IBM Magstar Model B16

Virtual Tape Server

Elements of Performance

Revised 7/10/98

IBM Performance Laboratory, Tucson
IBM Advanced Technical Support Center, San Jose
December 19, 1997

Introduction

Purpose of this Paper

This paper is a collaborative effort to assist IBM Field personnel and IBM customers in understanding the underlying elements that determine the performance of installed Magstar Virtual Tape Server subsystems. After reading this paper, the user should be able to:

1) describe the elements of the Magstar Virtual Tape Server (VTS) that are shared by the tasks running in the VTS controller
2) use the contents of the Type 94 SMF record to track the performance of the VTS
3) recognize the symptoms that indicate the VTS is at or near its maximum capacity
4) understand the options available to improve the throughput or performance of the VTS

This paper is distributed with the intent that it should be shared with anyone who is involved in the marketing, support, and operations of a Magstar Virtual Tape Server.

This paper applies to the 3494 Model B16 Virtual Tape Server.

The algorithms used in the VTS and described in this document are subject to change without notice.

Questions arising from reading this paper should be directed to Carl Bauske, Advanced Technical Support Center for Storage Products, Internet id cabauske@us.ibm.com, Lotus Notes id Carl Bauske/Princeton/IBMUS or J.D. Metzger, Internet id jdmetzg@us.ibm.com, Lotus Notes id J Metzger/San Jose/IBMUS.

VTS Unique Attributes

Non-Virtual vs. Virtual Tape Subsystem Performance

Historically, basic benchmark performance data on non-virtual tape subsystems was measured by running a series of well-defined workloads on a configuration consisting of a number of tape drives attached to a tape control unit. The benchmark configuration has typically been a control unit with one “control unit function” with a maximum of 16 tape drives, i.e., a “1X16” string. The workload that was usually considered to be the most definitive in expressing the maximum throughput of the subsystem was a job stream of multiple concurrent dump tasks. The “power” of the tape subsystems was expressed as the MB/sec that the subsystem could sustain when running various numbers of dump tasks. In the 3490E and earlier systems, the limit to performance was the ratio of control unit functions to drives, after the 2nd or 3rd active drive. Tuning for performance meant lowering the number of drives per control unit, or implementing “short strings”. This approach to tape performance was further encouraged by the 3490 C models, which, by design, limited the number of drives that could be attached to the control unit. The 3590 A00 and A50 continue this high performance approach of short strings.

A virtual tape subsystem, however, requires that we take a new view of performance. The Magstar VTS introduced several new elements into the performance equation, including Tape Volume Cache (TVC), a VTS controller running AIX and storage management software, virtual tape drives, virtual and logical volumes, etc. This new architecture can provide significant benefit in the tape processing environment, but a new approach must be taken in order to effectively size, monitor, and manage the performance of the VTS.

Seascape Architecture

The 3494 model B16 VTS is an IBM Seascape Architecture product. Seascape development processes enable products to be brought to market more quickly and at a lower cost by making use of off-the-shelf hardware and software components and Licensed Internal Code (LIC).

In the VTS, these off-the-shelf components are the RS/6000 Processor, the 3590 Magstar 128 track Tape Drive, the 7133 SSA Disk Storage System, the 3494 Automated Tape Library, the AIX operating system and storage management software. Figure 1 is a logical diagram of these components and the interfaces between them:

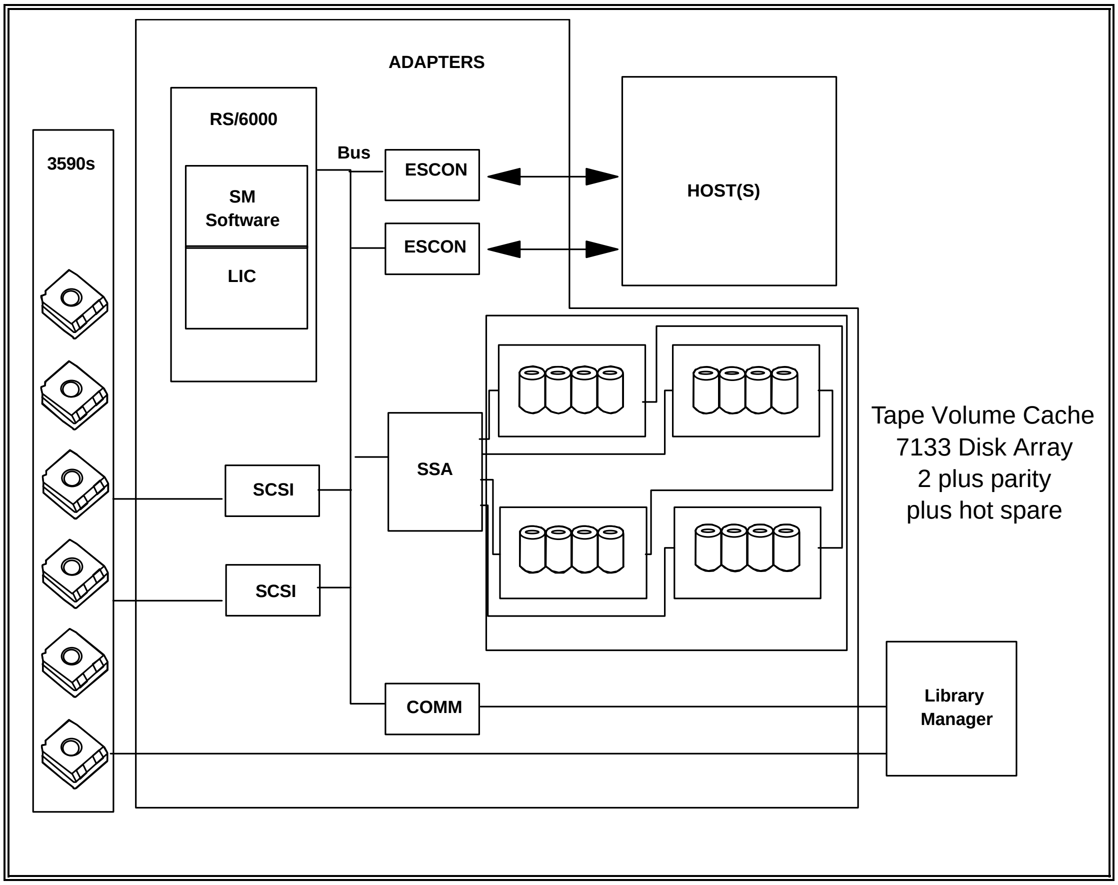

Figure 1.

Shared Resources

In the process of copying and recalling virtual volumes on stacked physical volumes, components are shared by tasks running in the VTS controller. Some tasks represent customer work and other tasks are associated with the internal operations of the VTS. All of these tasks must share the same resources, especially the RS/6000 processor and the 7133 disk arrays. Contention may occur for these resources when heavy demands are placed on the VTS subsystem. To manage the use of shared resources, the VTS uses various resource management algorithms which can have a significant effect on the level of performance achieved for a specific workload.

VTS Subsystem Performance Metrics

Channel Throughput

Factors that Influence Data Transfer

Channel Throughput is the most-often cited performance specification for tape subsystems. In the VTS there are many factors which affect channel throughput, most significantly Tape Volume Cache (TVC) space management.

The path for customer data written to the VTS includes the ESCON channels, and the RS/6000 bus, processor and storage. The instantaneous data rate for this whole path is a maximum that is much higher than the sustained data rate.

All tasks running in the VTS RISC controller require a share of processor cycles. These tasks include the emulation of each virtual drive in use, each copy task, and each recall task. If there are 8 active virtual drives, 2 copy tasks and one recall, that is a total of 11 tasks. The processor cycles would have to be shared by all of these tasks through a time-slicing multiprocessing algorithm.

All virtual volumes are written and read by the host into and out of the tape volume cache (TVC). The processes of copying virtual volumes to stacked tape and the recall of volumes from stacked tape to the TVC is transparent to the host. When a virtual volume is written by the host, either from beginning of tape or for file extension, it will be added to the queue of volumes to be copied to a stacked volume after a delay of 4 minutes from the time the volume is closed.

Write operations for new volumes and appending to old volumes, and read operations requiring the recall of a volume into the cache, require free space in the TVC for storage of data. When free space becomes small, or the space occupied by volumes which are closed and ready to be copied to stacked volumes (copy queue) becomes large, the VTS will increase the number of copy tasks allowed. Also, during periods of low host use of virtual tape drives, the VTS will take advantage of available processing power and copy volumes from the cache to stacked volumes. In addition, if free space becomes small, the VTS will reduce the number of recall tasks allowed, thus making more 3590 drives available for copying. Copied volumes are eligible for immediate *fragmenting*, which is the process of reducing the data portion of the virtual volume in the cache down to a small “fragment”, which contains information such as headers. Fragmenting creates the TVC free space required for new volumes to be written and old volumes to be recalled.

In order to make processing cycles available for cache management tasks such as copying volumes, recalling volumes, and fragmenting volumes, the completion of host write operations is delayed (throttled) when necessary. Throttling is described in a later section of this paper. The available 3590 drives are also managed for the mounts required for copying to stacked volumes, recalling volumes into the cache, and reclaiming stacked volume space for volume data which is no longer active.

Customer Sample Charts

Figures 2 and 3 are charts constructed from customer SMF type 94 data and illustrate how the Channel Throughput can fluctuate and how it can be limited by high sustained activity levels.

Figure 2 illustrates a 24 hour period *without* any apparent slowdown of throughput caused by either serious contention for resources or free space throttling due to a cache full condition. The pattern of sustained workload that would fill the cache IS NOT present.

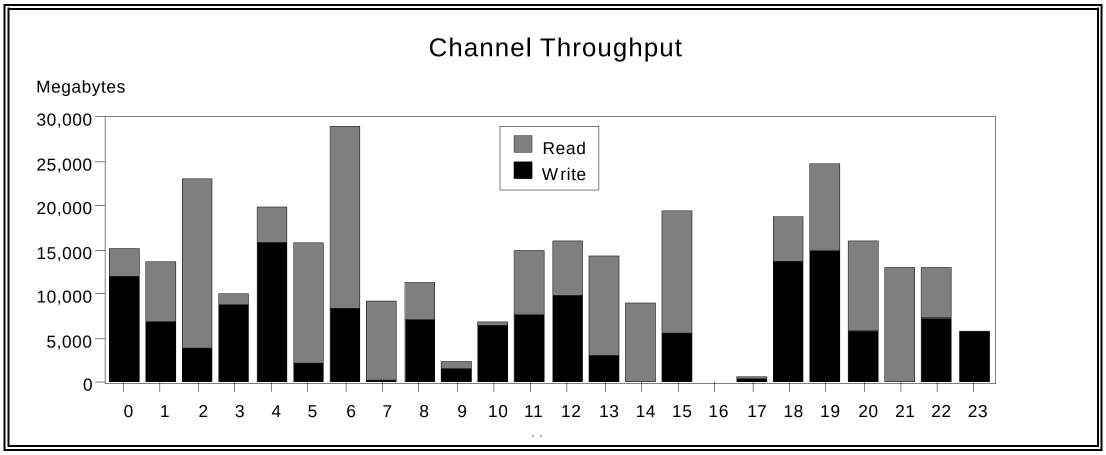

Figure 2

Figure 3 illustrates high batch window throughput demands starting at hour 19. The first 2 hours (19-20) show a greater throughput than the next 3 hours (21-23). This pattern is indicative of a full cache situation resulting in free space throttling. Note that hours 0-2 are probably the tail-end of a high-demand period started on the previous day. The pattern of sustained high workload over multiple hours IS present.

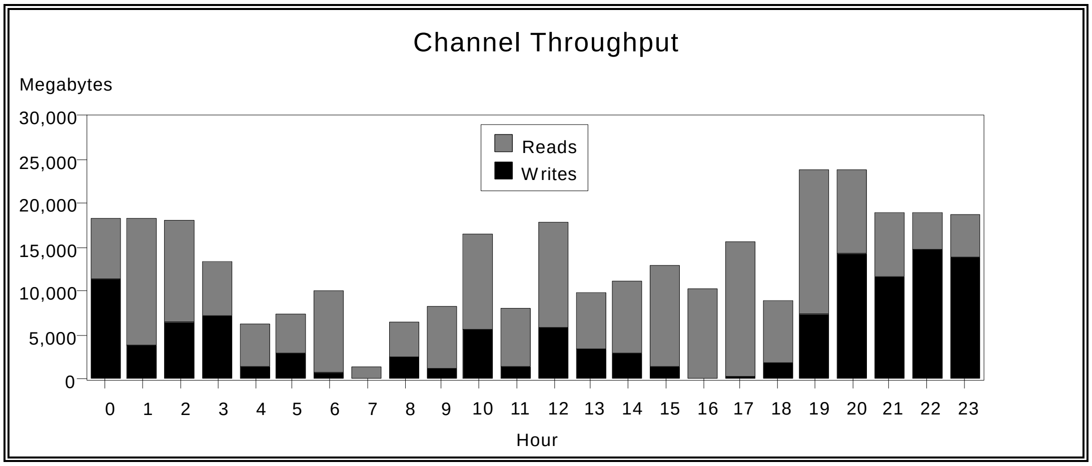

Figure 3

Virtual Mount Time

Average and Maximum Mount Time

The SMF type 94 record contains hourly statistics for virtual mount times, measured from the time the mount order is received until the not-ready-to-ready interrupt from the virtual drive is sent to the host signaling that the volume is mounted. These statistics are expressed as Maximum Mount Time, Minimum Mount Time, and Average Mount Time. Minimum Mount Time is not interesting because it is always 1 second. Fast Ready mounts and TVC hits for specific mounts are very fast. The values of interest are Maximum Mount Time and Average Mount Time for each hour. It is important to remain focused on the Average value because it is the more significant indicator of subsystem performance. Often too much attention is paid to the maximum value for each hour. In a complex queuing/priority algorithm there are always anomalies that deviate significantly from the mean, but we will discuss the use of both values.

Components of Virtual Volume Mount Time

The following figures will help to illustrate the various virtual mount scenarios.

Figure 4 illustrates a scratch mount using a FAST READY category. Two cases are shown: 1) the first use of a virtual volume, and 2) the re-use of a virtual volume, when mounted scratch. The difference being that on first use, the VTS will build the VOL1, HDR1, and initial tape mark, and on subsequent *scratch* mounts for the same virtual volume the VTS will use the existing "fragment" in the TVC, in order to avoid recalling the tape contents from stacked tape. In both cases, these mounts generally take 1-3 seconds to complete. They are never delayed by VTS resource management.

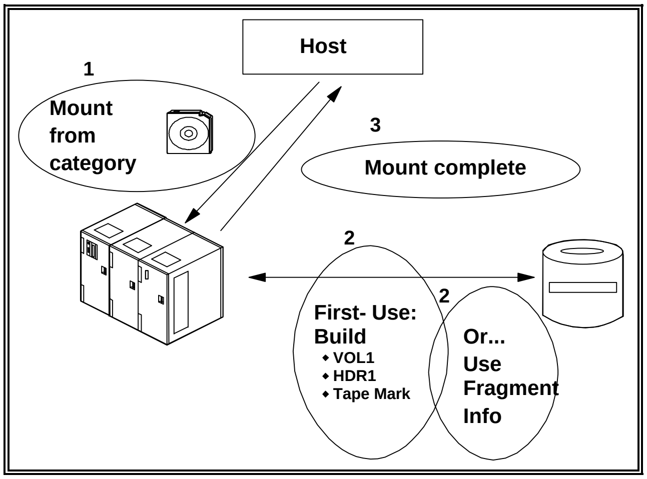

Figure 4

The more of these mounts that are done in an hour, the better the average mount time. These are the fastest mounts in the industry...faster than robotic mounts or mounts from a cartridge loader.

Figure 5 illustrates a specific mount request that is a *cache hit.*

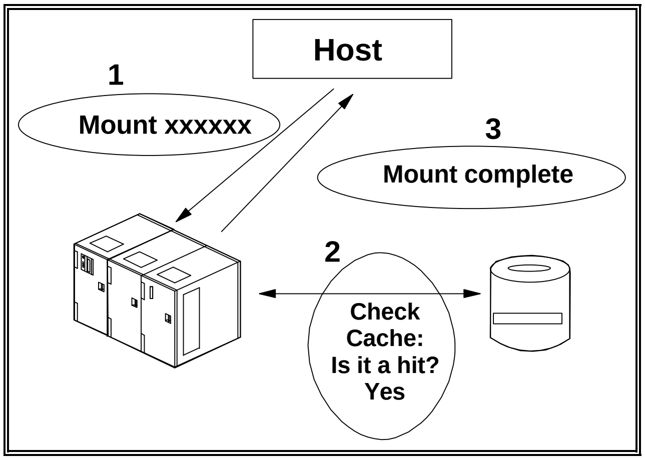

Figure 5

These mounts generally take 2-3 seconds to complete.

The more of these mounts there are in a given hour, the lower the average mount time will be. These mounts are not delayed by VTS resource management.

Figure 6 illustrates a specific mount request that is a *cache miss.*

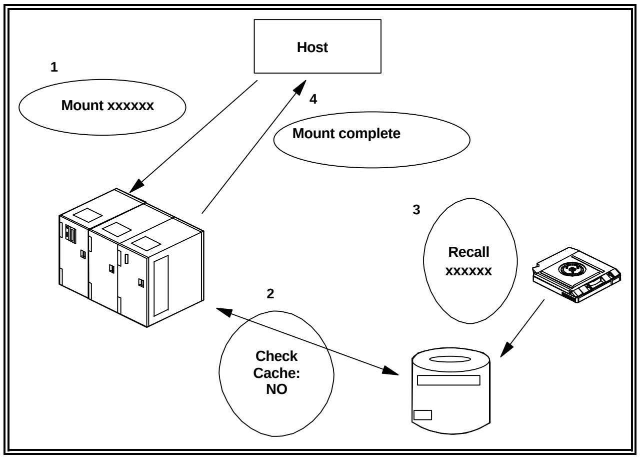

Figure 6

These mounts typically take 2-3 minutes. The two major portions of this time are the physical mount of the stacked volume in the 3590 drive, and the locate/recall of the virtual volume from the stacked volume back into the Tape Volume Cache. The *entire* virtual volume must be back in the cache before mount complete is signaled.

These mounts can be significantly delayed by one or more of the following factors:

- If there are no 3590 drives available, the recall mount will be queued until a drive becomes available.
- If the virtual volume to be recalled resides on a stacked volume that is in use by a copy task or another recall task, the recall mount will be queued until the copy or other recall task completes.
- If the virtual volume to be recalled resides on a stacked volume that is in use by a reclaim task (target or source), then the mount will be queued until reclamation completes processing the current logical volume , after which the reclamation task will be terminated, and the recall mount processed.

Because there is virtually no limit to the depth of the queues used for this operation, the length of the delay is limited only by how fast mount orders are sent to the VTS which are cache misses, and the number of virtual drives being used. Many virtual drives performing read/write operations which require recall from stacked volumes in use by other tasks will result in long virtual mount times.

Factors that influence mount time

Jobs requiring access to multi-volume data sets are likely to be involved in or cause delays. It is recommended that this type of work be directed to 3490E or 3590 native drives, rather than the VTS. If there are a limited number of these mounts required and some mount delay can be tolerated, then they can be directed to the VTS.

It is possible for the jobs that have access to the VTS to issue specific mounts very quickly, and if many specific mounts that all require a recall of a virtual volume are received in rapid succession, large queues can develop. These queues are against the resources of physical drives and/or stacked volumes. The length of time it takes for a recall task is unlimited, and can result in specific mounts of virtual volumes that take many minutes. This is normal for any hierarchical storage management system and should be expected when this type of rapid-fire mount requesting is part of the VTS workload.

Low numbers of scratch stacked volumes can also influence mount time, because of the increased probability of reclaim activity. By maintaining a scratch level of at least 50 scratch stacked volumes, the user can specify to the Library Manager times of the day when there can be no reclaim tasks started.

There is a strong correlation between the number of virtual mounts in an hour that are cache misses and the maximum mount time for that hour. These metrics go hand-in-hand. Both are statistics reported in the SMF type 94 record.

Customer Example Charts - Virtual Mount Time

The following charts illustrate the use of average and maximum mount times.

Figure 7 is a chart of virtual mount times showing Maximum and Average values.

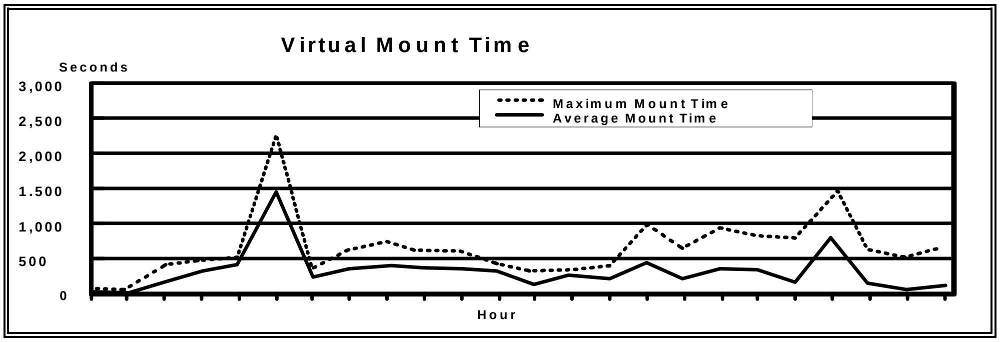

Figure 7

This is for one day in which the VTS was heavily loaded with recall-inducing cache misses. As can be seen, the maximum mount time can be significantly extended by queuing against the resources of physical drives and stacked volumes.

Figure 8 compares a similar day as Figure 7, but showing the relationship of maximum mount times to the number of virtual mounts that require recall. (The vertical scale is arbitrary.)

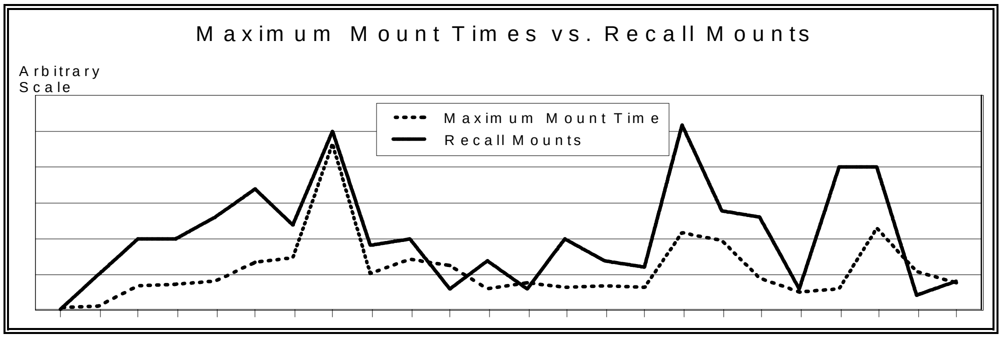

Figure 8

The chart shows a strong correlation between the two values. The SMF type 94 records contain the data required to produce such a chart.

Disconnect Time

Elements of Disconnect Time

Disconnect time is reported by RMF and not by the SMF type 94 records from the VTS.

Disconnect time is the time spent by the control unit/device in order to perform an I/O, which does not include the time the channel is connected to order the I/O or to transfer data. Disconnect time is the time it takes for the subsystem to prepare for data transfer.

In non-virtual tape subsystems, disconnect time has been a non-issue in performance except in the case of TWI (Tape Write Immediate) operations or during long running motion commands such as forward space file. It is largely the same for the VTS, except for the effects of throttling. High average disconnect time is an indicator of heavy throttling on a VTS.

Throttling in the VTS

Throttling is a delay on host write I/O completion introduced by the VTS to reduce the processor cycles required for I/O. Throttling values are calculated by the VTS for each of the following:

- Amount of free space in the TVC
- Number of recalls in process
- Size of the copy queue
- Number of entries in the copy queue

Throttling is calculated at regular intervals and the largest of the values determined for each of the factors above is used. Of the various throttling factors, TVC free space has the most effect. Each of the factors may have a calculated value of 0.

Free space Throttling: The amount of free space throttling is inversely proportional to the amount of TVC free space remaining, and at very low levels of free space can result in write activity coming to a virtual standstill, causing jobs to be swapped out. In addition, free space throttling will reduce the number of drives available for recalls, down to a minimum of one drive, in order to make drives available for copy.

Recall Throttling: the VTS will introduce host write throttling whenever recalls are in progress. The amount of delay is proportional to the number of recalls currently active.

Copy Queue Throttling: the VTS will introduce host write throttling if either the number of copy tasks, or the amount of data waiting to be copied exceeds certain thresholds.

VTS Performance Management

Monitoring Performance

SMF Type 94 Hourly VTS Statistics

SMF type 94 records are generated by all logical libraries including the VTS. They are sent to all attached hosts every hour on the hour (Library Manager clock). The type 94 record has VTS-unique fields that allow the user to monitor the activity and performance of the VTS subsystem.

See the ITSO Redbook, IBM Virtual Tape Server: Implementation Guide (SG24-2229) for the SMF type 94 record format and field descriptions.

The type 94 records should be collected by the VTS user for input to reports and graphs. Sample jobs for DFSort which extract the VTS fields from the records can be found on the MKTTOOLS disk in the package VTSRPTS. Sample SAS programs to produce reports from the type 94 records can also be found in the same package, and in the package VTSPWP on MKTTOOLS. MKTTOOLS is available to all IBM field personnel.

See “Guidelines for Gauging VTS Performance” below for the use of some of the fields.

RMF

RMF may be used to monitor Disconnect Time for the VTS virtual drives. Normally, the values for the VTS Logical Control Unit (LCU) are more meaningful than those for individual virtual drives.

Library Manager Panels

The Library Manager console provides graphical histograms of the Channel data rate (read and write), the number and type of virtual mounts (Hit, Miss, and Fast Ready), Physical Mount history, and also a graph of the total virtual mounts by hour. These are all represented as a rolling 24 hour graph.

Guidelines for Gauging VTS Performance

Channel Throughput

The Library Manager panel or, preferably, the SMF type 94 records may be used to monitor the channel throughput on an hourly basis. The reporting tools mentioned above can provide daily summary reports as well. The reports should be examined for periods greater than two hours (72 GB cache systems) or four hours (144 GB cache systems) where the throughput exceeds 18 GB per hour, read plus write, over the ESCON channels.

If this level of sustained demand never occurs, that is, there are no periods of multiple consecutive hours over 18 GB per hour data transfer, the VTS has room for workload growth.

If this level of activity occurs occasionally, such as on busy days only, or only once per day, then the VTS should be giving good performance overall, but there is likely little or no room for additional workload.

If this multi-hour, high level of demand is occurring multiple times per day, or every day for a longer period, then unless this is an exceptional workload which has a finite lifetime and the workload is expected to decrease, either the workload should be decreased by moving some to another tape subsystem such as native 3590 devices, or another VTS should be acquired.

Virtual Drive Mount Time

The SMF type 94 data provides the Maximum, Minimum, and Average mount times for the virtual drives in the VTS. The key fields are the Average and the Maximum. These can be used to gauge the impact of recall-inducing specific mounts (cache misses).

Examine the reports for hours which show maximum mount times exceeding 900 seconds (15 minutes) or average mount time exceeding 300 seconds (5 minutes). These indicators point to workload that is not cache friendly. Some consideration should be given to moving this workload to another type of subsystem such as native 3490E or 3590.

Apply the same guidelines as for the channel throughput above to the occurrence of these mount time events, in that an occasional appearance should not be alarming, but chronic or repeated appearances should be questioned.

VTS Disconnect Time

RMF reports or RMF monitors may be used to determine if there are periods when the disconnect time for the VTS LCU exceeds 500 ms. This is an indication that throttling is occurring.

In extreme cases, throttling will be manifested in very long or erratic job run times, and steps should be taken to immediately determine the offending workload and move it to another subsystem or acquire another VTS.

Percentage of Recalls

If the percentage of logical mounts that cause the recall of a virtual volume exceeds 20%, then the user may want to consider either moving some of this type workload to native drives, or upgrading the VTS to 144 GB of cache storage. While a percentage of recall-inducing mounts greater than 20% does not in itself imply that the VTS is stressed, experience to date has shown that users may be dissatisfied with VTS performance when this guideline is exceeded. The impact on performance appears as extended mount times and/or erratic job run times.

Improving Performance

Workload

Workload directed to a VTS should be screened for the following characteristics. Any workload will execute, but some workload takes advantage of the VTS architecture, which provides automated stacking of small volumes onto high capacity volumes. The following types of workload should be directed away from the VTS:

Multi-volume data sets. These are likely to take advantage of the capacity of native 3490E and 3590 drives, which is more cost effective use of tape. The recall of multi-volume data sets is considered to be a taxing workload for the VTS.

Heavy data recall activity where the data recalled is unlikely to be in the cache at the time of recall. The prime example of such a workload is DFSMShsm recalls of migrated data sets.

If these guidelines are followed, the VTS is more likely to provide excellent performance and a high level of satisfaction. Introduction of small amounts of workload with unfriendly characteristics is not likely to cause serious problems, but by monitoring the data in the SMF records, it is possible to see stress developing before there is a problem.

Hardware Capacity

If the workload is appropriate, or in any case, a given, and the VTS is indicating signs of stress then it may be necessary to add system capacity.

Upgrading the Tape Volume Cache (TVC) capacity from 72 GB to 144 GB can provide improvements in mount time, short term throughput, and offer protection against throttling. The SMF type 94 records provide virtual volume time in cache statistics, which can be used as an indicator of the need for more cache. If the virtual volume cache residency time is chronically less than 2 hours, then there is a strong likelihood that additional cache would be of benefit. At the present time there is no tool to accurately quantify the benefit of additional TVC capacity.

If the VTS subsystem is configured with less than six 3590 drives, and long mount times or throttling are occurring, consider installing the full complement of 3590s. Also note that if there are multiple VTS subsystems in the same SMS storage group with unequal numbers of 3590 drives, steps should be taken to balance this resource. All VTSs in the same storage group should have the same cache size and number of 3590 drives.

The Volume Mount Analyzer (VMA) which is supplied with DFSMS/MVS may be used to determine the workload characteristics of the jobs using the VTS. VMA Extract files may be used as input to the IBM internal tool VTSA. VTSA can be used to determine if additional VTS subsystems are required, and also allow the user to plan for growth.

If monitoring of the VTS indicates appropriate workload (or the workload cannot be changed) and the reports also indicate that the VTS may be at or near capacity, the VTSA should be run to determine additional requirements. VMA/VTSA also are helpful in determining appropriate workload for 3590 and 3490E native subsystems as well as the VTS.

END of DOCUMENT