Evaluating the Performance Characteristics and Resiliency of IBM DS8000 Global Mirror

IBM DS8000 Asynchronous Remote Mirroring

October 2009

Victor T. Peltz
International Business Machines Corporation
Dr. H. Pat Artis
Performance Associates Inc.

Legal Notices and Disclaimers

No part of this document may be reproduced or transmitted in any form without written permission from the IBM Corporation or Performance Associates, Inc.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information could include technical inaccuracies or typographical errors. IBM and / or Performance Associates, Inc. may make improvements and / or changes in the product(s) and / or program(s) referenced in this paper at any time without notice. Any statements regarding IBM's or Performance Associates' future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The performance data contained herein was obtained in a controlled, isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While IBM and Performance Associates have reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customer experiences described herein are based upon information and opinions which have been provided by the customer. The same results may not be obtained by every user.

Reference in this document to IBM or Performance Associates, Inc. products, programs, or services does not imply that IBM or Performance Associates, Inc. intends to make such products, programs or services available in all countries in which IBM or Performance Associates, Inc. operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead. It is the user's responsibility to evaluate and verify the operation on any non-IBM product, program or service.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR INFRINGEMENT. IBM and Performance Associates, Inc. shall have no responsibility to update this information. IBM products are warranted according to the terms and conditions of the agreements (e.g. IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The providing of the information contained herein is not intended to, and does not grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
USA

Performance Associates may be contacted at:
Performance Associates, Inc.
P.O. Box 5080
Pagosa Springs, CO 81147-5080
USA
(970) 731-3273
www.perfassoc.com

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Redbooks (logo)®	z/OS®	z/VM®
z9™	AIX®	DB2®
DFSMS/MVS™ DFSMSdss™	DFSMShsm™	
DS4000™	DS6000™	DS8000™
Enterprise Storage Server®	ECKD™	ESCON®
FlashCopy®	Geographically Dispersed ParallelSysplex™	
GDPS®	HyperSwap™	FICON®
HACMP™	IBM®	IMS™
Parallel Sysplex®	Redbooks®	System p™
System z™	System Storage™	Tivoli®
TotalStorage®	VTAM®	

PAI/O Driver® is a registered trademark of Performance Associates, Inc.

The following terms are trademarks of other companies:

SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Oracle, JD Edwards, PeopleSoft, Siebel, and TopLink are registered trademarks of Oracle Corporation and/or its affiliates.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Acknowledgements

The authors would like to thank the organizations who supplied the data used to characterize the workloads used in this study. We also would like to thank Performance Associates, Inc. and the IBM Tucson Performance team who executed the experiments discussed in this paper. Without their hard work, the study would not have been possible.

A note to the reader

This paper assumes the reader is familiar with the general concepts of remote data mirroring, the underlying concepts of synchronous and asynchronous mirroring and the terms Recovery Time Objective (RTO) and Recovery Point Objective (RPO). For a brief explanation of these terms and concepts, please refer to sections 7.1 and 7.2 in the Appendix. Additional information may be found in references [1] and [2] listed in the References section at the end of this paper.

The reader also will find it helpful to be familiar with the concept of Consistency Groups and how they are created as a means of preserving write order dependencies when performing remote mirroring. Readers unfamiliar with write order dependencies or Consistency Groups should consult reference [2] in the References section.

Synchronous and asynchronous data mirroring are two examples of a general set of storage-related functions known as Data Replication. For a discussion of where data mirroring fits in the hierarchy of data replication functions, the reader may want to consult reference [8].

Terminology

In this paper, the terms Remote Site and Secondary Site are synonymous.

Two terms are used more or less synonymously in the IT industry to denote the time lag between the Primary and Secondary storage systems, namely: “RPO” which is discussed in section 7.2.2, and “Consistency Group Formation Time (CGFT).” In this paper, we will use the term “CGFT” to denote the amount of time between when the last Consistency Group was formed and the most current updates at the remote site.

Abstract and Summary

Objectives of this Study

This paper describes the results of a joint study conducted by Performance Associates Inc. and the IBM Systems & Technology Group performance organization. The principal components used in the study were an IBM System z host, three IBM System Storage™ DS8300 storage systems, the IBM DS8000 Global Mirror remote replication function, and the Performance Associates, Inc. PAI/O Driver® for z/OS[1] which was used to simulate actual remote mirror production workloads.

The objectives of the study were to investigate two specific questions.

1. Host – Global Mirror interaction, namely:
    - does the I/O workload influence DS8000 Global Mirror performance, and
    - does Global Mirror affect DS8000 host response time?
2. How “business realities” affect Consistency Group formation time, for example:
    - planned and unplanned workload growth,
    - large temporary workload spikes, and
    - Primary-to-Secondary link outages.

Principal Results and Conclusions

The purpose of this study was to explore the robustness of IBM’s DS8000 Global Mirror function using workloads based on the characteristics of actual customer environments. The applications of interest were:

- Brokerage: the opening 15 minutes of trading on the New York Stock Exchange,
- Automotive: insurance policy management, and
- Credit Card: sale transaction authorization.

Three tests were performed:

1. An I/O test measuring host response time and Consistency Group formation time using the OLTP and batch workload profiles developed for each of the applications previously listed.
2. The same I/O test plus a 60 GB sequential workload burst injected into the mix at a rate of 1 GB/sec. This experiment tested the storage system’s ability to absorb this extra data, recover, and restart the process of forming Consistency Groups at the remote site.
3. The same I/O test with only 50% of the Primary-to-Secondary link bandwidth available. This experiment investigated the effect constrained link bandwidth would have on the formation of Consistency Groups.

[1] For additional information about the PAI/O Driver for z/OS, see: www.perfassoc.com.

The experiments in this study illustrate that DS8000 Global Mirror is a robust remote mirroring solution which is meeting its design goals.

1. **Protect host response time.** Do not allow the remote mirroring activity to elongate an application's I/O activity any more than absolutely necessary.
2. **Maintain data consistency.** Maintain data consistency at the remote site and minimize the time lag between when a data was modified at the Primary site and when the corresponding updates were secured at the Secondary site.
3. **Automatic recovery.** Execute and be able to recover from problems automatically with minimal or no operator intervention. Examples are:
    - **Large transient data spikes.** If the storage system is hit with a large transient data spike and is temporarily unable to form Consistency Groups, restart the process as soon as possible. This was the case with the second set of experiments. Because link bandwidth was artificially constrained, the storage system was temporally busy transmitting the entire set of sequential burst data, in addition to the ongoing OLTP write updates, to the Secondary and had insufficient link bandwidth to also transmit the data necessary to form Consistency Groups. Once the sequential burst data was completely transmitted to the Secondary, formation of Consistency Groups resumed automatically.
    - **Loss of a data link.** The storage system should automatically utilize link bandwidth as it becomes available. This was confirmed by the experiments where one of the two data links was disabled. The time required to form Consistency Groups increased temporarily during the interval when only one link was available. When the second link was re-enabled, the system started using it and Consistency Group formation time reverted to its original value.

Lastly, and equally important, this study demonstrates that it is possible to construct accurate test workloads which model an installation's live production I/O activity. This, in turn, makes it possible to install, test, and validate a remote mirroring solution without exposing the installation's actual production workload to risks inherent in implementing new functions and procedures without having thoroughly tested them first.

Based on our conclusions in this study, we believe installations should feel comfortable that the DS8000 Global Mirror replication function is a robust and effective solution that can be deployed with confidence for business-critical remote mirroring requirements.

Table of Contents

Figures

Tables

1 Introduction

The requirements imposed by modern 24x7 business environments necessitate that many IT systems must deliver very high, or nearly continuous, availability. For many installations, a key component of a system that can maintain high availability is the capability of replicating data to a remote site in as near to real time as possible.

This paper describes the results obtained from a joint study performed by Dr. H. Pat Artis, of Performance Associates Inc. and the Enterprise Storage Performance organization of the IBM Systems & Technology Group. The purpose of this study was to evaluate the IBM DS8000 Global Mirror remote replication function to assess its resiliency when performing asynchronous replication for a typical large 24x7 on-line transaction-driven application.

The objectives of the study were to investigate two specific questions.

1. Host – Global Mirror interaction, namely:
    - does the I/O workload influence DS8000 Global Mirror performance, and
    - does Global Mirror affect DS8000 host response time?
2. How "business realities" affect Consistency Group formation time, for example:
    - planned and unplanned workload growth,
    - large temporary workload spikes, and
    - Primary-to-Secondary link outages.

Three IBM DS8300 storage systems were used to run DS8000 Global Mirror. An IBM System z 2097, model 755 processor running z/OS® was used to host the I/O workload; the application workload was simulated using the PAI/O Driver® for z/OS developed by Performance Associates, Inc. The PAI/O Driver for z/OS is discussed in section 2.4. In section 3 we will describe: (1) the methodology used to characterize the workloads and, (2) the actual workloads used in the study.

The IBM DS8000 Global Mirror function is one example of hardware-based asynchronous storage mirroring. A brief explanation of synchronous and asynchronous mirroring is given in section 7.1.1. See also: reference [8] for a discussion of how hardware-based storage mirroring fits into the overall framework of data replication.

2 Principal components of this study

2.1 IBM Global Mirror Configuration Architecture

The version[2] of Global Mirror used in this study is implemented on the IBM DS8000, DS6000 and ESS800 storage systems. Refer to Figure 1. Note the storage at the Primary site is in a Master / Subordinate relationship. Each storage system at the Primary site sends data to its corresponding Secondary storage system at the Remote site. The Master controls the creation of Consistency Groups. When it is time to create a

[2] Other IBM storage products, for example the IBM DS5000 and the IBM SAN Volume Controller (SVC), offer Global Mirror asynchronous remote mirroring. However their implementations differ from the Global Mirror function used in this study.

Consistency Group, the Master signals the Subordinates to initiate the process [ref. 2] of creating a Consistency Group.

From a connectivity standpoint, there is no restriction regarding using any combination of DS8000, DS6000, and ESS800 at the Local site and connecting them to any combination of DS8000, DS6000, and ESS800 at the Remote site. In the example shown, we have a DS8000 Master sending data to a DS8000 Secondary at the Remote site, a DS6000 Subordinate sending data to an ESS800 at the Remote site, and an ESS800 sending data to a DS6000 at the Remote site.

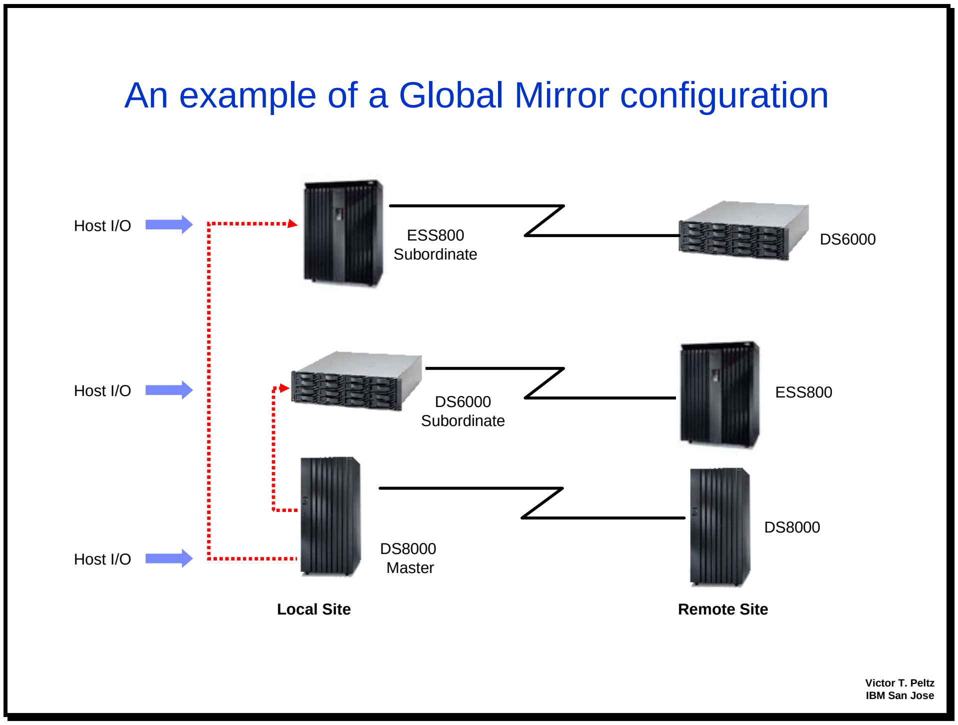


Figure 1: A sample Global Mirror configuration

2.2 Hardware configuration for this study

The hardware configuration used in this study is shown in Figure 2. The salient features of this configuration are as follows:

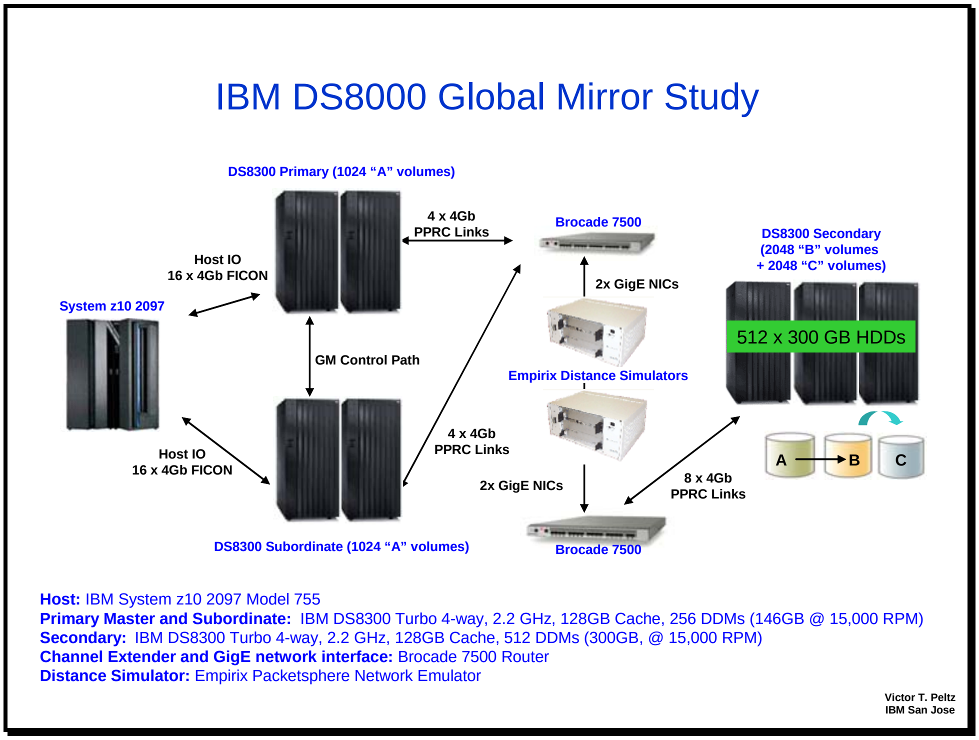
1. The Primary storage consists of a DS8300 Master and a DS8300 Subordinate. We wanted to test a configuration that included a Subordinate to verify that communication between the Master and Subordinate, or other as yet unknown factors, would not hinder the configuration's ability to form Consistency Groups.

2. The DS8300 Secondary at the Remote site was deliberately configured with the same number of physical disks as the Primary DS8300 storage systems. Each physical disk on the Secondary DS8300 is double the raw capacity of the disks on the Primary DS8300s. This was done to match the practice of many actual installations, whereby they try to minimize the number of physical disks used on the Secondary storage systems for budget reasons.

   In this configuration, a logical volume "A" on the Primary is paired with a logical volume "B" on the Secondary. Logical volume "B" in turn acts as the Source volume for the FlashCopy operation [ref. 2] at the Secondary. The "C" volumes (i.e., the Targets of the FlashCopy operation) also reside on the same physical 300 GB disks as the "B" volumes.

3. Two Gigabit Ethernet lines provided the remote links connecting the Primary and Secondary storage systems. There are four components in the remote link data paths: two Brocade 7500 protocol converters to convert from / to FCP and Gigabit Ethernet protocols, and two Empirix Distance Simulators, one for each link, which injected delay into the network corresponding to the Primary and Secondary being a given distance apart.

4. The host system was an IBM System z 2097, model 755 processor. The IBM System z 2097 hosted the PAI/O Driver for z/OS [ref. 3]; the PAI/O Driver for z/OS was used to reproduce the customer I/O workloads.



IBM DS8000 Global Mirror Study

Host: IBM System z10 2097 Model 755
Primary Master and Subordinate: IBM DS8300 Turbo 4-way, 2.2 GHz, 128GB Cache, 256 DDMs (146GB @ 15,000 RPM)
Secondary: IBM DS8300 Turbo 4-way, 2.2 GHz, 128GB Cache, 512 DDMs (300GB, @ 15,000 RPM)
Channel Extender and GigE network interface: Brocade 7500 Router
Distance Simulator: Empirix Packetsphere Network Emulator

Victor T. Peltz
IBM San Jose

Figure 2: Hardware Configuration for the DS8000 Global Mirror Study

2.3 DS8000 Secondary FlashCopy

As part of the process of creating Consistency Groups at the Secondary, Global Mirror utilizes the internal FlashCopy™ function of the DS8000 storage system. DS8000 FlashCopy can operate in one of two ways[3]:

1. "Traditional" FlashCopy which will allocate sufficient space to hold an entire logical volume, and

2. "Space-efficient" FlashCopy which will only allocate a subset of the space for an entire logical volume which is sufficient to hold the Global Mirror write updates.

[3] See reference [2] for additional information about DS8000 FlashCopy operation.

One of the decisions which had to be made prior to starting the data collection process was whether to perform each workload simulation using Traditional FlashCopy, "Space-efficient" FlashCopy, or both at the Secondary storage system. A previous study of DS8000 Global Mirror performance, shown in section 7.6, indicates that the performance and behavior of Global Mirror is almost identical whether one uses Traditional FlashCopy of "Space-efficient" FlashCopy at the Secondary. Hence, the decision was made to use only Traditional FlashCopy for this study.

2.4 PAI/O Driver for z/OS

The Performance Associates Inc. PAI/O Driver for z/OS is a vendor and technology independent set of tools and services for evaluating the performance characteristics of storage subsystems. Licensed enterprise users employ the host software test component of the product and the software is site licensed to storage subsystem vendors worldwide. The PAI/O Driver for z/OS software is part of the overall process Performance Associates uses to develop an I/O profile for a particular storage system. This process is discussed in section [3] .

3 Host workloads

3.1 Host workload characterization methodology

Much of the text in this section is taken from a paper written by Artis and available on Performance Associates web page: www.perfassoc.com/Published-Papers.html. See also: reference [4].

When designing a workload to drive an I/O subsystem, one must decide what values to assign to several important parameters. For example:

- The I/O request transfer size: i.e.,
  - the number of bytes per block to be transferred during each read or write, and
  - whether the size of the blocks is constant or varies according to some predefined distribution.
- The distribution of the I/O request rate: will it be constant or vary according to a predefined distribution.
- The percentage split between random vs. sequential reads and writes.

Using the PAI/O Driver for z/OS workload characterization facilities, profiles were developed for actual workloads from Fortune 500 companies. The I/O data from these installations was captured using standard z/OS functions, namely:

- System Measurement Facility (SMF),
- Report Measurement Facility (RMF), and
- the DCOLLECT facility.

Based on these data sources, representative I/O workload profiles were developed for each of the environments along with the total replication bandwidth requirements. Having collected and analyzed the relevant data, it is possible to program the PAI/O Driver for z/OS to simulate the desired workload at any desired total I/O rate.

3.1.1 z/OS primary data sources for workload characterization

There are three primary data sources that support the remote copy workload characterization process:

- SMF type 42 subtype 6 dataset activity (**TYPE42DS**) records,
- RMF type 74 device activity (**TYPE74**), type 74 subtype 5 cache statistics (**TYPE74CA**), and type 74 subtype 7 switch statistics (**TYPE74SW**) records, and
- DCOLLECT volume (**DCOLVOLS**) and dataset (**DCOLDSET**) records.

These data sources are used to characterize volume, dataset, and subsystem level activity and will be discussed in the following sections.

3.1.1.1 Volume and Dataset level records for workload characterization

Figure 12 provides an overview of how the TYPE74, TYPE74CA, TYPE42DS, DCOLDSET, and DCOLVOLS are employed to analyze volume level activity. The figure summarizes the activity for a hypothetical volume, DBX232, for one RMF interval.

To analyze at the workload component level, there are five critical data sources:

1. **TYPE74 (RMF 74-1)**. Each z/OS image in the Sysplex creates a TYPE74 (RMF 74-1) record for each volume. These observations need to be summed for each RMF interval to determine the number of I/O operations per volume. The 74-1 record does not give counts of read/write and random/sequential SSCH operations.
2. **TYPE74CA (RMF 74-5)**. The TYPE74CA (RMF 74-5) record provides the total number of read/write and random/sequential SSCH operations as well as the cache-hit counts for the volume.
3. **TYPE42DS (SMF 42-6)**. For each RMF interval, there will be a number of TYPE42DS (SMF 42-6) records, one for each open dataset. These records provide read/write and random/sequential SSCH and block counts. The ratio of block and SSCH counts provides the average chain length. The number of blocks times the blocksize yields data transfer.

   Unfortunately, there are some workloads that do not record SMF 42-6 records. Examples of workloads that do not produce 42-6 records include VTOC and catalog I/Os, applications that issue their own CCWs, and software virtual tape products.
4. **DCOLDSET**. The DCOLLECT dataset records (DCOLDSET) provide the dataset size and the blocksize for each dataset.
5. **DCOLVOLS**. The DCOLLECT volume records (DCOLVOLS) provide the size of each volume.

Section [7.4] has additional information concerning use of these records to characterize dataset level I/O activity.

3.1.1.2 System-level records for workload characterization

To address the potential of missing SMF 42-6 records, one can reconcile the data bytes transferred based on the SMF and RMF device records with the total read and write data transfer for the subsystem. This can be done by using RMF TYPE74SW

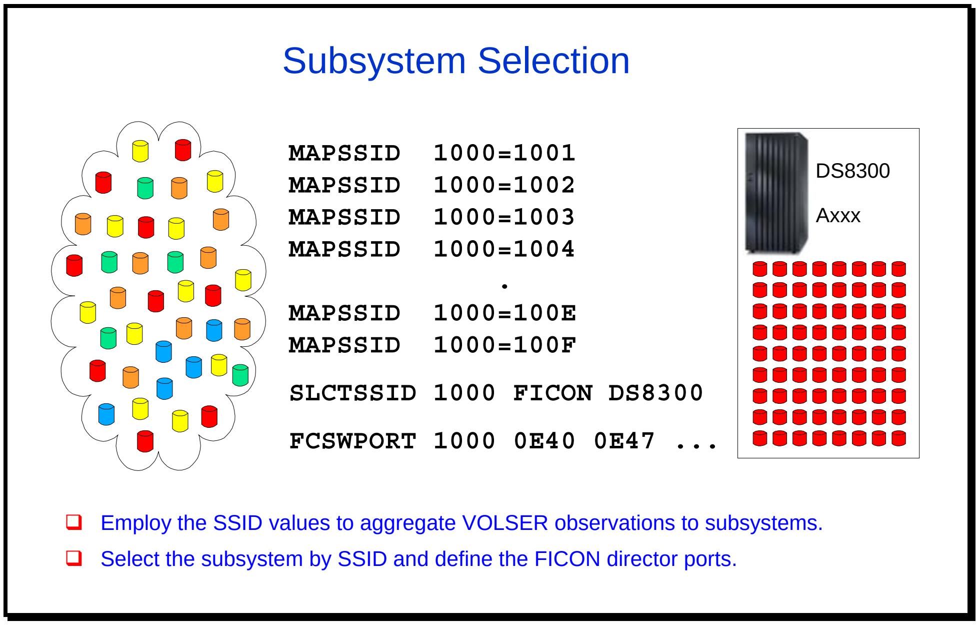
(RMF 74-7) records. These records provide the read and write byte counts for the subsystem. For subsystems connected via FICON switches, the RMF TYPE74SW observations provide precise measures of the actual read/write data transfer activity.

The first step in characterizing subsystem activity and reconciling the TYPE42DS dataset level data transfer values with the FICON switch statistics is the association of the dataset/volume level information for the subsystem with the FICON switch statistics. Figure 3 provides an overview of the subsystem selection and definition process. Specifically, the selection and definition process must provide three mappings:

1. Map the multiple SSIDs that represent a modern storage subsystem into a single SSID.
2. Select one or more of the aggregated SSIDs for analysis. That is, extract all of the applicable TYPE74, TYPE74CA, TYPE42DS, DCOLDSET, and DCOLVOL information for the subsystem.
3. Map the FICON switch port data transfer statistics to the selected aggregate SSID(s).

Once these three mappings have been completed, the volume and dataset level data transfer statistics may be reconciled with the aggregate read/write data transfer FICON switch statistics.

For further detailed information about the workload characterization methodology, see section 7.4.2 and reference [4].



Subsystem Selection

```
MAPSSID  1000=1001
MAPSSID  1000=1002
MAPSSID  1000=1003
MAPSSID  1000=1004
                .
MAPSSID  1000=100E
MAPSSID  1000=100F

SLCTSSID 1000 FICON DS8300

FCSWPORT 1000 0E40 0E47 ...
```

❑ Employ the SSID values to aggregate VOLSER observations to subsystems.

❑ Select the subsystem by SSID and define the FICON director ports.

Figure 3: Association of Dataset / Volume Statistics with FICON Switch Statistics

3.2 Weighted Average Host Response Time

For all of the measurements discussed when the term Host Response Time is used, it refers to the weighted average host response time for the two Primary storage systems. The weighted average host response time (RT) is calculated as follows:

$$RT = \frac{(RT_{Master} \text{ x I/O Rate}_{Master}) + (RT_{Subordinate} \text{ x I/O Rate}_{Subordinate})}{\text{I/O Rate}_{Master} + \text{I/O Rate}_{Subordinate})}$$

Figure 4: Calculation of Weighted Average Host Response Time

3.3 Host workloads used in this study

By using the workload characterization methodology previously described, it was possible to program the PAI/O Driver for z/OS to simulate five workloads:

- **Brokerage "Market Open."** This workload is based on the opening 15 minutes of trading activity of the New York Stock Exchange.
- **Brokerage Batch.** This workload simulates the overnight batch operations associated with the daily trading activity. This is typically when large batch database updates are performed.
- **Credit Card.** A major worldwide credit card organization was used to derive the profile for I/O activity for online credit card authorizations.
- **Automotive Insurance Online.** This workload was derived from a national automotive insurance company. It represents the peak on-line activity resulting from insurance claims processing.
- **Automotive Insurance Batch.** This workload captures the I/O activity for the nightly batch updates and backups resulting from the online activity during the day.

In addition to the Brokerage "Market Open" workload discussed in the body of this paper, four other workloads were examined. These workloads included Brokerage Batch, Credit Card, Auto Insurance Online, and Auto Insurance Batch. The behavior of DS8000 Global Mirror with these workloads is very similar to the behavior observed with "Market Open." While these results are included in section 7.5, they will not be discussed in detail in the body of the paper. Inspection of the graphs will confirm that similar comments apply to these workloads as for the corresponding "Market Open" experiments.

The characteristics of each of these workloads are summarized in Table 1.

z/OS Workload	Time Interval	Data Selection Criteria	Primary to Secondary Distance	Avg. I/O transfer size (bytes)	Random			Sequential	
					% of I/Os	% Read Hit	% Write	% of I/Os	% Write
Brokerage Market Open	9:30 9:45	NYSE Open	1,800 mi.	35K	57%	96%	22%	43%	52%
Brokerage Batch	0:45 01:45	Peak Seq. Write	1,800 mi.	47K	40%	97%	12%	60%	29%
Credit Card	00:00 23:59	Full day	700 mi.	50K	61%	94%	15%	39%	37%
Auto Ins. Online	10:30 11:45	Max. Tx. Rate	1,400 mi.	27K	72%	95%	5%	28%	29%
Auto Ins. Batch	21:00 21:45	Peak Seq. Write	1,400 mi.	21K	62%	91%	9%	38%	51%

Table 1: z/OS workloads used in this study

	"Market Open"	"Market Open" + 60 GB Burst	"Market Open" with 50% link bandwidth	Overnight Batch	Overnight Batch + 60 GB Burst	Overnight Batch With 50% link bandwidth
Brokerage	Page 19	Page 20	Page 22	Page 25	Page 37	Page 38
Credit Card	Page 39	Page 40	Page 41	n/a	n/a	n/a
Auto Insurance	Page 42	Page 43	Page 44	Page 45	Page 46	Page 47

Table 2: Workload - Experiment cross-reference

3.4 Experimental Tests

Three tests were performed:

1. An I/O test measuring host response time and Consistency Group formation time using the OLTP and batch workload profiles developed for each of the applications previously listed using aggregate I/O rates from 50% to 150% of the observed RMF interval peaks for the customer workload environments.
2. The same I/O test plus a 60 GB sequential workload burst injected into the mix at a rate of 1 GB/sec. This experiment tested the storage system's ability to absorb this extra data, recover, and restart the process of forming Consistency Groups at the remote site.

3. The same I/O test with only 50% of the Primary-to-Secondary link bandwidth available. This experiment investigated the effect constrained link bandwidth would have on the formation of Consistency Groups.

4 Global Mirror Study Results

4.1 Brokerage: "Market Open"

This workload represents the opening 15 minutes of trading on the New York Stock Exchange. Refer to Table 1. The Primary DS8000 storage systems that support the I/O activity related to recording stock trades are mirrored to Secondary DS8000 systems located 1,800 miles away. Note the size of the average transfer is 35k bytes, which is significantly larger than a block size of 4k bytes that is used in many typical OLTP workload studies. The larger transfer size[4] is what was observed in the I/O activity of the production systems handling the stock trading activity. Note this workload consists of both random I/O (57%) and sequential I/O (43%). The total write portion of the workload is the sum of the random and sequential write components. Using the data in Table 1, we calculate the total write portion for "Market Open" as:

(22% of 57%) + (52% of 43%) = 35% writes

Referring to Figure 5, we can make the following observations about DS8000 Global Mirror performance with this workload:

- **Average Host Response Time.** Average Host Response Time remains essentially flat[5] despite the aggregate I/O rate almost doubling over the measurement interval. This indicates that Global Mirror is meeting one of its principal design goals: to minimize impact of mirroring on Host Response Time and hence, minimize the impact of the remote mirroring I/O on the business-critical application itself.
- **Consistency Group Formation Time (CGFT).** CGFT increased slightly as the workload (i.e., I/O rate) increased but did not exceed about 2.5 seconds. As with Average Host Response Time, this behavior for CGFT shows that Global Mirror is achieving one of its design goals for this configuration: the data currency at the Secondary storage system at the Remote site is kept within approximately 2.5 seconds of the production data at the Primary site. This result also suggests that sufficient bandwidth was available in the data link between the Primary and Secondary such that the link itself did not become a bottleneck. Estimating data link bandwidth requirements is an important aspect of designing a remote mirroring configuration.

[4] Transfer size = blocksize x CCW chain length.

[5] Studies over the years have shown that from a human factors standpoint, it is more important to provide consistent response time rather than fast but erratic response. As a metric: the standard deviation of response time as measured at the end-user's terminal is more critical than the mean (i.e., average) response time; assuming the mean is less than some value which is deemed acceptable by the end-user.

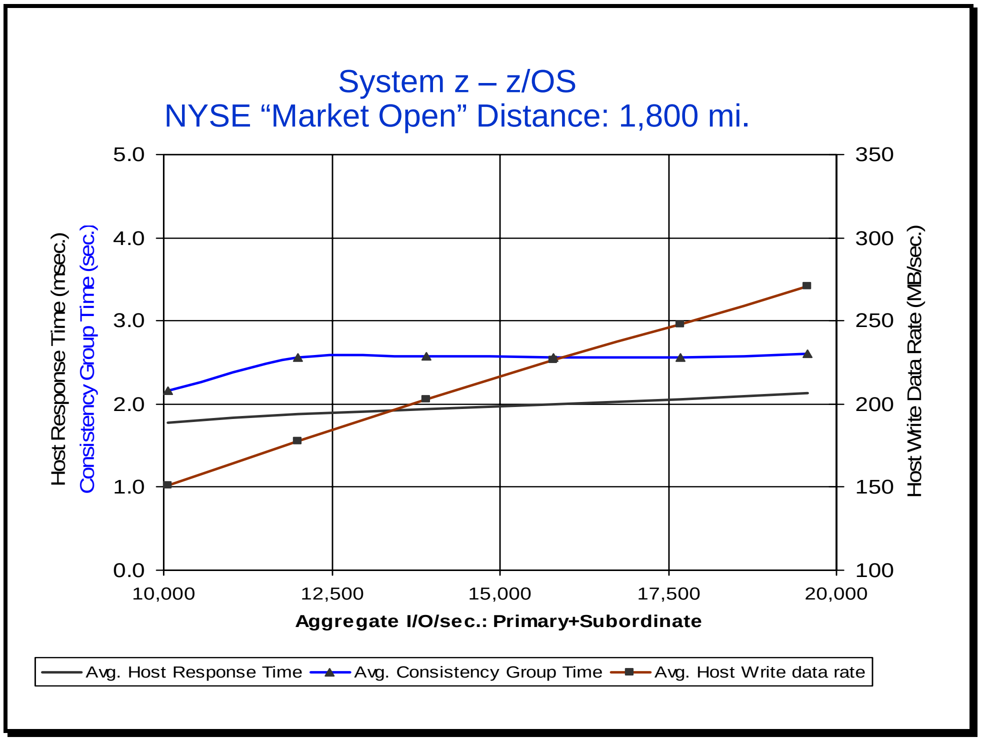


Figure 5: NYSE "Market Open"

- **Average Host Write Data Rate.** The rate at which data was transferred from the System z host to the DS8000 Primary storage systems increased linearly as the I/O rate increased. The DS8000 Primary storage systems, Master and Subordinate, were able to absorb the workload as it increased while maintaining an almost constant host response time.

These three results illustrate that DS8000 Global Mirror was able to achieve the business objectives set by the customer:

1. Protect host response time; do not allow the remote mirroring activity to elongate the application's I/O activity any more than absolutely necessary.
2. Maintain data consistency at the remote site and minimize the time lag between when a data was modified at the Primary site and when the corresponding updates were secured at the Secondary site. In the event of an unplanned outage or other incident at the Primary site which results in the installation having to fail over to the Secondary site, keeping data consistent at the Secondary and as up-to-date as possible helps minimize the amount recovery effort.

4.2 Brokerage: "Market Open" + 60 GB Sequential Burst

Often real life I/O workloads are not strictly transaction-based random workloads or sequential batch workloads. As shown in Table 1, the "Market Open" workload is already a mixture of OLTP random and sequential batch I/O. In this study, we also wanted to explore what would happen if in addition to the "Market Open" workload, Global Mirror suddenly had to handle a radically different workload, in this case a very large sequential burst of write data. This could happen, for example, if a system were handling multiple different applications with different I/O workload profiles.

The following two charts illustrate just such a test. For clarity the experimental results are shown as Figure 6 and Figure 7; Figure 6 shows CGFT and Host I/O rate (I/O/sec.), Figure 7 shows CGFT and Host Write Data Rate (MB/sec.). The red line on both figures shows CGFT for the experiment; the timeline of the experiment is as follows:

1. The "Market Open" workload is started and allowed to run for approximately 9 minutes.
2. At the 9-minute mark, shown by the vertical lines on Figure 6 and Figure 7, a second workload is added: namely 60 GB of data written at a rate of 1 GB/sec. to the Primary DS8000s.
3. Injection of this large additional sequential workload causes the remote links to be saturated due to insufficient bandwidth to handle this increased amount of data. Global Mirror suspends attempting to form Consistency Groups. This is shown by the break in the red line on both figures.
4. At the 18-minute mark, enough of the write updates have been transferred across the Primary-to-Secondary link to enable Consistency Groups to begin forming again.

Note that throughout this sequence, host I/O rate, shown on Figure 6, remained relatively constant. The response and consistency group formation time perturbations resulting from the 60 GB sequential write workload by the Primary DS8000s are shown clearly in Figure 7. Up to the 9-minute mark, the DS8000s are handling the write content at roughly 250 MB/sec. At the 9-minute mark the 60 GB sequential write workload is introduced, causing a pronounced spike in the Host Write Data Rate – briefly to over 1,250 MB/sec. The Host I/O Rate and Write Data Rate measurements show that while Consistency Group formation had to suspend temporarily, host I/O was largely unaffected. This experiment confirms two of the DS8000 Global Mirror principal design objectives.

1. Minimize the impact to host application I/O. Host I/O continued largely unaffected despite the injection of the 60 GB sequential write workload and the resulting temporary suspension of forming Consistency Groups.
2. Minimize the necessity to tune Global Mirror or require operator intervention. When sufficient bandwidth became available Consistency Group formation started automatically – as shown at the 18-minute mark.

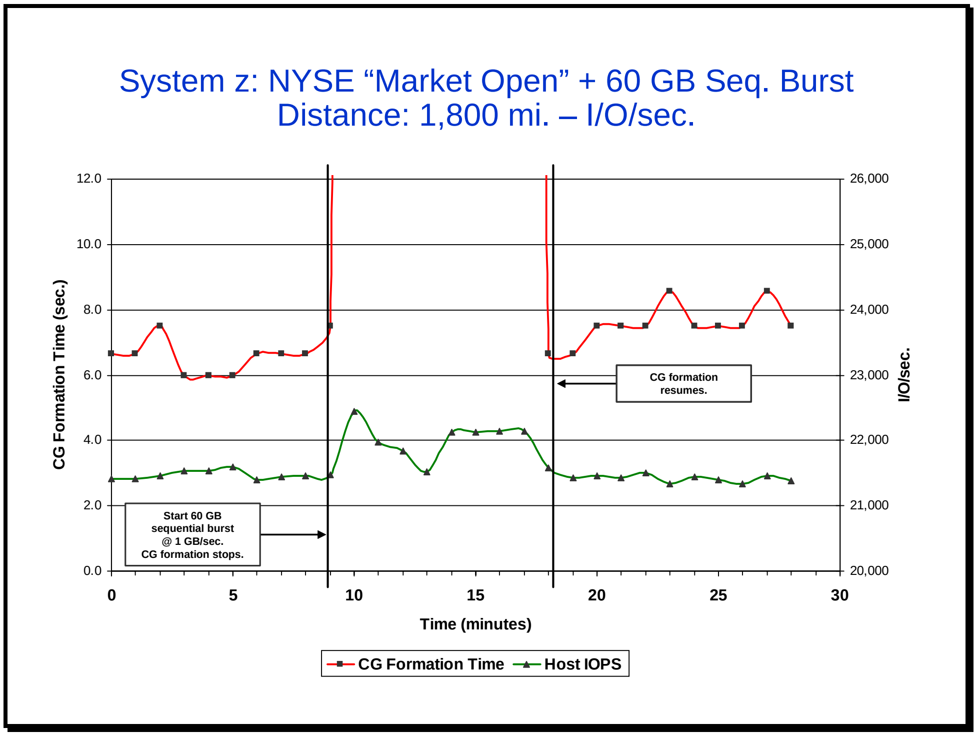


Figure 6: "Market Open" + 60 GB Sequential Burst - I/O/sec.

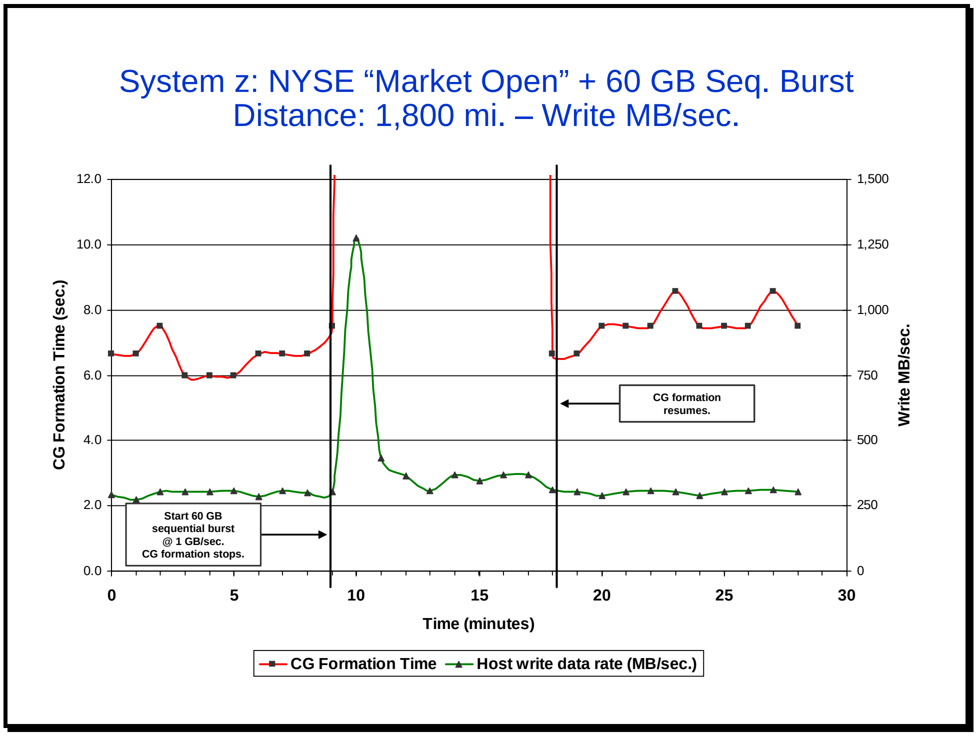


Figure 7: "Market Open" + 60 GB Sequential Burst - Write MB/sec.

4.3 Brokerage: “Market Open” with 50% link bandwidth

The experiment described in section 4.2 shows what can happen when a heavy write workload is injected into the I/O stream. Another “event,” which, hopefully, will not happen often, is the loss of a portion of the Primary-to-Secondary link bandwidth. Since most customers employ two or more disparate replication paths, which may be supplied by multiple telecom vendors, a loss of 50% of the total link bandwidth is probably a worst-case scenario. Hence, a 50% reduction in bandwidth was selected for these tests.

For this experiment, one of the two data links was disabled temporarily. The results are shown in Figure 8 and Figure 9 . As with the previous experiment, for clarity, we are showing the results in two figures: Figure 8 shows CGFT and Host I/O Rate, Figure 9 shows CGFT and Host Write Data Rate. The red line on both figures shows CGFT for the experiment; the timeline of the experiment is as follows:

1. The “Market Open” workload was started and allowed to run for approximately 9 minutes.
2. At the 9-minute mark, shown by the vertical lines on Figure 8 and Figure 9, one of the two data links was disabled.
3. The experiment was allowed to run with only one data link until approximately the 14-minute mark.
4. At the 14-minute mark the second data link was enabled again.

During the interval with only one data link, Host I/O Rate and Host Write Data Rate were reduced slightly and Consistency Group Formation Time increased. This is to be expected given that link bandwidth was reduced by 50%.

This experiment verifies another of DS8000 Global Mirror’s design goals: automatically utilize available bandwidth. When additional bandwidth became available (i.e., when the second data link was enabled again) Global Mirror began using it, CGFT returned to approximately 6 seconds, the value prior to the loss of one link, and Host I/O reverted to performing as it did prior to the loss of a link.

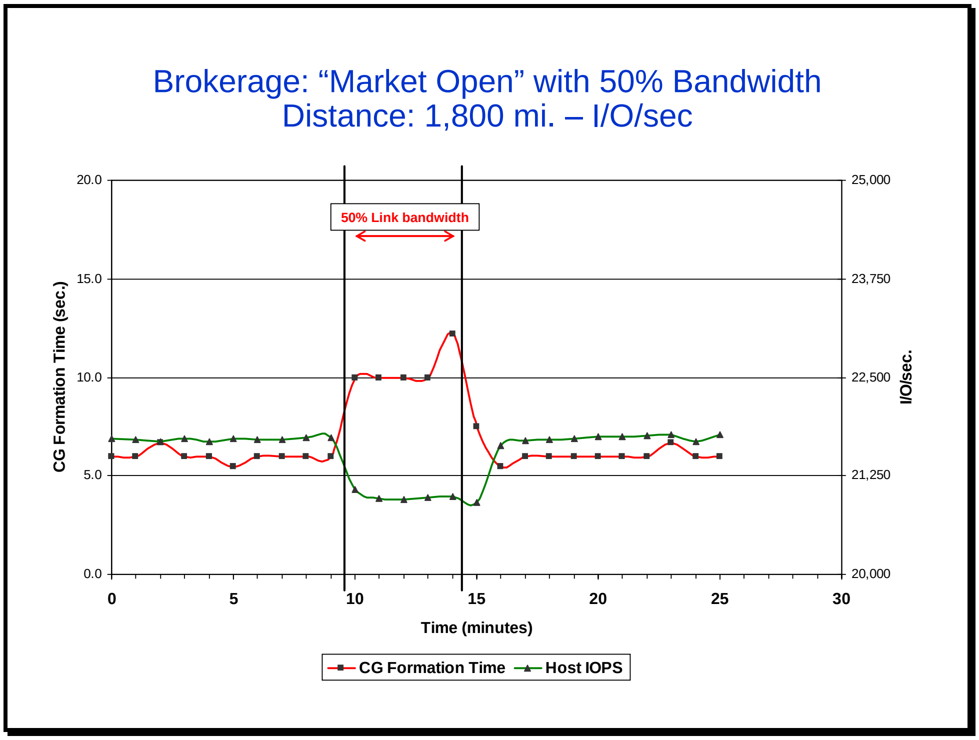


Figure 8: "Market Open" with 50% link bandwidth - I/O/sec.

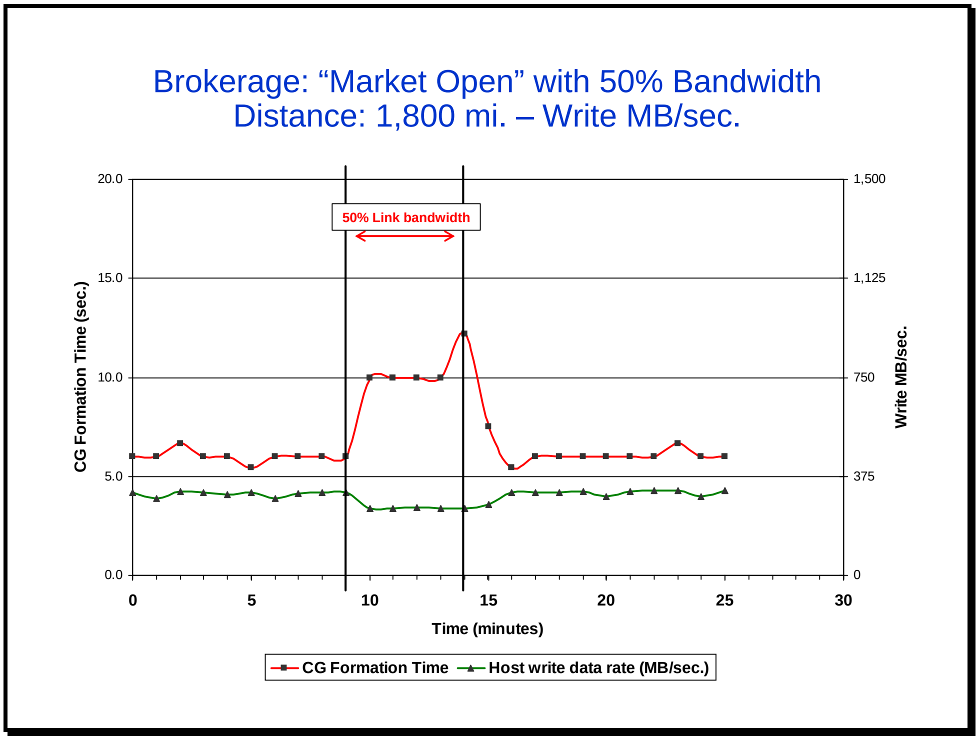


Figure 9: "Market Open" with 50% link bandwidth - Write MB/sec.

4.4 Brokerage: Overnight Batch

In section 4.1, a critical period of the NYSE prime shift OLTP workload was discussed. Data also was collected for the overnight batch processing; the results are shown in Figure 10. Similar observations to those made for the OLTP workload can be made for the behavior of Global Mirror when handling the batch workload.

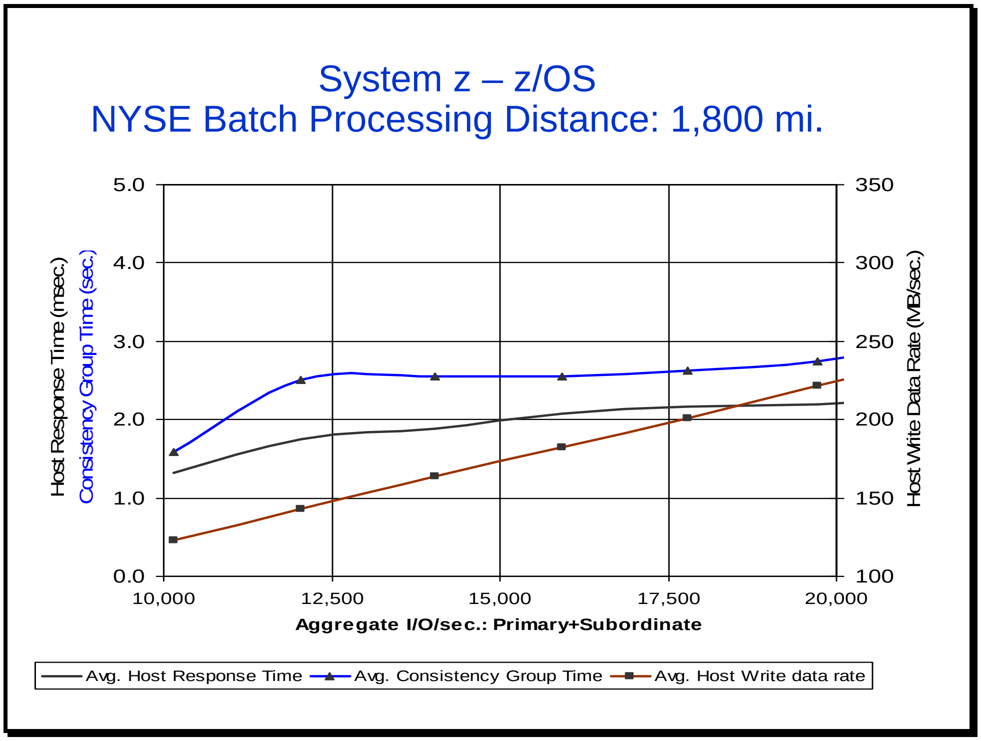


Figure 10: NYSE Overnight Batch Processing

- **Average Host Response Time.** Average Host Response Time did not remain quite as flat as with the OLTP "Market Open" workload, however this is not as critical with batch processing. So long as the batch processing workload completes within the allotted overnight window, some variation in the applications' Host Response Time typically is not a concern.

- **Consistency Group Formation Time (CGFT).** As with the prime shift OLTP workload, CGFT increased slightly as the batch workload I/O rate increased and then more-or-less flattens out until reaching the higher end of the workload's I/O rate at about 19,000 IOPS. At around 19,000 IOPS, we note that CGFT is starting to increase slightly. This is likely because the limitation imposed by the capacity of the Primary-to-Secondary link bandwidth is starting to become a bottleneck. This is typical of CGFT behavior for batch processing. The average transfer size of I/O requests for batch jobs is usually significantly larger than corresponding average transfer size for transaction-based I/O requests.

  Because batch processing typically requires more bandwidth than OLTP workloads, one may be tempted to size the Primary-to-Secondary link bandwidth to meet the batch processing requirement, which may exceed the OLTP bandwidth requirement. However, this need not always be the case, one can sometimes cheat. So long as the CGFT does not grow beyond what the business deems to be an acceptable amount of data which would have to be recovered in the event of an unplanned outage, then one may be able to employ links with sufficient bandwidth to ensure good and consistent CGFT for the OLTP workload while letting the batch CGFT increase somewhat. An example of this trade-off is shown in section 7.7.

5 DS8000 Global Mirror performance at other distances

For this study, all of the workloads assume specific distances between the Primary and Secondary storage systems; for example: in the case of Brokerage all of the experiments previously described assume the Primary-to-Secondary distance is 1,800 miles. Specific distances were used because these are the distances between the respective Primary and Secondary storage systems at the actual customer installations from which the I/O data was collected.

However, it is reasonable to ask "what is the behavior of Global Mirror at other distances?" While this study did not investigate Global Mirror behavior at distances other than those documented in this paper, a previous Global Mirror study[6] did investigate Global Mirror performance at various distances using a workload similar to the OLTP workloads used in this study.

In this previous study, a workload, known as "70/30/50," was used. This workload is characterized by having 70% reads, 30% writes, and 50% read-hits. By comparison, the "Market Open" workload has 35% writes. The "70/30/50" workload is one of several "Open Systems" workloads used by IBM to analyze the performance of storage configurations when operated in an AIX environment.

The conclusions of this previous study show that DS8000 Global Mirror can perform well at distances greater than those used in this study. An example of how Global Mirror performs at longer distances is shown in Figure 11.

While "70/30/50" is not a z/OS-based workload, we believe it is reasonable to assume that DS8000 Global Mirror also can perform well at distances greater than 2,000 miles in System z environments.

[6] See reference [7]. This is the same study that investigated DS8000 Global Mirror performance using Traditional vs. "Space-efficient" FlashCopy (see: section 2.3).

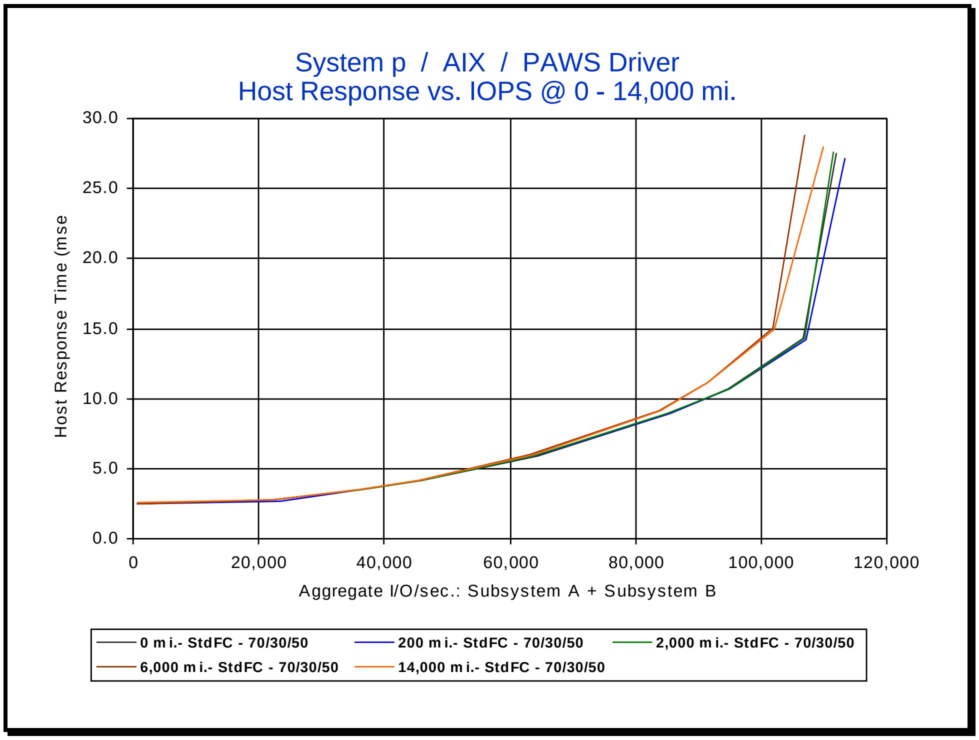


Figure 11: Global Mirror performance at distances greater than 2,000 miles

6 Study Conclusions and Recommendations

The purpose of this study was to explore the robustness of IBM's DS8000 Global Mirror function using workloads that accurately model several real customer environments. The applications of interest were:

- Brokerage: the opening 15 minutes of trading on the New York Stock Exchange,
- Automotive: insurance policy management, and
- Credit Card: sale transaction authorization.

Three tests were performed:

1. An I/O test measuring host response time and Consistency Group formation time using the OLTP and batch workload profiles developed for each of the applications previously listed.
2. The same I/O test plus a 60 GB sequential workload burst injected into the mix at a rate of 1 GB/sec. This experiment tested the storage system's ability to absorb this extra data, recover, and restart the process of forming Consistency Groups at the remote site.
3. The same I/O test with only 50% of the Primary-to-Secondary link bandwidth available. This experiment investigated the effect constrained link bandwidth would have on the formation of Consistency Groups.

In the case of the Brokerage and Automotive applications, measurements were made for the OLTP and overnight batch I/O activity. For the Credit Card application, only OLTP I/O activity was measured.

All of the experiments confirmed that DS8000 Global Mirror is a robust remote mirroring solution that is meeting its design goals.

1. **Protect host response time.** Do not allow the remote mirroring activity to elongate an application's I/O activity any more than absolutely necessary.
2. **Maintain data consistency.** Maintain data consistency at the remote site and minimize the time lag between when a data was modified at the Primary site and when the corresponding updates were secured at the Secondary site.
3. **Automatic recovery.** Execute and be able to recover from problems automatically with minimal to no operator intervention. Examples are:
   - **Large transient data spikes.** If the storage system is hit with a large transient data spike and is temporarily unable to form Consistency Groups, restart the process as soon as possible. This was the case with the second set of experiments. Because link bandwidth was artificially constrained, the storage system was temporarily busy transmitting the entire set of sequential burst data, in addition to the ongoing OLTP write updates, to the Secondary and had insufficient link bandwidth to also transmit the data necessary to form Consistency Groups. Once the sequential burst data was completely transmitted to the Secondary storage system, formation of Consistency Groups resumed automatically.

- **Loss of a data link.** The storage system should automatically utilize link bandwidth as it becomes available. This was confirmed by the experiments where one of the two data links was disabled. The time required to form Consistency Groups increased temporarily during the interval when only one link was available. When the second link was re-enabled, the system started using it and Consistency Group formation time reverted to its original value.

Lastly, and equally important, this study demonstrates that it is possible to construct accurate test workloads to model an installation's live production I/O activity. This, in turn, makes it possible to install, test, and validate a remote mirroring solution without exposing the installation's actual production workload to risks inherent in implementing new functions and procedures without having thoroughly tested them first.

Based on our conclusions in this study, we believe installations should feel comfortable that the DS8000 Global Mirror replication function is a robust and effective solution that can be deployed with confidence for business-critical remote mirroring requirements.

7 Appendix

7.1 Remote Mirroring

7.1.1 Types of Remote Mirroring

There are two categories of remote mirroring functions found in the marketplace today.

1. **Synchronous mirroring.** Mirroring which is typically used for storage systems separated by campus distances or up to a few hundred kilometers. This type of remote replication is known as synchronous mirroring[7]. Synchronous mirroring is designed to facilitate zero data loss at the storage hardware level in the event of a Primary storage system failure.

2. **Asynchronous mirroring.** Mirroring used when the storage systems are separated by distances greater than a few hundred kilometers up to distances of thousands of kilometers or more. This type of remote replication is known as asynchronous mirroring. With asynchronous mirroring some data may be lost in the event of a failure at the Primary site, but the data at the Secondary is maintained in a consistent state to facilitate application recovery and restart at the Secondary site.

IBM's hardware synchronous remote replication function for DS8000 is named Metro Mirror; the IBM asynchronous replication function is named Global Mirror. In addition to storage hardware mirroring, host-based software mirroring is available in the marketplace; however, it is not discussed in this paper.

7.1.2 Factors determining the choice of Remote Mirroring

Several factors influence the choice of whether an installation runs synchronous or asynchronous replication, the most important being:

- the distance between the Local (Primary) site which is the source of the data, and the Remote (Secondary) site which is the recipient of the data,
- the host I/O workload characteristics of the data being replicated, in particular the writes per second,
- the desired performance of the storage system and, in particular, the remote replication function, and
- the ongoing operational costs for the remote replication function. Ongoing remote replication operational costs are often cited as a major component of the Total Cost of Ownership (TCO) of a corporate business continuity implementation.

In many cases, the data link bandwidth required to achieve the required level of performance for the given workload introduces a significant portion of the overall operational cost. Hence, it is important to estimate bandwidth requirements accurately. Tools are available to assist in calculating link bandwidth requirements; consult Performance Associates, Inc. and / or IBM for additional information.

[7] IBM DS8000 synchronous mirroring can be used for distances up to 300 km. In some instances it may be acceptable to use DS8000 synchronous mirroring at distances greater than 300 km, however, IBM should be consulted first.

7.2 Principal Business Continuance Metrics

Two metrics are commonly used to establish IT system requirements for a Business Recovery plan: Recovery Time Objective (RTO) and Recovery Point Objective (RPO).

7.2.1 Recovery Time Objective (RTO)

Recovery Time Objective (RTO) stipulates how long the business can tolerate an IT application being inoperable. RTO values can range from a few minutes to multiple hours. Surveys across multiple industry sectors have established that RTOs typically range from a few minutes to multiple hours, depending on the industry and application.

7.2.2 Recovery Point Objective (RPO)

Recovery Point Objective (RPO) is used in two slightly different ways that are related to each other.

1. RPO stipulates the amount of data, expressed in units of time, which an installation considers acceptable to lose at the Primary site in the event of an unplanned outage. The data will be recovered at the Secondary site as part of the application recovery and restart process at the Secondary site.
2. RPO measures the currency of the data at the Secondary compared to the Primary, typically in some number of seconds. For example, if a host application is processing 1,000 transactions per second and the data at the Secondary is 5 seconds older than the data at the primary, roughly 5,000 transactions worth of data will have to be recovered at the Secondary in the event of an unplanned outage.

The time lag measurements discussed in this paper use the second interpretation of RPO previously described. However, the term Consistency Group Formation Time (CGFT) will be used to denote the time lag between the Primary and the Secondary. CGFT will be defined in section 7.3. The term “RPO” is used to denote the installation’s business objective for how far behind the Primary the Secondary should be allowed to lag.

7.2.3 RPO for Synchronous vs. Asynchronous mirroring

7.2.3.1 Synchronous Mirroring operation

When a Primary storage system receives a write command for a z/OS volume that is part of a remote copy pair, the Primary initiates the process of transmitting the data to the Secondary storage system at the remote site. This process, including actual data transmission, takes time. If the remote mirror process is operating synchronously, the host is informed the I/O operation is complete only after the Primary system receives notification from the Secondary that it has received the data. Hence, for synchronous mirroring, the host application will experience an elongated I/O time compared to the I/O time if the operation had only written to locally attached storage.

Because the host is informed that the write operation is complete only after the Primary receives notification from the Secondary, in this case the RPO for a given write operation is zero (i.e., RPO=0).

7.2.3.2 Asynchronous Mirroring operation

In contrast to Synchronous Mirroring, if the remote mirror process is operating asynchronously, the Primary system immediately informs the host that the I/O operation is complete and subsequently manages the data flow operation to the Secondary.

Hence, for asynchronous mirroring, at any given point in time the currency of the data at the Secondary storage system may lag the Primary. In this case we have RPO>0. If an installation desires to minimize the data loss at the Secondary, it will specify that the RPO be as small as possible. This time lag is typically a few seconds to some tens of seconds, depending on the workload at the Primary and Secondary storage systems and the remote data link bandwidth. In cases where an installation does not require that data loss be minimized, an RPO of several minutes or even longer may be appropriate.

7.2.4 Application recovery time as a function of RPO

Since RPO measures the amount of time the Secondary storage system lags behind the Primary, the amount of data that must be recreated at the Secondary in the event of an unplanned system outage is directly proportional to the RPO. Hence, RPO is an important measure of how fast an installation can recover from a system outage or disaster and hence, RPO has a direct influence on each application's RTO and on the RTO of the business as a whole.

7.3 DS8000 Global Mirror Consistency Group Formation

Regardless of what remote replication technique an installation utilizes, the set of data used by applications at the Secondary site must be consistent to a given point in time. Without this point-in-time consistency, it is very difficult, and in some cases impossible, to recover any data lost in flight due to an unplanned system outage.

IBM's DS8000 Global Mirror replication function is designed to keep data consistent at the Secondary by treating the collection of data involved in the remote replication process as a group and periodically confirming that all updates associated with a particular point in time has been transmitted to the Secondary storage system. This set of data is called a Consistency Group.

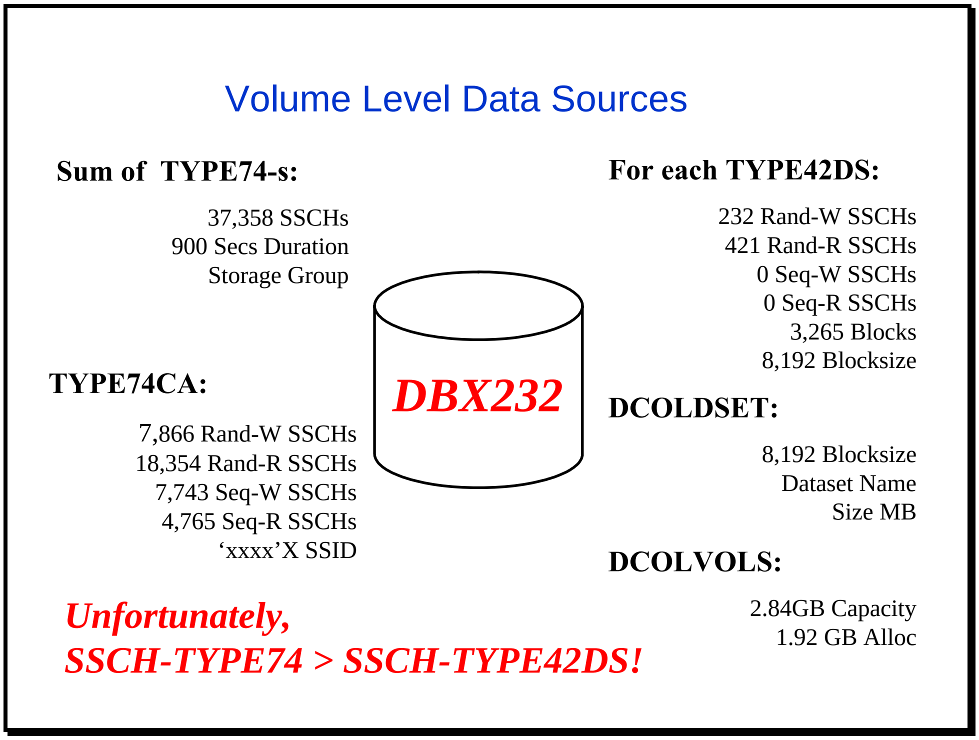
The objective for how frequent Consistency Group formation should occur is governed by a DS8000 Global Mirror parameter specified by the installation. The default is to form Consistency Groups as often as possible. With sufficient bandwidth DS8000 Global Mirror is designed to form Consistency Groups about every 3–5 seconds. Additional information about write order consistency and formation of Consistency Groups is given in reference [2].

7.4 Characterization of application workloads

Material in this section is taken from reference [4].

7.4.1 Records yielding volume and dataset-level information

This section discusses the data collection and reduction considerations for each data source of the five data sources shown in Figure 12. Special attention should be given to sources that commonly suffer from problems associated with missing or invalid data.



Volume Level Data Sources

Sum of TYPE74-s:

37,358 SSCHs
900 Secs Duration
Storage Group

TYPE74CA:

7,866 Rand-W SSCHs
18,354 Rand-R SSCHs
7,743 Seq-W SSCHs
4,765 Seq-R SSCHs
‘xxxx’X SSID

DBX232

For each TYPE42DS:

232 Rand-W SSCHs
421 Rand-R SSCHs
0 Seq-W SSCHs
0 Seq-R SSCHs
3,265 Blocks
8,192 Blocksize

DCOLDSET:

8,192 Blocksize
Dataset Name
Size MB

DCOLVOLS:

2.84GB Capacity
1.92 GB Alloc

Unfortunately,
SSCH-TYPE74 > SSCH-TYPE42DS!

Figure 12: z/OS Volume-level I/O activity data sources

The five data sources are:

1. **∑TYPE74-s**: a TYPE74 device activity record is generated for each RMF interval by every z/OS image that shares a volume. The I/O supervisor counts the number of Start SubChannel (SSCH) instructions issued to each device. The service time components (PEND, DISC, and CONN) are reported by the channel subsystem to z/OS for each SSCH operation. Hence, the TYPE74 SSCH counts and service time components are highly reliable metrics[8]. For a shared volume, the RMF interval TYPE74 observations from each z/OS image must be summarized to produce a complete picture of the volume's activity. One should note that the SSCH count is simply a total value; there is no detail about the read/write - random/sequential components of the metric. As can be seen in Figure 12, DBX232 processed 37,458 SSCH commands over the 900 second RMF interval and is a member of SMS storage group PRODDB2.

2. **TYPE74CA**: the TYPE74CA cache statistics record provides counts of the read/write - random/sequential SSCH requests processed for each volume. If enabled in a z/OS image's *ERBRMFnn* Parmlib member, the metrics in the TYPE74CA are solicited from the subsystem using the LISTDATA CCW. To avoid needless duplication of data and an increase in the SMF/RMF volume of data collected, cache reporting should be enabled for just a single z/OS image within a Sysplex.

   For volume DBX232, there were 7,866 random writes, 18,354 random reads,

[8] See: reference [5].

6,743 sequential writes, and 4,765 sequential reads. While there may be small differences between the total counts in the TYPE74 and the TYPE74CA, ratio and proportion may be employed to determine the read/write - random/sequential components of the TYPE74 reported SSCH count. In addition, the TYPE74CA record includes a 4-byte hex SubSystem ID (SSID) value that allows volume level observations to be aggregated into subsystem statistics. As can be seen in Figure 12, DBX232 is a member of SSID '1000'x.

3. **TYPE42DS**: a TYPE42DS ASID dataset activity record is created for each dataset in use by every z/OS ASID for each SMF recording interval or when an ASID terminates. The TYPE42DS records are a rich source of dataset level statistics[9]. They include read/write - random/sequential SSCH counts, cache statistics, maximum observations for service and response time, and a count of the number of blocks transferred. Using the SSCH and block counts, the average chain length for data transfers can be calculated. However, there are a number of serious issues associated with the processing of TYPE42DS records. They include:

   - Missing blocksize values. A small percentage of TYPE42DS records have missing or zero blocksize values. These values can be resolved using the blocksize values contained in DCOLDSET records,

   - Missing observations. TYPE42DS records are not created for a variety of types of datasets. These dataset types include page/swap, catalog, HSM ML1, and observations for a host of ISV products. For example, neither VTAPE nor COPYCROSS products produce TYPE42DS records for the virtual tape datasets that they redirect to DASD,

   - Incomplete support. Some database products do not update the control blocks continuously - which are the source of the TYPE42DS statistics. Rather, they simply update the control blocks when the dataset is closed. Hence, all but the last interval observation for the dataset will have zero counts and the final observation will report the total counts for the entire time the dataset was active.

   - z/OS support issues. At the current time, there is a known problem with missing observations for PDSE datasets.

   Figure 12 also includes one TYPE42DS observation for dataset PROD.INDEX with a blocksize of 8K. The observation indicates that 232 random read and 421 random write SSCH operations were issued to the dataset and a total of 3,265 blocks were transferred.

4. **DCOLDSET**: the Access Method Services DCOLLECT utility provides a snapshot of the characteristics of the datasets on each of the volumes. Since it is a point-in-time image, there may be datasets reported by the TYPE42DS records that are not reflected by the DCOLDSET observations. However, the DCOLDSET records provide a valuable source of data for resolving blocksizes that may be missing in the TYPE42DS records as well as providing the size of the

[9] See: reference [6].

dataset. Figure 12 includes one DCOLDSET observation for the dataset PROD.INDEX that was previously discussed. The dataset is 193 MB in size.

5. **DCOLVOLS**: the Access Method Services DCOLLECT utility also provides a snapshot of the characteristics of each volume when it is executed. The DCOLVOLS observations provide the capacity of the volume as well as measures of allocated space at the point-in-time the utility was executed. As can be seen in Figure 12, the capacity of volume DBX232 was 2.84 GB.

Subject to the processing issues associated with the vagaries of TYPE42DS observations, these five data sources can be processed to produce a detailed workload characterization of the read/write - random/sequential activity at the dataset, volume, and/or subsystem level. While the missing observations can be assigned default values based on a user's knowledge of the workload characteristics, subsystem level metrics must be used to validate the user's assumptions. Subsystem level statistics will be discussed in the following section.

7.4.2 System-level I/O workload statistics

As was previously discussed, the vagaries of the TYEP42DS observations present the greatest problems during the dataset / volume level workload characterization process. While a complete discussion of this issue is beyond the scope of this paper, the analysis may be approached with two selection criteria. They are:

1. the percentage of SSCH operations reported by RMF for a volume that is represented by valid TYPE42DS records, and
2. the percentage of the RMF-reported SSCH operations that are represented by valid TYPE42DS records for the subsystem as a whole.

For the purpose of this discussion, these two values are defined as %VOL and %SSID, respectively. Based on these two values, there are four algorithms for the assignment of the SSCH I/O activity as reported by RMF that are not represented by TYPE42DS observations.

1. In the ideal case, the percentage of the TYPE74 SSCH operations represented by the TYPE42DS SSCH counts for each RMF interval for the volume exceeds the %VOL value. In that case, the predominant random and sequential blocksize and chain length values for the volume may be cloned to assign the random and sequential SSCH operations that are not represented by TYPE42DS observations.
2. Should the ideal case not be true, then the user-assigned override random and sequential blocksize and chain length values may be used to assign the unknown SSCH operations. In the event the user does not wish to assign override values, then the analysis continues at the subsystem level.
3. At the subsystem level, the percentage of the TYPE74 SSCH operations represented by the TYPE42DS SSCH counts for each RMF interval for the subsystem as a whole may exceed the %SSID value. In that case, the predominant random and sequential blocksize and chain length values for the subsystem as a whole are employed to assign the random and sequential SSCHs that are not represented by TYPE42DS observations.

4. Finally, for unknown SSCH operations not addressed by any of the three prior algorithms, the user-assigned default random and sequential blocksize and chain length values may be used to assign the unknown SSCH operations.

While the application of these algorithms is not a trivial process, it is essential to characterize dataset and volume level activity to avoid problems during the implementation of remote copy. For a further detailed discussion of the analysis methodology, consult reference [4].

7.5 Other host workloads used in this study

7.5.1 Brokerage: Overnight Batch Processing

7.5.1.1 Brokerage Batch + 60 GB Sequential Burst

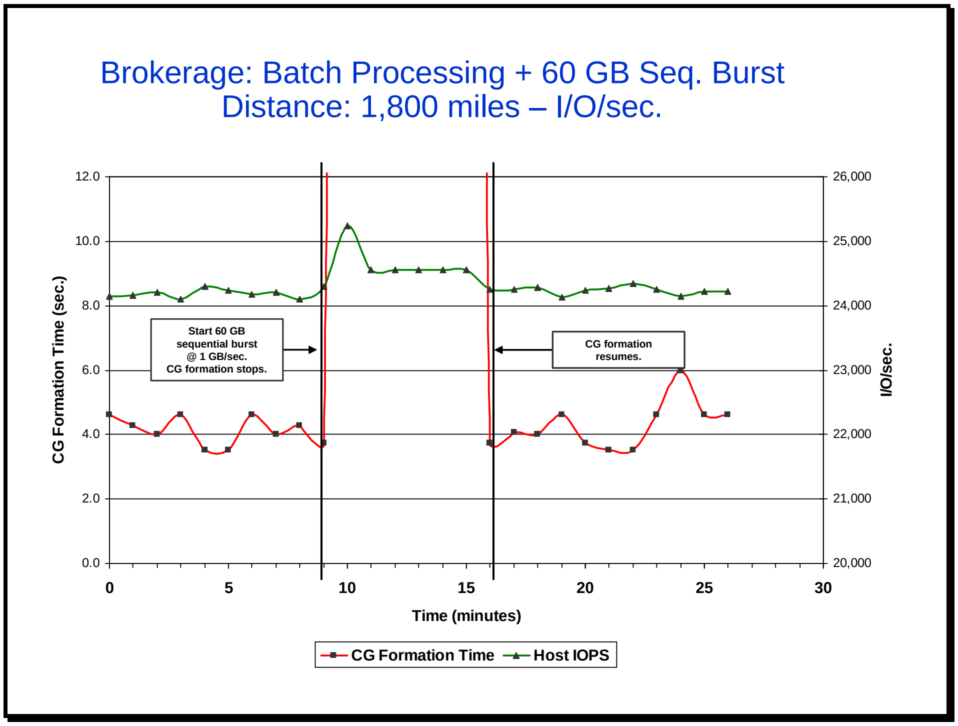


Figure 13: Brokerage Overnight Batch + 60 GB Sequential Burst – I/O/sec.

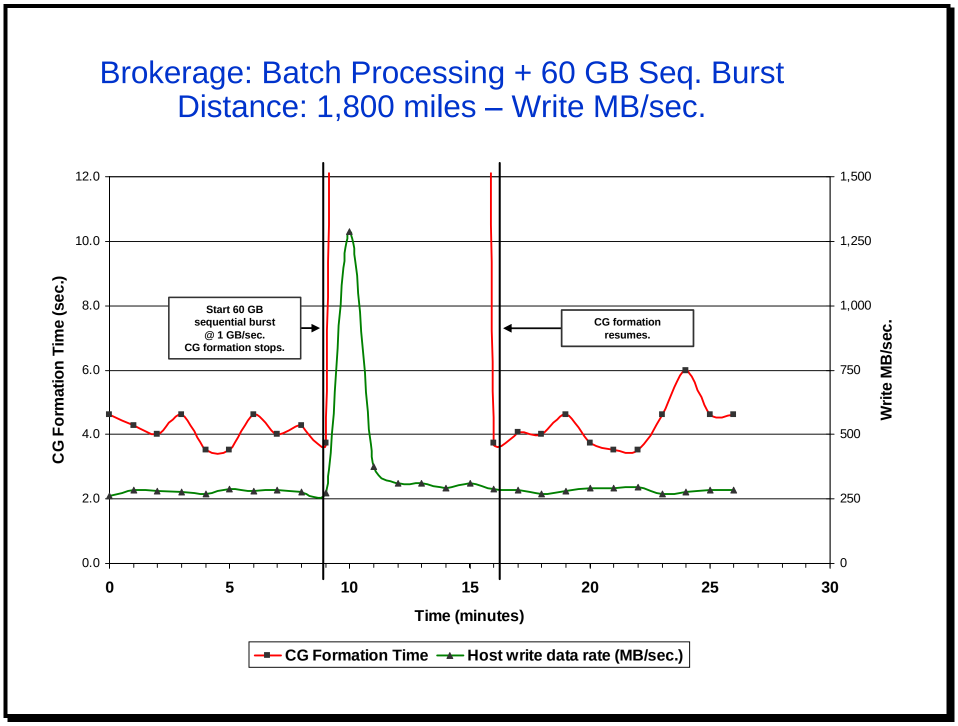


Figure 14: Brokerage Overnight Batch + 60 GB Sequential Burst - Write MB/sec.

7.5.1.2 Brokerage Batch: 50% link bandwidth

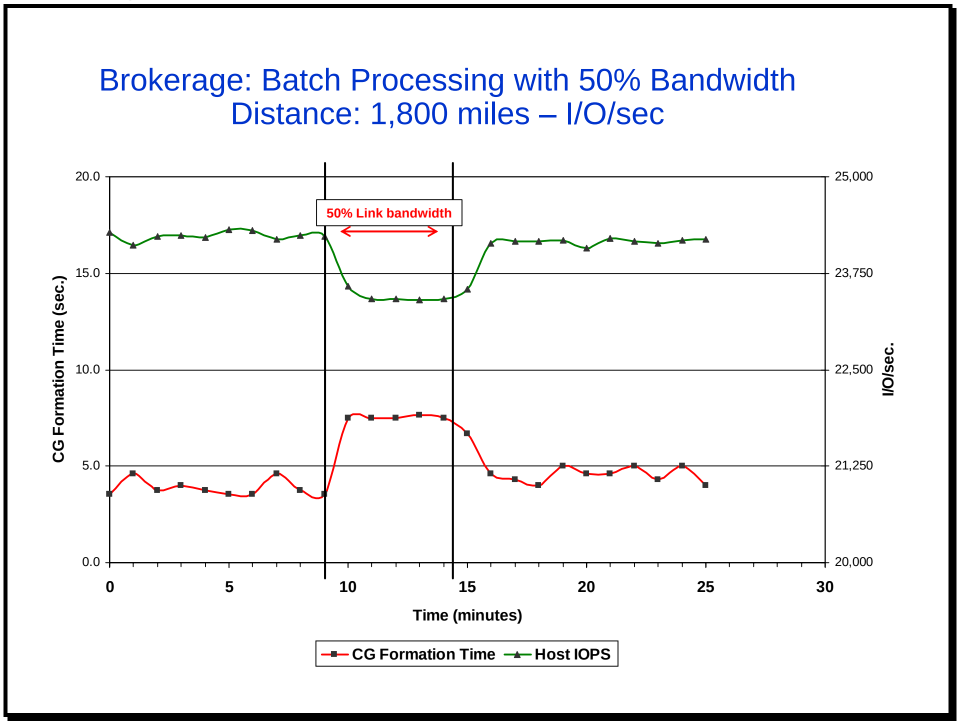


Figure 15: Brokerage Batch - 50% link bandwidth - I/O/sec.

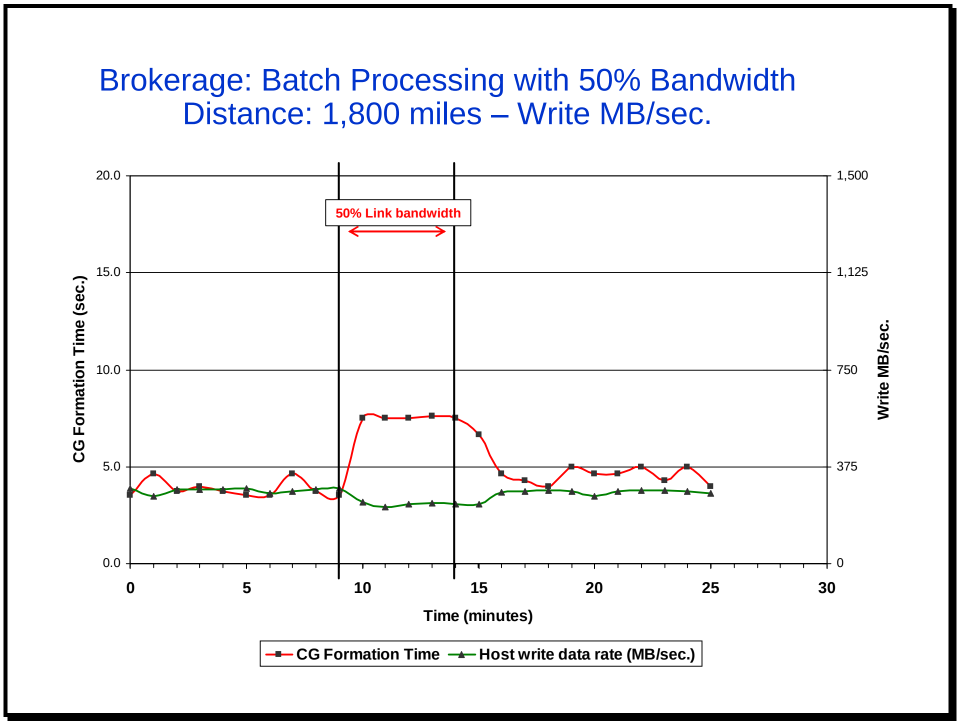


Figure 16: Brokerage Batch - 50% link bandwidth - Write MB/sec.

7.5.2 Credit Card Authorization

7.5.2.1 Credit card On-line processing

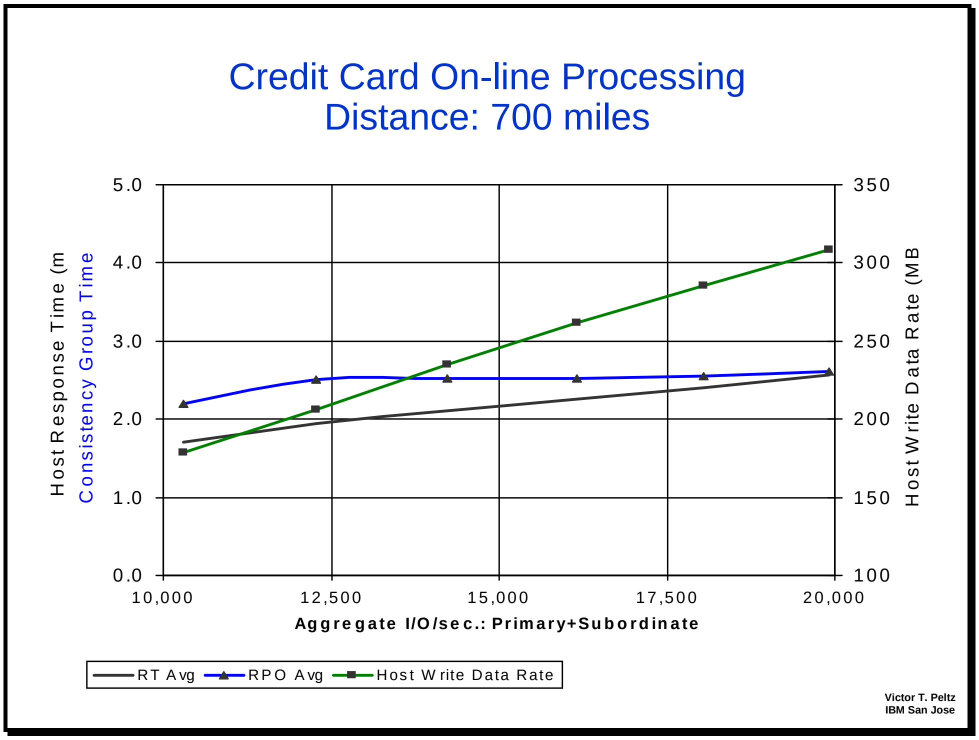


Figure 17: Credit Card On-line Processing

7.5.2.2 Credit Card: On-line processing + 60 GB Sequential Burst

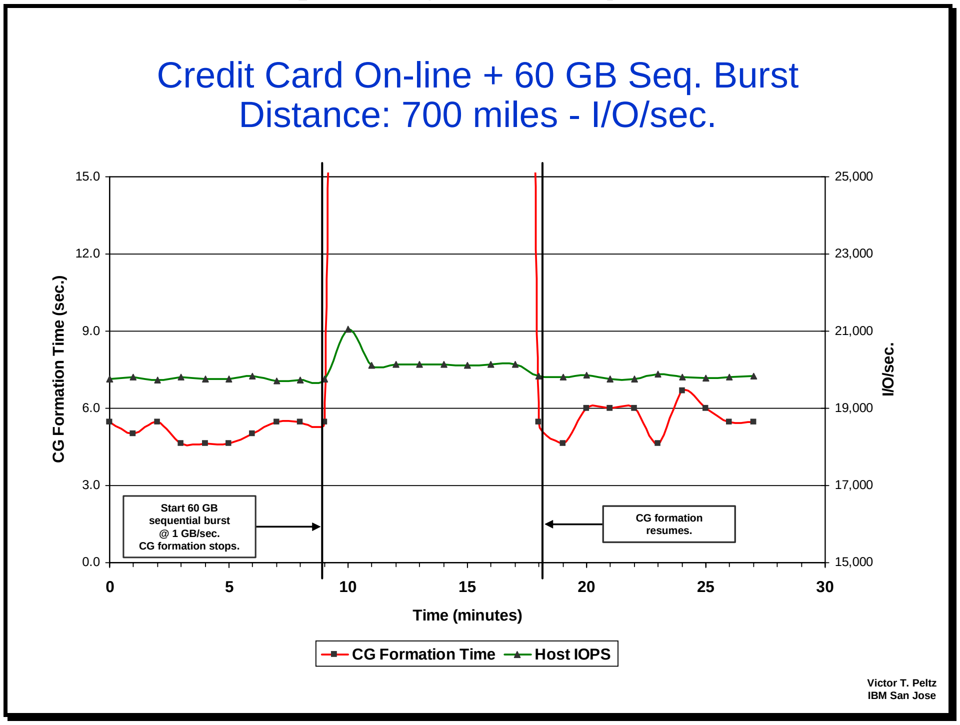


Figure 18: Credit Card On-Line processing + 60 GB Seq. Burst - I/O/sec.

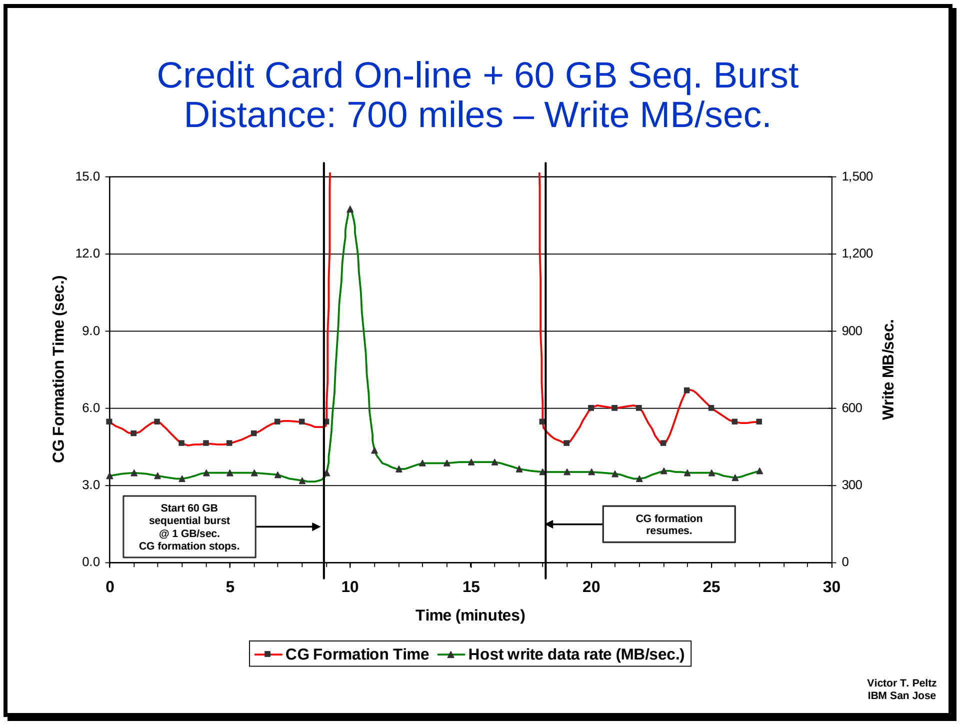


Figure 19: Credit Card On-Line processing + 60 GB Seq. Burst - MB/sec.

7.5.2.3 Credit Card On-Line processing: 50% link bandwidth

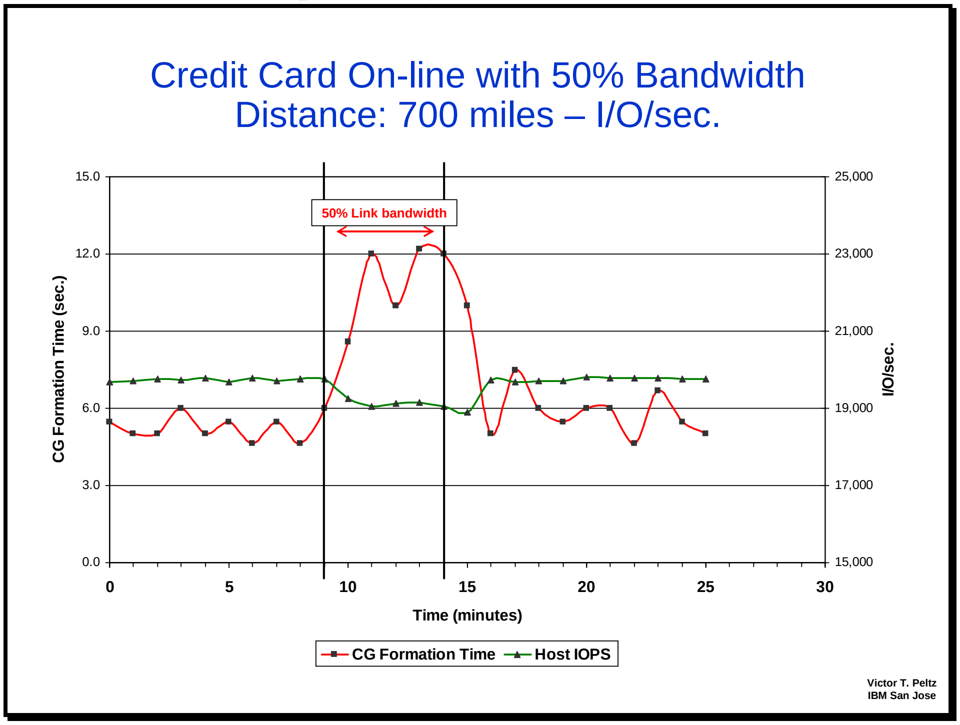


Figure 20: Credit Card On-Line processing with 50% link bandwidth - I/O/sec.

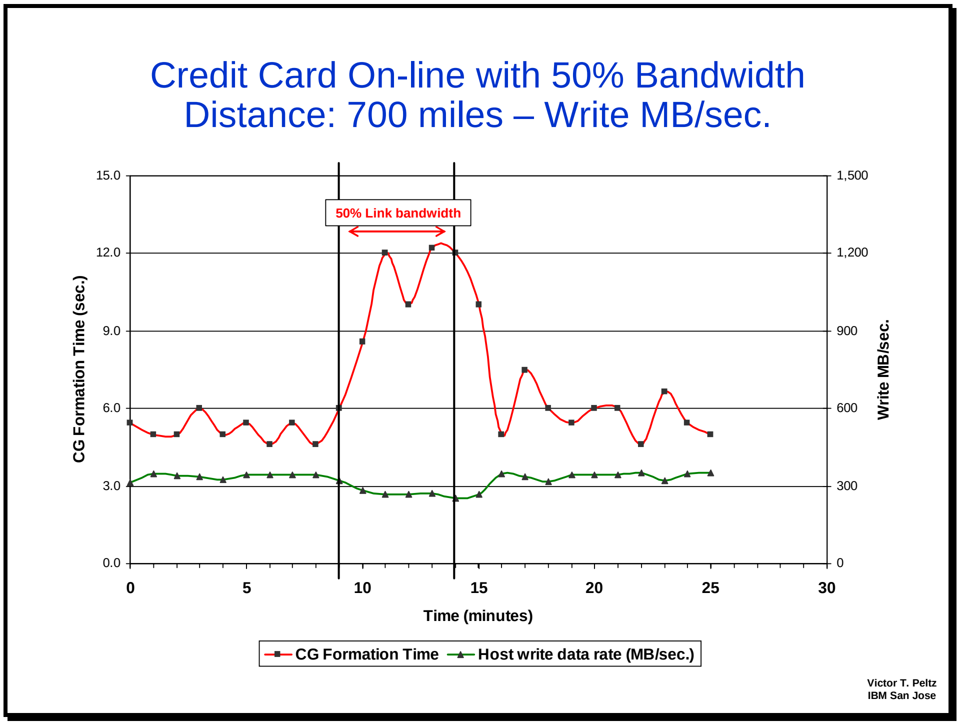


Figure 21: Credit Card On-Line processing with 50% link bandwidth - MB/sec.

7.5.3 Automotive Insurance: On-line

7.5.3.1 Automotive Insurance On-line Processing

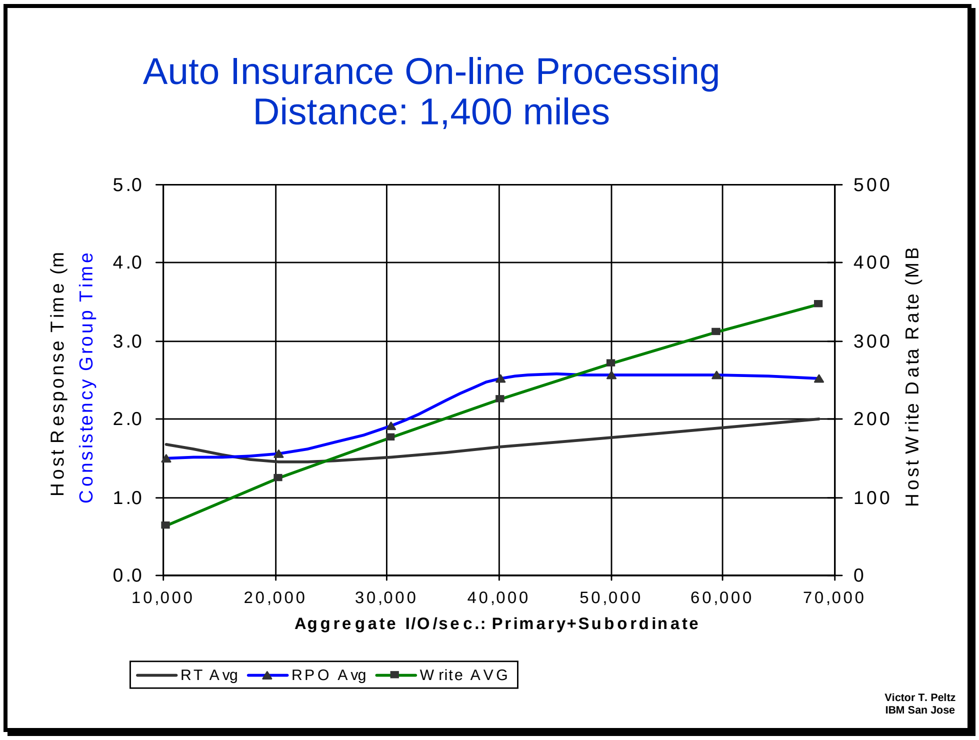


Figure 22: Auto Insurance On-Line Processing

7.5.3.2 Automotive Insurance On-line + 60 GB Sequential Burst

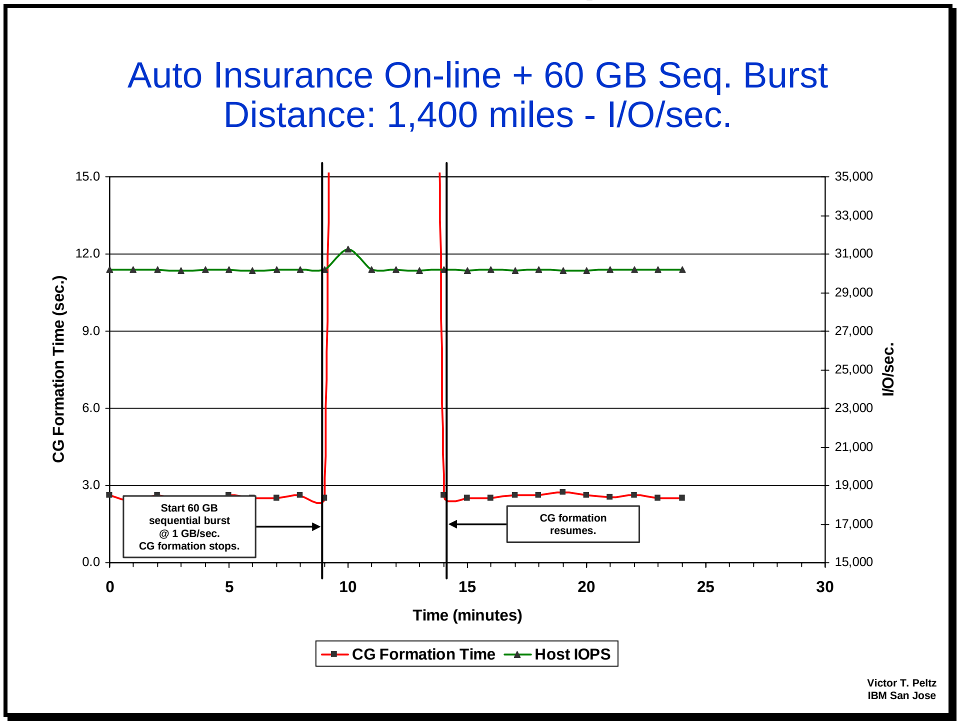


Figure 23: Auto Insurance On-Line Processing + 60 GB Seq. Burst - I/O/sec.

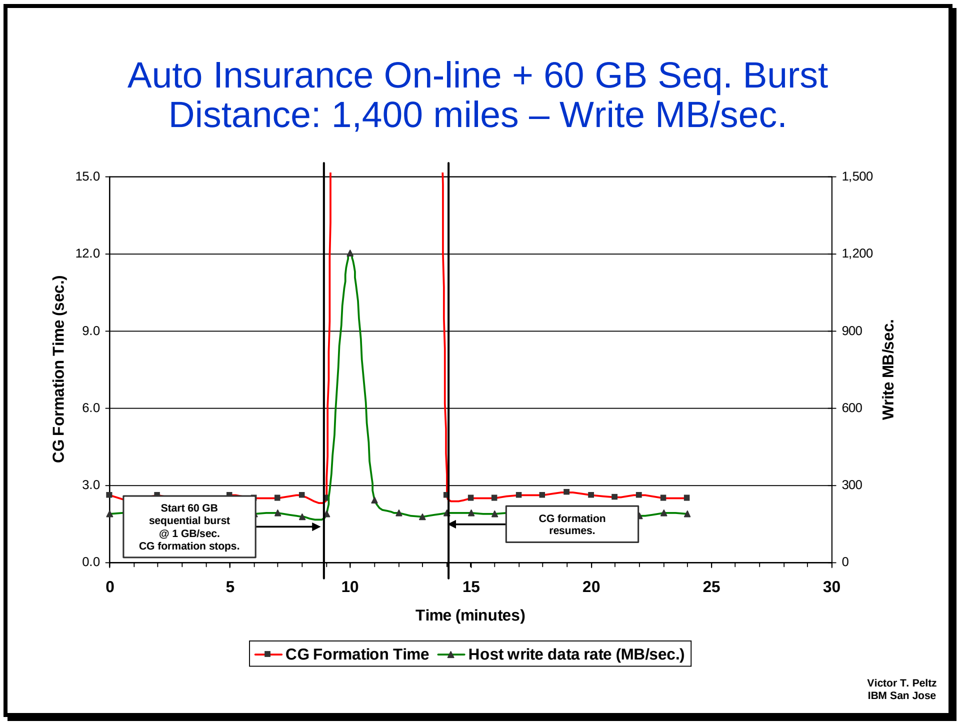


Figure 24: Auto Insurance On-Line Processing + 60 GB Seq. burst - MB/sec.

7.5.3.3 Automotive Insurance On-line: 50% Link Bandwidth

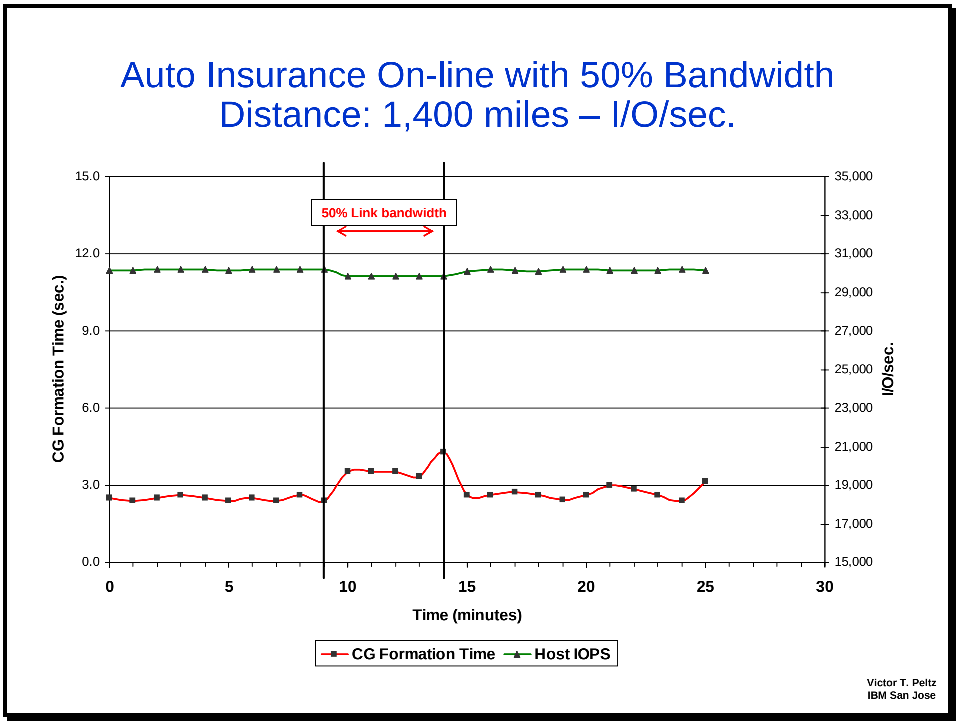


Figure 25: Auto Insurance On-Line processing - 50% link bandwidth - I/O/sec.

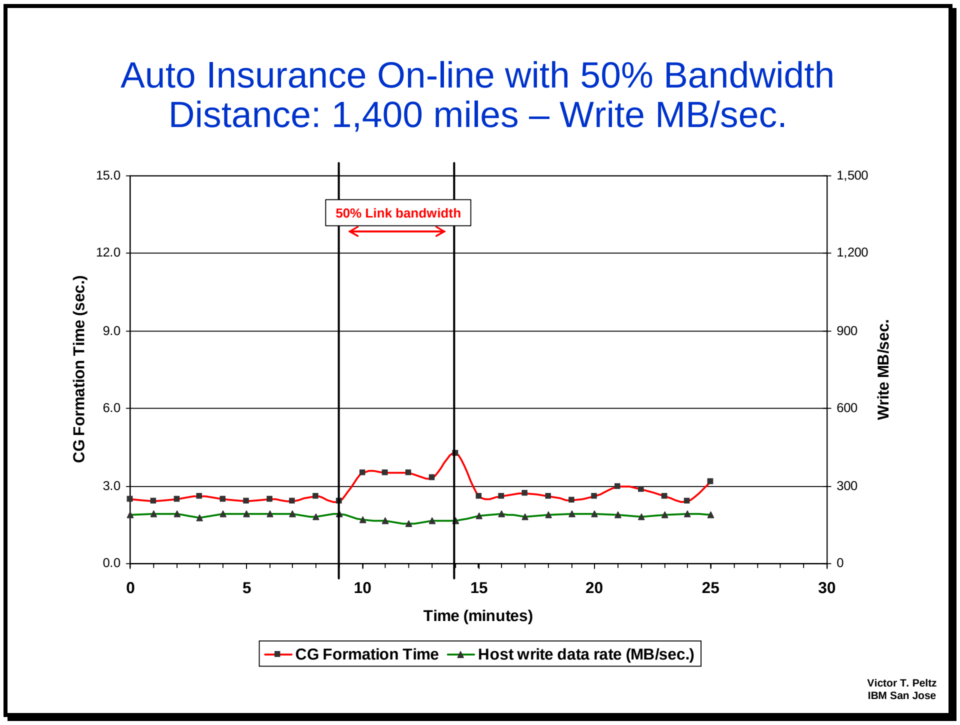


Figure 26: Auto Insurance On-Line processing - 50% link bandwidth - MB/sec.

7.5.4 Automotive Insurance: Batch

7.5.4.1 Automotive Insurance Batch Processing

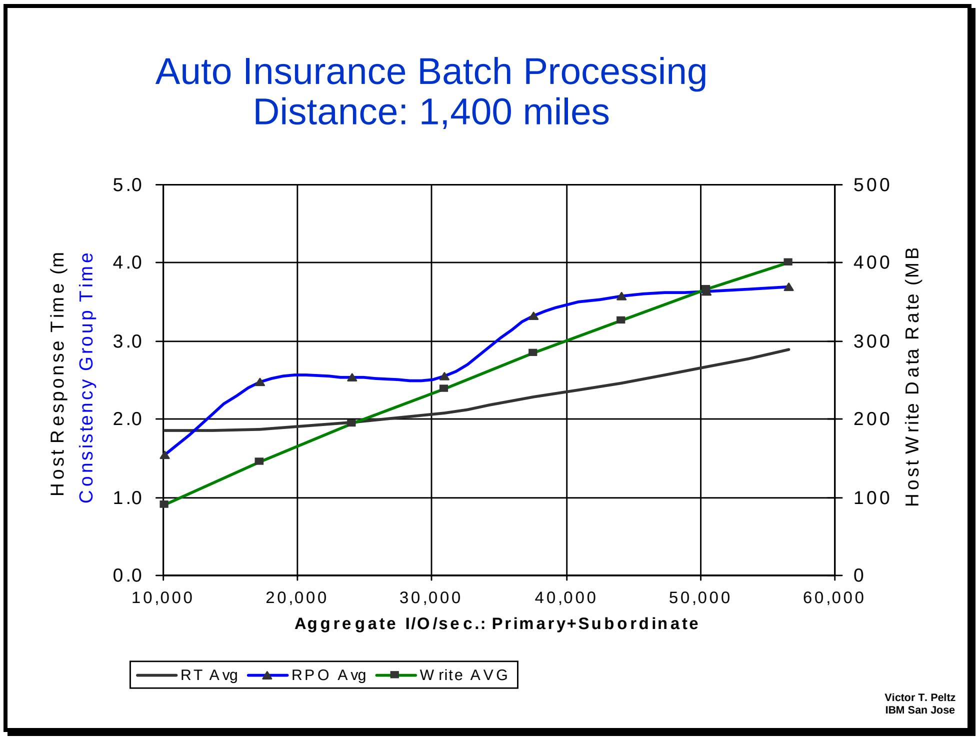


Figure 27: Auto Insurance Batch Processing

7.5.4.2 Automotive Insurance Batch + 60 GB Sequential Burst

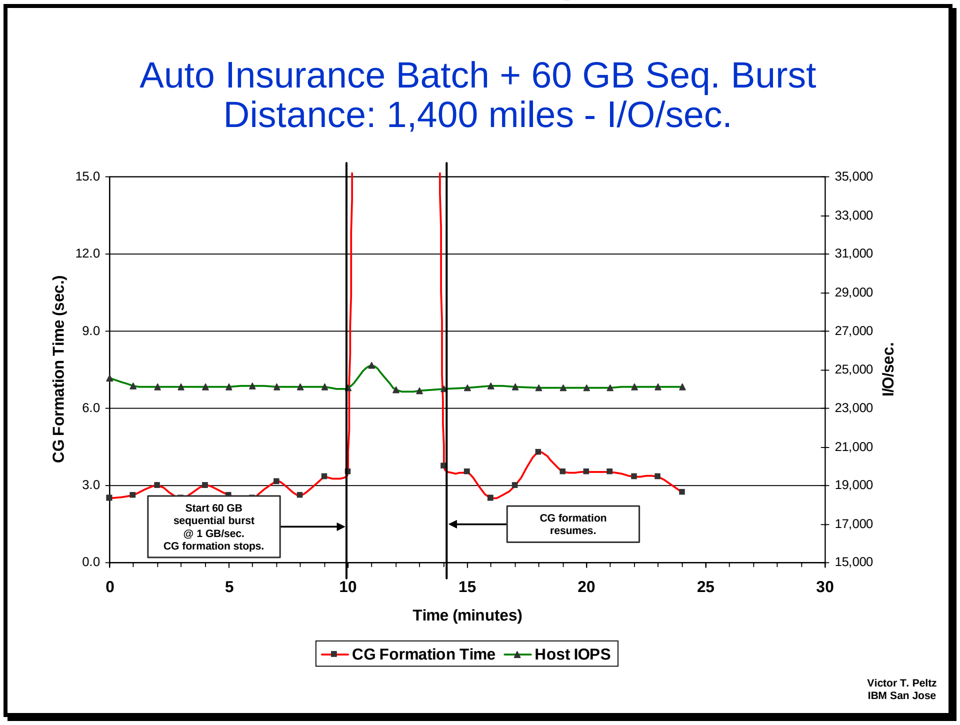


Figure 28: Auto Insurance Batch Processing + 60 GB Seq. Burst – I/O/sec.

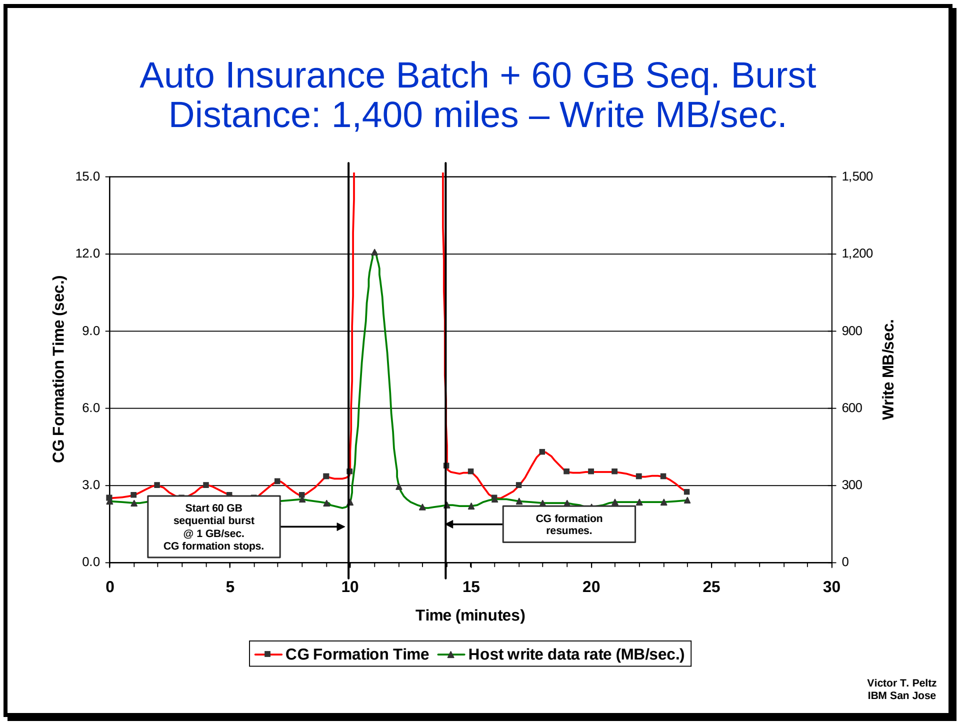


Figure 29: Auto Insurance Batch Processing + 60 GB Seq. Burst – MB/sec.

7.5.4.3 Automotive Insurance Batch: 50% Link Bandwidth

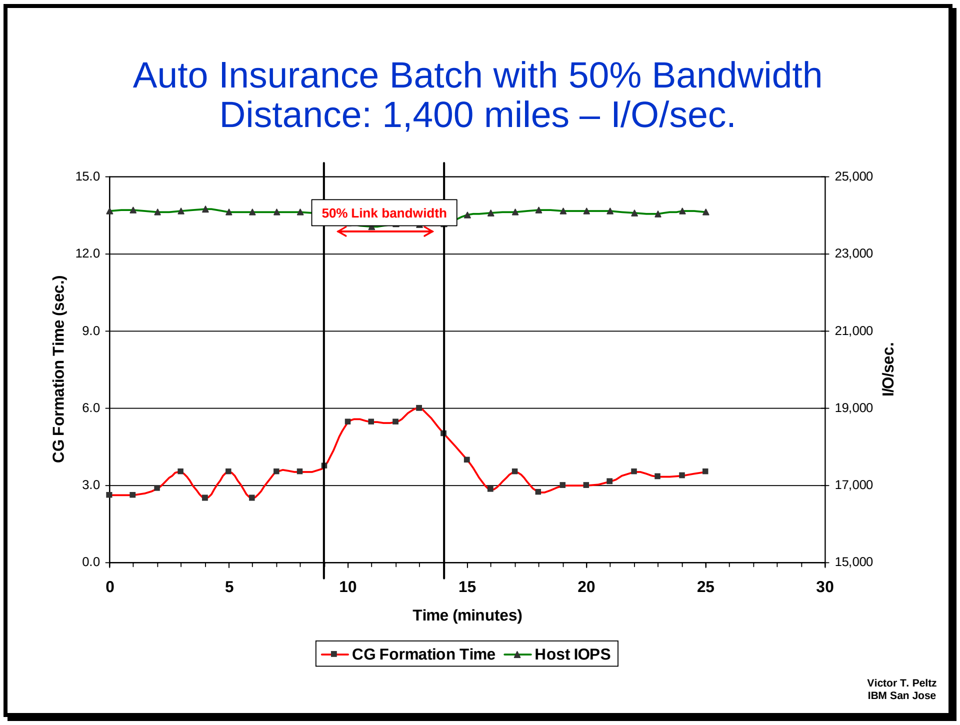


Figure 30: Auto Insurance Batch Processing - 50% Link Bandwidth - I/O/sec.

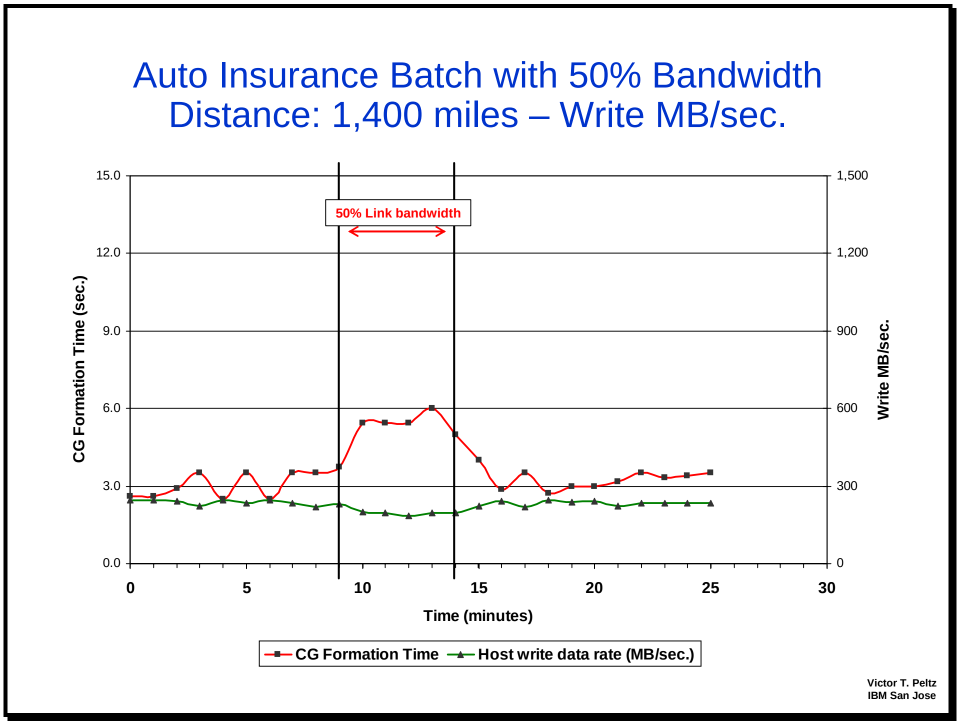


Figure 31: Auto Insurance Batch Processing - 50% Link Bandwidth - MB/sec.

7.6 DS8000 Global Mirror performance with FlashCopy

For this study, it was possible to employ either Traditional FlashCopy or “Space-efficient” FlashCopy[10] at the Global Mirror Secondary. Figure 32 and Figure 33 show the results of a previous Global Mirror study using a System p server and AIX. These measurements demonstrated that DS8000 Global Mirror performs essentially the same using either version of FlashCopy.

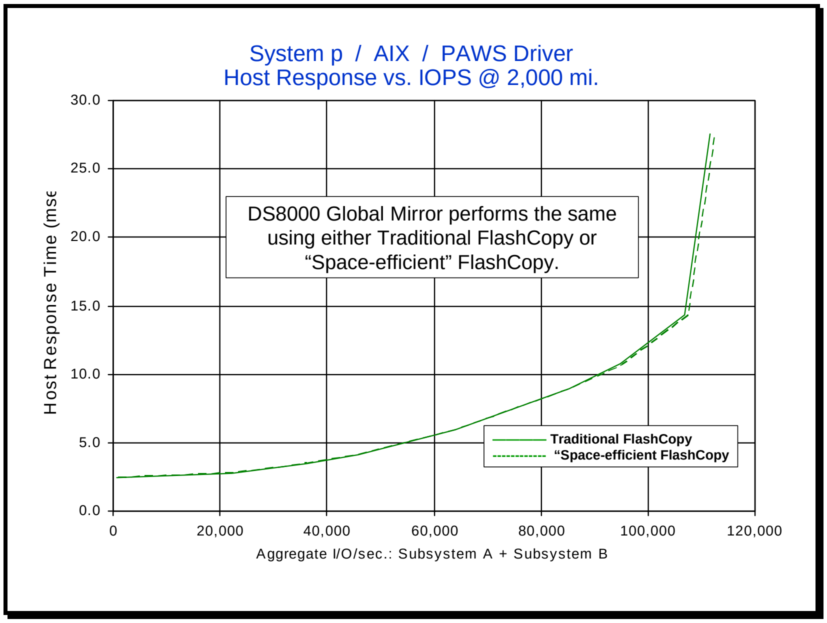


Figure 32: Host Response Time: Traditional vs. "Space-efficient" FlashCopy

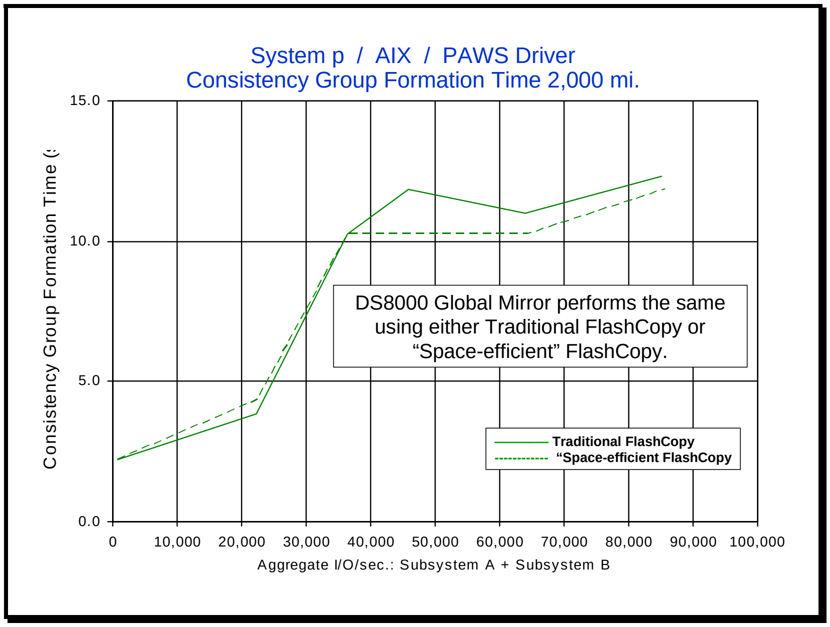


Figure 33: Consistency Group Formation Time: Traditional vs. "Space-efficient" FlashCopy

[10] See: reference [2] for additional information about Global Mirror operation and DS8000 FlashCopy.

7.7 Consistency Group Formation Time at a large financial institution

The pie charts in Figure 34 illustrate how an installation running DS8000 Global Mirror allows CGFT to increase for overnight batch processing. By allowing the CGFT for the overnight batch to increase compared to the prime shift workload CGFT, rather than merely installing links with more bandwidth, it is possible to achieve some savings in the Primary-to-Secondary link bandwidth costs.

In this example, for the 24-hour period, including the OLTP prime shift, the CGFT was less than 7 seconds 50% of the time, less than 15 seconds 21% of the time, etc. However, during the 10:00 p.m. to 2:00 a.m. overnight batch processing shift, the CGFT was less than 7 seconds only 14% of the time, and so forth.

This example illustrates how one may trade-off CGFT for link bandwidth savings. However, as with any trade-off, customers must evaluate their own needs carefully to determine the bandwidth that is needed consistent with their business objectives and requirements.

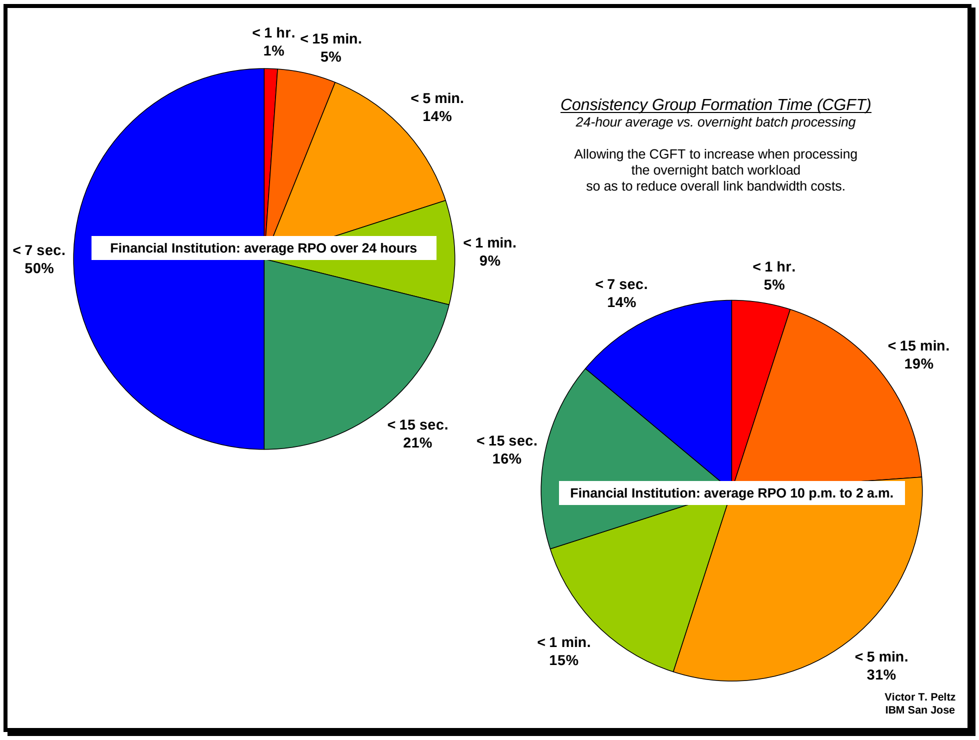


Figure 34: Financial Institution CGFT - 24-hour vs. overnight batch

References

1. An explanation of Recovery Point Objective (RPO) and Recovery Time Objective (RTO) may be found in "IBM System Storage Business Continuity Overview", SG24-6684.
2. Information about DS8000 Global Mirror may be found in "DS8000 Copy Services for System z", SG24-6787.
3. More information about Performance Associates, Inc. PAI/O Driver® for z/OS and their methodology for characterizing and simulating actual workloads can be found at: www.perfassoc.com.
4. Artis, H.P., "Workload Characterization Algorithms for Remote Copy Planning", Performance Associates, Inc., 2006. Available at: www.perfassoc.com/Published-Papers.html.
5. Artis, H.P. and Ross, R., "FICON I/O Measurements: Everything You Know is Wrong", CMG Proceedings, 2001.
6. Sherkow, Alan M., "Does Your Installation Have Half-Second I/O Response Time: A Closer Examination of Disconnect Time", CMG Proceedings, 1998.
7. Lin, A.W. and Peltz, V.T., "IBM Global Mirror Performance Study", August 2009. Available on the IBM Technical Sales Library, document WP101553.
8. Kern, R.F. and Peltz, V.T., "IBM Storage Infrastructure for Business Continuity", IBM Redpaper REDP 4605-0, October 2009.

About the Authors

Dr. H. Pat Artis
Performance Associates Inc.
P.O. Box 5080, Pagosa Springs, CO 81147-5080
drpat@perfassoc.com

Dr. H. Pat Artis is a recognized authority in workload characterization, forecasting, simulation modeling, I/O subsystem design, and capacity planning. He is the author of more than 100 papers and has lectured internationally on these subjects. His most recent investigations have concentrated on business reengineering, client/server solutions, management reporting, I/O configurations, graphic presentation techniques, storage management, and strategic planning issues for data processing installations. Dr. Artis is president of Performance Associates, a private consulting firm that specializes in serving the performance evaluation and strategic planning needs of large corporations. He holds degrees in Engineering Mechanics, Computer Science, and Information Sciences.

In addition to his other honors, Dr. Artis received the A.A. Michelson Award in 1984 for his fundamental contributions to computer performance evaluation. He is active in numerous performance-oriented organizations and is a Past President of the Computer Measurement Group. He is the co-author of the text "MVS I/O Subsystems: Configuration Management and Performance Analysis".

Victor T. Peltz
IBM Systems and Technology Group
Almaden Research Center
650 Harry Road, San Jose, California, 95120
vpeltz@us.ibm.com

Vic is a Consulting IT Architect with the IBM Corporation. He holds an M.A.Sc. in Electrical Engineering from the University of Toronto, joined IBM Canada Ltd. in 1975, and moved to the IBM San Jose Development Laboratory in 1992. Vic has worked for many years with IBM's Finance and Insurance customers, and in IBM's storage product marketing organizations in a variety of positions, including as a Product Manager and Enterprise storage competitive analyst. He has co-authored IBM Redbooks dealing with storage hardware, storage software and security. Vic also has published papers on system performance relating to DASD queuing algorithms and computer systems applications of magnetic bubble memories.

Vic is actively involved in the areas of Enterprise storage performance and Business Continuance and has been directly involved in working with organizations to help them recover from computing center disasters. The first occurred in 1969 and resulted in the total destruction of a university computer center as a consequence of a student riot. Over the years he has worked with numerous companies and governments to help them develop IT disaster recovery plans.

End of Document