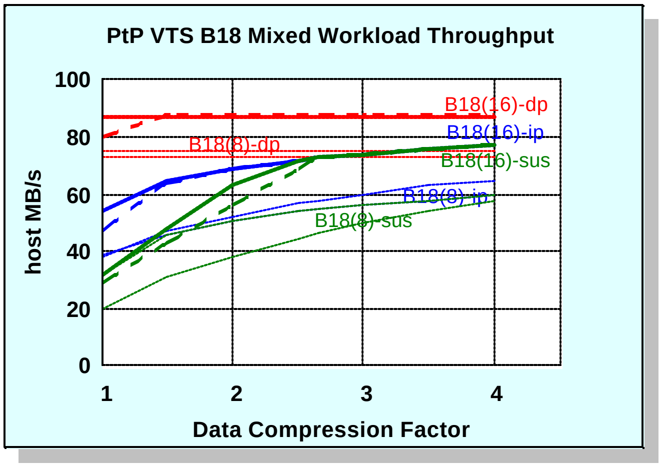

Fig. 6. *Modeled PtP VTS B18 maximum throughput in host MB/s for the mixed workloads defined in Table 2. The figure shows both the new PtP VTS B18 with sixteen host channels, designated as B18(16) and the older PtP VTS B18 with eight host channels, B18(8). Mix II throughput is given as solid lines while the Mix I throughput is shown as dashed lines. These data assume that all VTCs and VTSs are local. (The B18(8)-ip and B18(8)-sus curves for Mix II are largely coincident.) The labels are defined in Fig. 3.*

Figure 6 shows the same mixed workload performance as in Fig. 5, except for the existing PtP VTS B18 with eight host channels and the new PtP VTS B18 with sixteen host channels, the latter available as a new product or as an upgrade of an existing eight channel PtP VTS B18. The eight channel B18 throughput is comparable to the B10 with eight channels, while the sixteen channel B18 throughput is intermediate between the maximum B10 and B20 configurations.

Peak Write Time and TVC Capacity Planning

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

There are two factors that determine how long the *peak* rate in *deferred copy* mode can be sustained. One is the TVC capacity; the other the *deferred copy priority threshold* that specifies the maximum age in the TVC of data un-copied to the peer VTS (in integral 0 to 24 *hours*). When this *hours* age has been reached by a tape volume, its priority for being copied is increased.

Once the TVC capacity threshold triggered peer-to-peer copy process begins, the PtP VTS is functioning effectively in the *sustained* mode; namely, the rate at which data is copied between peers and to tape has to be occurring at least at the rate at which write data is coming in from the host.

In addition, if the *deferred copy priority threshold* is reached it is possible for the *deferred copy* mode throughput to dip below the *sustained* rate. This is because the rate of data reaching this threshold can exceed the peer-to-peer copy requirement in the *sustained* state. Once the backlog of un-copied data has been worked through, the *deferred copy* mode write throughput will rise to the *sustained* rate.

If a management goal for the PtP VTS is to have maximum throughput available on demand in *deferred copy* mode, the *hours* parameter should be kept large enough so that most peer-to-peer copies occur naturally before reaching the *deferred copy priority threshold*. Thus the *hours* parameter should be greater than the expected daily peak period duration.

How long can I run in maximum throughput deferred-copy write mode ?

For the PtP VTS B20, initial tests at a data compression factor of 2.66 indicate that the ***deferred copy peak*** **write rate can be maintained for about seven hours** from an empty 864 GB TVC (i.e., fully copied) state, and that **the maximum write throughput after the TVC full condition is reached is approximately 20 percent below the *sustained* rate (about 80 MB/s)**. It is estimated via modeling that if the PtP VTS B20 is operated at the maximum *deferred copy peak write* rate (100% write, compression factor three) for eight hours in a twenty-four hour period, then the average write demand during the remainder of the twenty-four hour period should average no more than about 75 MB/s for the new cycle to begin with all peer-to-peer copies done.

Similarly, the peak period with a 1.7 TB TVC, although not measured, is expected to be approximately twice as long, and requires that there be essentially zero host demand on the PtP VTS for the remainder of the twenty-four hour cycle for the next cycle to begin with no uncopied data.

Projections indicate that the PtP VTS B10 with a 432 GB TVC should have approximately a six hour *peak deferred copy* write period (at a compression factor of three) to the threshold for the maximum of uncopied data; about 2.5 hours with a 216 GB TVC. The maximum average write rates during the remainder of the twenty-four hour cycle are approximately 45 and 50 MB/s, respectively.

All of the data and recommendations in this section should be taken as planning guideline approximations. It is suggested that a safety factor be included as some variation from these numbers can be expected with usage patterns and the specific workload. This does not create a warranty or guarantee of actual 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.

Remote Dual Copy Performance

The performance cited in the prior section 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 the 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 (1), a common remote configuration (2), and a configuration with most of the hardware “local” except for one of the VTSs at a disaster recovery site (3). (In config. (3) VTS-B may be configured with four standby AX0 VTCs at the remote Location.)

The underlying consideration 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. This delay is augmented by the fact that periodically the transmitter has to wait for an acknowledgment from the receiver that a data block has been received without error. Thus, for every block transmitted, there is a round trip delay before the next block can be sent. The resulting ESCON data rate is determined by a relationship 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 is user selectable except the choice of distance and channel extender, when required.

Remote Operation Configurations

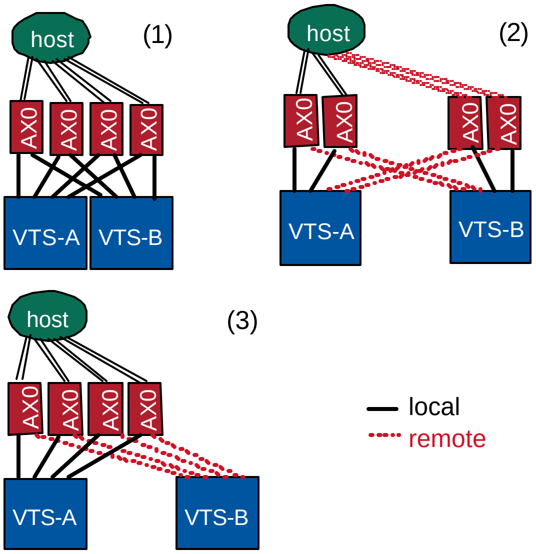

Fig. 7. *Common configurations for an* ***eight channel*** *PtP VTS. A sixteen channel PtP VTS would have twice as many AX0s, connected similarly. All remote distances are assumed to be mediated via IBM 2029 Fiber Savers.*

Another factor that affects performance at a distance is the mode in which the VTCs are set. They can operate in *preferred* VTS or *no preference* mode, defining which of the peer VTSs they send an host I/O to. In *preferred* mode the host I/O is sent to the VTS specified, usually the local one (there is no guarantee, however, that the I/O will actually occur at the VTS specified). In *no preference* mode, the VTC sends the I/O to the VTS having the smaller number of I/Os outstanding; in effect, it tries to balance the load on the two VTSs.

The choice of *no preference* or *preferred* VTS mode is made at the VTC. It is a static choice requiring a power-down of the VTC.

The following remote VTS performance projections are based on transmission characteristics obtained from a measurement at a distance of 25 km, where the remote connection between the AX0 VTCs and the VTSs is via IBM 2029 Fiber Savers. The configuration used was (3) in Fig. 7

The principal result is for configuration (3) in which the workload is balanced over four AX0 VTCs. Functionally the VTC/VTS operation of configurations (2) and (3) are equivalent. In configuration (2) the workload (i.e., the number of I/O per second issued to the local and remote VTCs) can be skewed if the host to VTC distance is large enough (more than about 10 km for *deferred copy* write). It is only in the fact that configuration (2) can have an input skew between the local and remote VTCs that the configurations differ in performance.

The remote configurations in Fig. 7 can also be achieved using ESCON Directors (IBM 9032) instead of the IBM 2029 Fiber Savers. The former extended distance operation is possible up to 26 km, while the latter can operate at a distance up to 50 km. Their performance is similar over their common range.

The elapsed time, as viewed from a VTC, is generally shorter for an I/O operation on the local VTS than on the remote VTS. This *response time* advantage will be apparent for read and deferred copy mode write I/Os if they are directed preferentially to the local VTS. However, as described below, operation in the *preferred* VTS mode will generally result in overall reduced *throughput* performance.

Effect of extended distances on remote dual copy operation

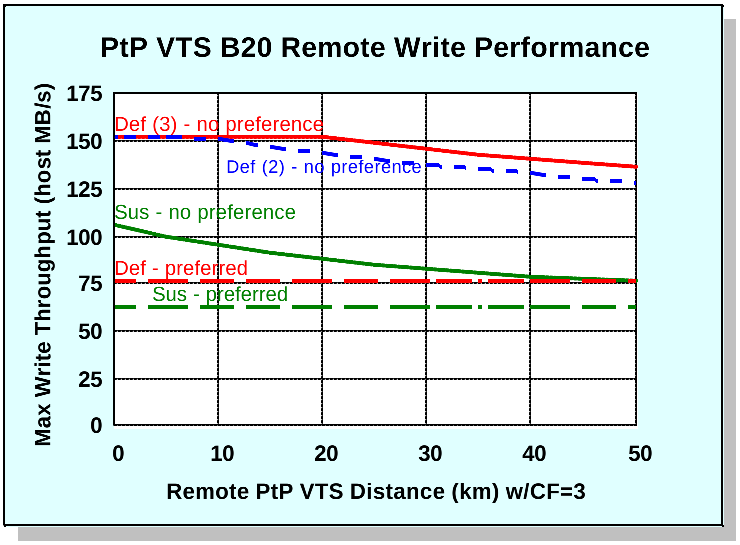

Fig. 8. *The modeled remote* sustained *and* deferred copy *mode write throughput performance of the PtP VTS B20 using the IBM 2029 Fiber Saver or IBM 9032 ESCON Director. For* "deferred copy *mode* - no preference" *the numbers (2) and (3) refer to the configurations in Fig. 7. This figure applies to data with a compression factor of three.*

The at-distance PtP VTS write throughput performance is shown in Fig. 8 for the PtP VTS B20 and in Fig. 9 for the PtP VTS B10. The characteristics shown apply to both configurations (2) and (3), except for the *deferred copy* write mode where throughput curves for both configurations are shown. All of the modeling has assumed 32 KB blocking and a BUFNO=20 (a smaller BUFNO will yield a somewhat smaller throughput rate).

The throughput of the *no preference* mode is always higher than that of 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 a VTC can leave one path to a VTS underutilized while the other is operating at maximum throughput, for example.

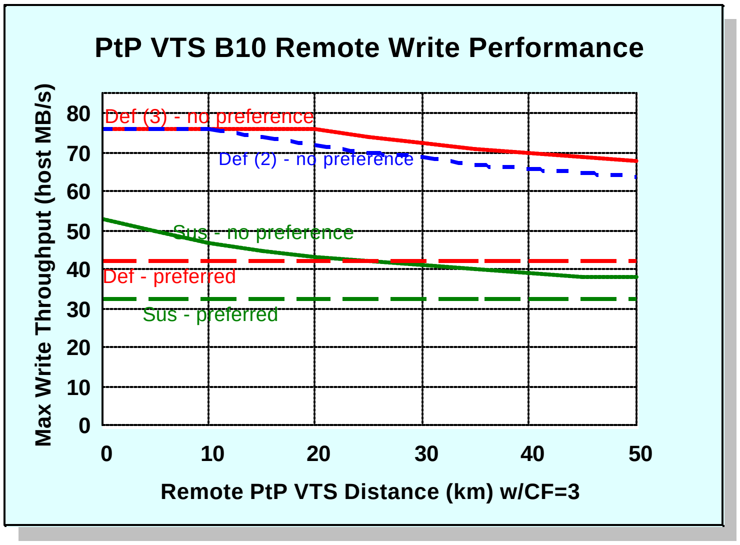

Fig. 9. *The modeled remote* sustained *and* deferred copy *mode write throughput performance of the PtP VTS B10 using the IBM 2029 Fiber Saver or IBM 9032 ESCON Director. For the* deferred dopy *mode* no preference *the numbers (2) and (3) refer to those configurations in Fig. 7. This figure applies to data with a compression factor of three.*

The *no preference* mode throughput is reduced and becomes 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 principally by the local ESCON paths, which remain constant in length. They have about half the throughput of the *no preference* mode because each VTC essentially has only a single ESCON channel to the *preferred* VTS.

The reason the *no preference deferred copy* mode write performance is lower for configuration (2) than for configuration (3) is that the host to VTC ESCON distance causes the input to the remote VTCs to fall below that which those remote VTCs and VTS can handle.

As a result of the performance characteristics exhibited in Figs. 8 and 9, **the general recommendation is to operate the PtP VTS in *no preference* mode.** Only in cases where the 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.

Note that in the "Sus - no preference" 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 VTC, 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 *preferred* mode leaves the extended distance ESCON channels underutilized at distances within about 25 km.

Performance of Mixed Workload in remote dual copy operation

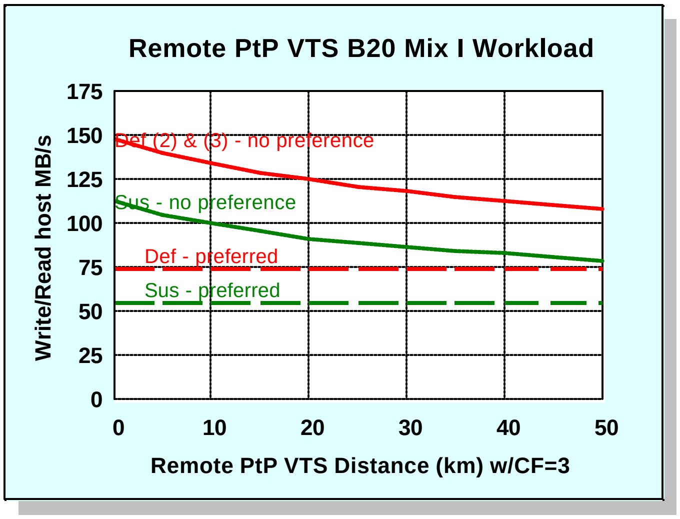

Fig. 10. *The remote Mix I (60% Write, 40% Read) throughput performance of the PtP VTS B20 in* sustained *and* deferred copy *mode using the IBM 2029 Fiber Saver / IBM 9032 ESCON Director. For the* deferred copy *mode* no preference *the numbers (2) and (3) refer to those configurations in Fig. 7. The remote throughput for Mix II (55% Write, 45% Read) is similar with distance.*

For practical workloads that involve a mix of read and write I/Os we have modeled the throughput performance of the **Mix I workload** (60% writes, 50% read hits). These results again show performance reduction for configuration (2) as shown in Fig. 10. 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.

Mixed PtP Configurations

The standard PtP VTS configurations are symmetric; that is, they comprise two B10s, B18s, or B20s. The licensed internal code, however, does not preclude some mixed configurations. A number of mixed configurations, the specifics available from IBM tape storage specialists and business partners, are supported. The performance of such configurations can be evaluated on a case by case basis.

What about mixing B10s, B18s, and B20s in a PtP VTS?

For example: If a current installation has a B18 VTS and were to purchase a B20 VTS, could they be configured as a mixed PtP VTS B20/B18? The answer is yes, but the problem is that the performance for much of the time would be equivalent to that of a PtP VTS B18; except for the *deferred copy* write period, and the recovery period would be extended in time.

However, there are situations where such mixed configurations might make sense. For example, if the intention is to run the PtP VTS in preferred mode, then with the B20 local and the B18 (with eight ESCON) remote the projected performance is estimated to be approximately equivalent to a symmetric PtP VTS B20 (assuming a compression factor on the order of three). That is, in the preferred VTS mode, the B18 can approximately keep up with the task of handling copies from the B20 in sustained operation.

Single Host-Job Performance

Single host job throughput can be significantly lower than the aggregate rate for multiple jobs

Throughput is generally defined in terms of the aggregate number of MB that can be written and read by a large number of jobs acting concurrently during a period of time; and reduced to MB/s or GB/hr. This aggregate throughput capability is not directly useful for estimating the data transfer rate of a single job because a single job can only use one host channel, passes through one AX0 VTC, and can only use one VTS to handle the primary input. In addition, while with parallel processes a given resource can be kept near 100% busy, with a single job some resources are utilized to process a packet at a time with a latency period between packets. For this reason, it is sometimes useful to know single job data rates to estimate job completion time.

On writes, in Fig. 11, we see that single job throughput is unaffected by distance up to 25 km as long as the channel extension (Fiber Savers or ESCON Director) is between the host and AX0 VTCs. Putting the channel extension between the AX0s and B18 causes the PtP VTS single job throughput to decrease systematically with distance.

For reads, having the channel extension between the host and AX0 VTC also has minimal throughput impact at 5 km. However, as the distance increases the read throughput is affected significantly more than the write throughput. By the time the host to AX0 VTC channel extension reaches 25 km, it has (slightly) more impact on read throughput than having the extension between the AX0 VTCs and the B18 VTSs.

Single host job read/write throughput in local and remote operation

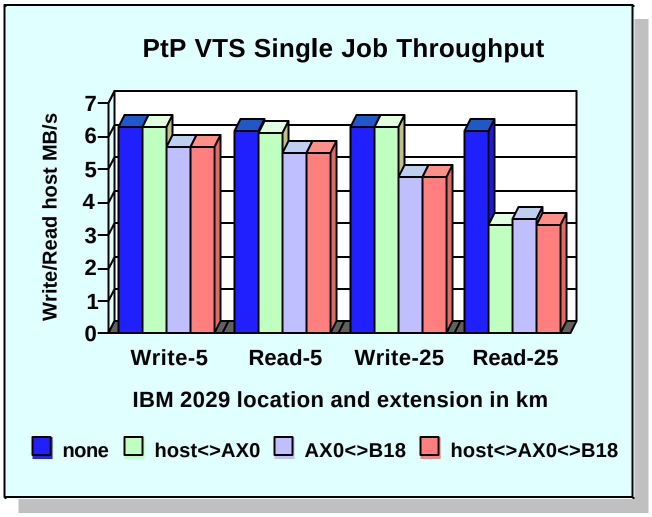

Fig. 11. *The measured single job write or read throughput on a PtP VTS in a remote configuration at 0, 5, and 25 km IBM 2029 extension. The lozenge in each configuration description in the legend represents the location of the IBM 2029. The zero distance ("none" = no IBM 2029) throughput is shown at the left of each configuration/distance group. The data are for 32 KB blocking and BUFNO=6. The data used had an average compressibility of 2.66.*

Support of Remote Operation

Current product support for extended distance (remote) operation with Fiber Savers is for up to 50 km (26 km using ESCON Directors) within the PtP VTS (i.e., between AX0s and B10, B18, or B20s) and up to 75 km between the host and AX0 VTCs using IBM 2029 Fiber Savers (43 km with ESCON directors). Configurations over longer distances require the use of channel extenders such as the products of INRANGE® and CNT®.

Generically there are two types of such extension devices: (1) signal repeaters and (2) channel extenders.

Some details on Remote Operation

Signal repeaters take an incoming signal, amplify it, and send it along in the direction it was propagating. There is minimal propagation delay through the repeater, but the full round trip propagation delay is experienced by the sender before an acknowledgment is received that a block of data arrived without error. At extended distances this delay per block can significantly affect the data rate of the channel. This kind of channel extension technology has been found to be acceptable in some applications (at reduced throughput) for distances up to about 100 km. An example of such a signal repeater for ESCON channel extension is the IBM 2029 Fiber Saver. (Note that for PtP VTS operation at over 50 km distance more is required than simply a minimum acceptable data rate. The synchronization timing between the peer VTSs also puts requirements on maximum signal delay.)

A channel extender works in pairs of identical hardware devices. One device is at the sending location and the other is at the remote location. Each device buffers incoming data and immediately returns an acknowledgment to the local ESCON connected to it. The communication over such an extension appears to be local. Of course, the extended distance link between the pair of channel extenders still introduces response time delays. However, as long as the distance link bandwidth is adequate, it can appear as if the ESCON channel has been extended without affecting its bandwidth. This benefit, although not measured, is expected to extend at a reduced level even to single job throughput. Such extension can, in principle, occur over continental distances (say 3000 miles).

At this time, using channel extension technology within the PtP VTS other than the IBM 2029 to 50 km (or IBM 9032 to 26 km) is considered to be a custom installation not covered by the standard installation agreement.

A number of effects can reduce the PtP VTS throughput if the tape volumes are small

Small Volume Effects on Throughput Performance

The performance information presented up to this point has been based on measurements performed with full 3480E logical volumes (800 host MB), and 250 host MB logical volumes in the modeling for the Typical Mix workload. There are special performance considerations that need to be made if the average volume size in the workload is smaller.

There are three VTS internal per-volume overheads that become significant for small volumes. All host volume sizes are quoted in MB before a data compression by a factor of three is applied (CF=3):

- There is about a 1.5 second Library Manager overhead in cataloging volumes in the tape library. This translates to a rate of about 2400 volume mounts/hr or a minimum host volume size of about 114 MB required to achieve the maximum PtP VTS write throughput.
- There is a fixed overhead associated with freeing space of the volumes in the TVC which have been copied to physical tape. For small volumes, on the order of 100 host MB or less, the rate at which this can be done will limit the throughput of the VTS.
- There is a latency associated with the beginning of data transfer in copying volumes from the TVC to physical tape. The effect is that tape volumes need to be at least 300 host MB in size (or 100 host MB in size at CF=1) to achieve the maximum sustained 100% write throughput described earlier in this paper. For example, if the host volume size is reduced from 300 MB to 150 MB the sustained write throughput is reduced to about 60% of maximum.

Performance Tools

Help in capacity planning and resolving performance issues

Tape Magic is a high-level tape subsystem configurator available to IBM customer representatives and business partners 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.

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 tape specialists and business partners, is called **Consul Batch Magic (CBM)** 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. IBM storage specialists have access to CBM.

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.

The **IBM StorWatch Expert for Enterprise Tape Library (ETL Expert)** provides asset, capacity, and performance reporting for the IBM TotalStorage tape library solutions :

IBM TotalStorage 3494 Tape Library,
IBM TotalStorage Virtual Tape Server,
IBM TotalStorage Peer-to-Peer Virtual Tape Server.

The IBM StorWatch Expert is a program product in the IBM StorWatch software family. It helps in the management of the Enterprise Storage Server (ESS) and the Enterprise Tape Library (ETL) using a Web bowser interface. The StorWatch ETL Expert provides a single Web-based console to monitor all 3494 tape libraries, VTSs, and PtP VTSs in the enterprise anywhere, anytime, anyplace. The ETL Expert helps answer questions such as: How much free space is there? What tape drives are available? What is the cache miss percentage? In addition it provides a Health Monitor which takes a heartbeat of the tape libraries every 10 minutes. The ETL Expert monitors twenty-two key indicators of tape library performance (i.e.: average virtual mount time, overall throttling value, etc.) for which thresholds can be set to fit the particular tape environment. The ETL Expert takes over the monitoring task and issues an alert when the set thresholds are exceeded.

Conclusions

The PtP VTS, featuring a new level of data protection, has an ancestry of continuous performance improvement

The virtual tape server, beginning with the VTS model B16, has addressed 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, offered significant improved throughput performance. The first Peer-to-Peer VTS was built on the B18 base, offering 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. The current release of the PtP VTS B10 and B20 and the enhancement of the B18 extend the PtP VTS throughput performance by up to approximately a factor of two. The B10 and B20 models also introduce a new hardware architecture base with more powerful processors and enhanced I/O connectivity. 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.

References

1. IBM Corporation, “IBM TotalStorage Peer-to-Peer Virtual Tape Server Performance with Model B18 VTSs,” 30 May 2001. (PtP VTS 5301.pdf)
2. IBM Corporation, "IBM TotalStorage Virtual Tape Server Performance," 28 Aug 2001. (VTSperf8281.pdf)
3. IBM Corporation, “IBM Magstar 3494 Virtual Tape Server Performance White Paper, v. 4.0,” 28 July 2000. (VTSpwp40.pdf)

Acknowledgments

The authors thank Dennis Haight and Christine Hong for their contributions to the performance tools section. The comments of the reviewers Joseph Metzger, Barbara McNaughton, William Greeley, and Jay Hayes are gratefully acknowledged.

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.

The use of this information or the implementation of any of these techniques is the reader's responsibility and depends on the reader's ability to evaluate and integrate them into their operating environment. Persons or entities attempting to adapt these techniques to their own environments do so at their own risk.

References in this document to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM program product in this document is not intended to state or imply that only IBM's program product may be used. Any functionally equivalent program may be used instead.

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

IBM Storage System Group
9000 S. Rita Road
Tucson, AZ 85744