IBM® System Storage™ DS8000™ Track space Efficient (TSE) Implementation Considerations and Recommendations

Lee LaFrese
Carol Mellgren
Shachar Fienblit

Updated January 3, 2017

Henry Sautter
Paul Spagnolo
Yan Xu

Systems Technology Group
December 5, 2007
Revised March 7, 2008
Revised September 5, 2008
Revised March 28, 2016

© 2016, International Business Machines Corporation

Notices, Disclaimer and Trademarks

Copyright © 2016 by International Business Machines Corporation.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation. Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. 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 NON-INFRINGEMENT.

IBM 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. 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 has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Statements regarding IBM's future direction and intent are subject to change or withdraw without notice, and represent goals and objectives only. The provision 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
U.S.A.

IBM, Enterprise Storage Server, ESCON, FICON, FlashCopy, System Storage, System p, z/OS, and DS8000 are trademarks of International Business Machines Corporation in the United States, other countries, or both. Other company, products or service names may be trademarks or service marks of others.

Abstract

This paper provides updated guidance and recommendations for using the FlashCopy SE (track space efficient) function on the DS8000. IBM announced FlashCopy SE in October 2007 and has introduced further enhancements in March 2008. This point in time copy licensed function is offered in addition to traditional IBM FlashCopy on the DS8000 (DS8000 models DS8100, DS8300, DS8700, DS8800 and DS8870). It allows the target capacity to be thinly provisioned (e.g. smaller than the full capacity of the source volumes). This is an attractive cost proposition but there are tradeoffs in terms of performance and function. This paper provides recommended use cases and information on sizing and configuring FlashCopy SE. It also includes recommendations about using FlashCopy SE with RAID-6 which became available on the DS8000 in September 2008. Do not confuse this function with other Thin Provisioning functionality such as Extent Space Efficient volumes provided for example in the DS8880 models. The DS8880 does not support Track Space Efficient (TSE) FlashCopy.

This paper may be updated to reflect further improvements to the product over time.

Important Differences between FlashCopy SE and Standard FlashCopy

The most important difference between FlashCopy SE and standard FlashCopy is space efficiency. With a FlashCopy SE relationship, disk space will only be consumed for the target when a write to source needs to be written to disk or when a write is directed to the target. For a source volume without much write activity during the duration of the Flash Copy SE relation, the target volume may consume significantly less physical space than the source. If a large portion of the source volume is written, there is little space savings over standard FlashCopy and additional overhead in management of the physical data. For this reason, background copy is not allowed on FlashCopy SE relations. Since there is no background copy, FlashCopy SE relationships can be used to recover from logic failures or database corruption, but not from hardware failures.

Because of space efficiency, data is not physically ordered in the same sequence on the target disk drives as it is on the source. A mapping structure is created to keep track of where the data on the target is physically located. When data is read from the target volume it may be retrieved from the source if it still resides there, just as it would in standard FlashCopy. If the data resides on the target, the mapping structure is used to locate it.

Since space is allocated as needed on the targets, data location on the target volumes is independent of data location on the source volumes. Essentially, all data on FlashCopy SE targets is stored in a random fashion and does not benefit from locality of data that may be expected on a standard FlashCopy target. Processes that might access the source data in a sequential manner may not benefit from sequential processing when accessing the target.

FlashCopy SE relationships may be failed when no space is left in the storage allocated for the targets or if the storage allocated for the targets becomes unusable. This is to prevent I/O failures on the source volumes when this occurs.

General Principles

Flashcopy SE is designed for temporary copies. Copy duration should generally not last longer than 24 hours unless the source and target volumes have little write activity. FlashCopy SE is optimized for use cases where a small percentage of the source volume is updated during the life of the relationship. If much more than 20% of the source is expected to change, there may be tradeoffs in terms of performance versus space efficiency. In this case, standard FlashCopy may be considered as a good

 © Copyright IBM Corporation, 2016

alternative. Since background copy would update the entire target, it would not make much sense and is not permitted with Flashcopy SE. Likewise, establishing a FlashCopy SE relationship just prior to running applications that make widespread changes to the source volumes (e.g. database reorgs, formats, full volume restores from tape, etc.) is not advisable.

When configuring space efficient storage in an extent pool two components are created, virtual space and repository space. Virtual space is the addressable capacity of all SE volumes which can be defined in an extent pool. Repository space is the physical disk capacity that serves as backing storage for the virtual space. The virtual space allocated must be at least two times the size of the associated repository space. When physical space for FlashCopy SE target volumes is needed, it is allocated from a repository common to all the space efficient targets in an extent pool.

FlashCopy SE relationships are made by creating SE volumes and then establishing FlashCopy using the SE volumes as targets. The SE volumes have no space allocated for user data until space is needed for destaging from cache. When space is needed for a SE volume, it is allocated on demand from the repository space.

Standard FlashCopy will generally have superior performance to FlashCopy SE although this does not mean that performance will be poor when using the function. Nearly all additional overhead to manage the FlashCopy SE relationship occurs when data is written to disk and not when host commands are issued so writes will generally not be delayed. However, if performance on the source or target volumes is of primary importance, standard FlashCopy is strongly recommended.

There are a number of alternative methods for creating and managing FlashCopy SE and standard FlashCopy relationships. These methods include DFSMSdss, TSO commands, ICKDSF, TPC for Replication, the DS Storage Manager (GUI) and API calls. FlashCopy Manager is a new product for managing FlashCopy under System z but it does not support FlashCopy SE at this time. Also please note that FlashCopy Manager does not currently support creating or managing FlashCopy relationships that include Extended Address Volumes (EAVs).

FlashCopy SE Repository

The FlashCopy SE repository provides the physical disk capacity that is reserved for space efficient target volumes. When provisioning a repository, Storage Pool Striping will automatically be used with a multi-rank extent pool to balance the load across the available disks. FlashCopy SE is optimized to work with repository extent pools consisting of 4 RAID arrays. In general, it is recommended that the repository extent pool contain between 1 and 8 RAID arrays. **Extent pools larger than 8 RAID arrays are not recommended for the repository.**

Once the physical and virtual space associated with an extent pool is configured, you are not able to expand them. It is important to configure as much physical and virtual space as you expect to need for all of the anticipated FlashCopy SE relationships. The only downside to configuring more virtual space than is actually needed is that a small amount of real space is consumed on the repository extent pool. This overhead is less than 1% of the virtual space capacity. The entire physical space specified for the repository will be deducted from the available capacity in the extent pool so there is a more noticeable cost for over configuring this.

Easy Tier is an automated tiering capability introduced on the DS8000 with the DS8700 models. All DS8700 and later models of the DS8000 support Easy Tier. Once the repository is defined, it cannot be moved, promoted nor demoted with Easy Tier. Also any track space efficient (SE) volume that is created in the repository is also not managed by Easy Tier since the volume is a virtual volume and thus cannot be moved out of the repository.

 © Copyright IBM Corporation, 2016

Since the repository space is backing storage for thinly provisioned volumes, it is important that adequate disk resources are configured to avoid creating a performance bottleneck. It is advisable to use the same disk rotational speed or faster (10K RPM or 15K RPM) for the target repository as for the source volumes. It is not recommended to ever use 7200 RPM disks for a space efficient repository. It is also recommended that the repository extent pool have as many disk drives as the source volumes do. One way to accomplish this is to use smaller capacity disks of the same speed for the repository compared to the source. This is applicable when dedicated repository capacity is desired to isolate it from performance sensitive application data. For example, if the source volumes are provisioned with 32 drives of 300 GB 15K RPM disks, space efficiency may be achieved by using 32 drives of 73 GB 15K RPM disks in the repository without sacrificing performance.

Note that the repository extent pool may also contain non-repository volumes. Although contention may arise if the extent pool is shared, this is often the most efficient way to implement FlashCopy SE in terms of overall performance. It is important to consider disk activity from FlashCopy SE as well as host activity when planning a configuration. If an extent pool will contain repository space and production volumes, then similar size, speed and quantity of disk drives is recommended for both the source volumes and the repository.

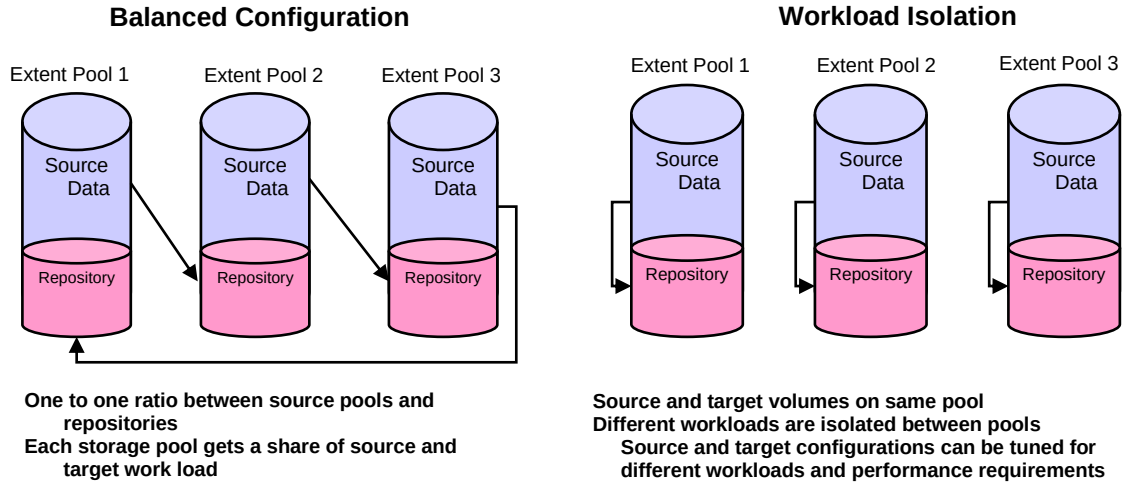

Figure 1 – FlashCopy SE Repository Recommended Configuration Alternatives

In most cases, using a balanced configuration of extent pools for source and repository is recommended as depicted on the left side of Figure 1 above. Note that each extent pool contains both source data and repository space but every relationship is defined across extent pools. In some cases it may be desirable to isolate workloads by extent pool as shown on the right side of figure one above. This is inherently less efficient than the balanced configuration but is another reasonable strategy when it is desirable to protect performance sensitive applications from being influenced by other applications or FlashCopy activity.

To be consistent with space efficiency goals, repository space will most commonly be configured with RAID-5 ranks. RAID-10 may actually perform better than RAID-5 in a repository since it is faster for random writes. RAID-10 should not be completely ruled out if performance is critical.

In September 2008 IBM introduced RAID-6 for the DS8000. RAID-6 is a double parity implementation designed to protect against data loss even after two concurrent drive failures or media errors on a disk array. This can be a significant reliability enhancement but there is a trade-off in terms of write performance. In general, a RAID-6 array has about two-thirds of the write throughput capability of a

 © Copyright IBM Corporation, 2016

RAID-5 array. There is no difference in read performance between RAID-5 and RAID-6. RAID-6 is typically most effective for use with high capacity disks holding primary, mission critical data. It is important to properly size a RAID-6 implementation for the expected write demand.

Workload planning is especially important when implementing RAID-6 for write intensive applications including copy services targets and FlashCopy SE repositories. For these and other heavy write use cases, you should expect to configure 50% more RAID-6 arrays than you would use for RAID-5. Thus you should carefully consider whether or not the additional reliability afforded by RAID-6 is really necessary in your environment. If FlashCopy SE is being used for temporary copies of transient data, does it really need RAID-6 reliability? Can the data be recovered in other ways in the very unlikely event of two concurrent failures? If indeed RAID-6 is required it can be configured to provide the necessary performance. If you need technical assistance planning for FlashCopy SE with RAID-6 repositories, please contact your IBM Storage sales representative.

Performance will be best when the number of source extent pools is the same as the number of associated repository extent pools. For example, each source extent pool should use target volumes on one repository extent pool that is not shared by other source extent pools. **Excessive fan in with multiple source extent pools using a single repository extent pool is not recommended** as it can create a performance bottleneck on the source due to contention on the repository. See Figure 2 below for examples of repository configurations to avoid.

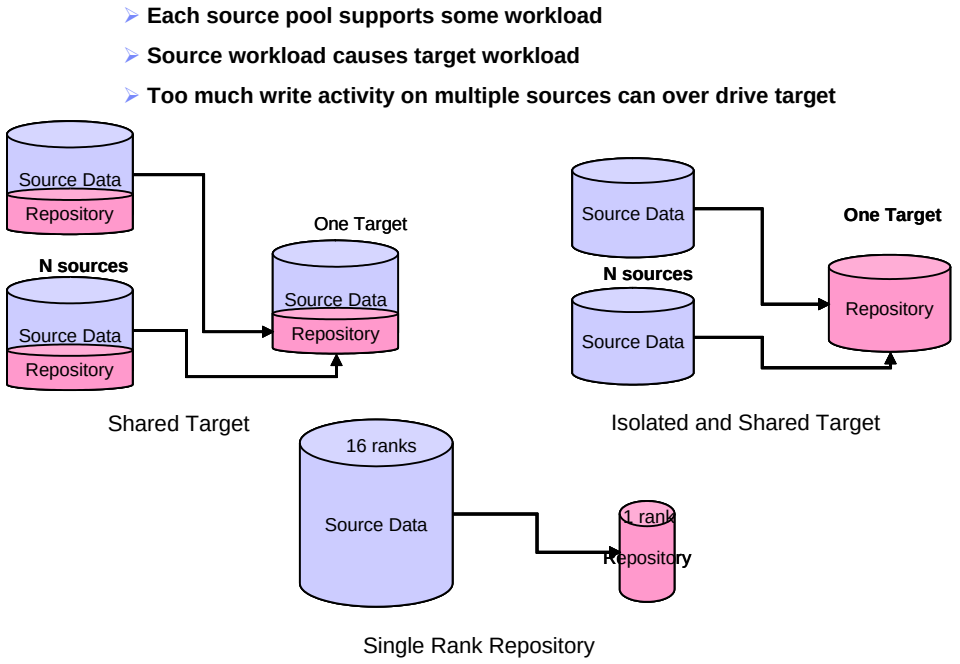

Figure 2 – Problematic FlashCopy SE Repository Configurations to Avoid

As mentioned before, once the repository is defined in the extent pool it cannot be expanded so planning is important to assure that it is configured to be large enough. If the repository becomes full, the Flashcopy SE relationships will be failed. After the relationship is failed, the target will become unavailable for reads or writes, but the source volume will continue to be available for reads and writes.

Repository space should be monitored to prevent filling it. When the free repository space goes below 15%, a warning is given to the user. This is done so the user may be proactive in releasing space. Notification may be requested for any percentage of free repository space, with a default notification at

 © Copyright IBM Corporation, 2016

15% free and 0% free. The 15% and 0% free notification will be presented, even if no user specified notification is requested. The notification will be via an SNMP trap. On System z, the notification will be sent via a console message. It is also possible to monitor the space usage of the repository and the space usage of each FlashCopy SE relation using queries.

The user must determine the amount of free repository space that is adequate. If the free repository space gets too low, action should be taken to free up space to avoid performance issues and failed relationships. Space in the repository is marked free when one of the following occurs:

- Withdraw SEFLC relationship with release space
- Release space command
- Establish SEFLC relationship (marks free any space on the target volume)
- Init CKD volume for a SE volume (may or may not be a target volume)

Space in the repository is freed asynchronously via a background reclaim task so it may not be immediately available. This allows withdraw, release, establish or init commands to complete quickly. While the background process to reclaim the free space is active, there may be a performance impact to the source volumes with targets on the same repository as the active reclaim as well as on the remaining target volumes.

Concurrent Code Load, Quiesce LPAR and Resume LPAR should not be initiated when a command that releases space has recently been issued. If a command that releases space has recently been issued over a large number of volumes, it is recommended not to create new SE relationships using that repository until some free space has become available. The time needed for the repository free space to be reclaimed varies, depending on the workload and the size of the repository.

If the workload changes on the source volume of a FlashCopy SE relation, or if the target volume is using more space than desired, it is possible to convert a Space Efficient target to a fully provisioned target using PPRC. This requires additional configuration steps and host ports to create a PPRC path with both source and target on the same storage subsystem. A complete procedure is beyond the scope of this paper.

You can estimate the physical space needed for a repository by using historical performance data for the source volumes along with some knowledge of the duration of the FlashCopy SE relationship. In general, each write to a source volume will consume one track of space on the repository (57 KB for CKD, 64 KB for FB). Over time there will likely be substantial rewriting - as much as 50% is typical. Thus the following calculation can be used to come up with a reasonable size estimate:

IO Rate x (% Writes/100) x ((100 – Rewrite %)/100) x Track Size x Duration in seconds x ((100+Contingency%)) = Repository Capacity Estimate in KB

Since it is critical to not undersize the repository, a contingency factor of up to 50% is suggested.

You can also use standard FlashCopy or PPRC Force Failover to estimate the change rate for the source data to be used with FlashCopy SE. This requires that you have the appropriate feature codes installed on an existing DS8000 system. For example, when running classic FlashCopy, the "lsflash" DSCLI command may be used to query the number of out-of-synch tracks on a volume at a point in time. This data can be used to estimate the number of tracks that have been copied since the FlashCopy was started. Ask your IBM sales representative for assistance if you are unsure of how to proceed.

Some Recommended Use Cases

The following are the most commonly recommended use cases to consider for FlashCopy SE. Remember that there may be tradeoffs in terms of performance for some of these use cases. A good test plan is

 © Copyright IBM Corporation, 2016

invaluable to validate that FlashCopy SE will meet all the requirements for a specific production environment.

- Create a temporary copy with FlashCopy SE to dump it to tape. This will likely be the most widely used example of FlashCopy SE.
- Temporary snapshot
  - Application development
  - DR testing
  - Testing with new hardware/software (system updates)
  - Data mining
- Online backup for different points in time
  - Multiple versions online at the same time
  - Multiple backups at periodic intervals (hourly, twice a day, multiple backups per day)

  Note: Online backups with a SE target can be used to recover from logical failures (data corrupted, unintentional deletions or updates). SE relations cannot be used to recover from hardware failures (i.e., source volume lost due to multiple DDM failures) since the SE target volume is not a self contained copy of the source volume.
- Checkpoints (only if the source volumes will undergo moderate updates).
- A remote point-in time copy can be created by establishing a FlashCopy SE relationship to an SE volume that is a source in a Metro Mirror relationship.
- The "C" volume in a Metro Mirror relationship. This is especially useful when Metro Mirror is suspended and then re-established after a relatively short time (up to several hours). The "C" volume serves as a safety copy in case of disaster during the re-establish process. **However, the re-establish process may take longer than it would if standard FlashCopy were used.**
- With the availability of new licensed internal code in March, 2008, FlashCopy SE may be used for the "C" or "D" volume in a Global Mirror (GM) relationship and also for the "D" or "E" volume in a Metro Global Mirror (MGM) relationship. See the section titled "Using FlashCopy SE with GM or MGM" for more implementation details.
- With the availability of new licensed internal code in March 2008, FlashCopy SE may be used to make up to twelve copies of a single target. Using multiple targets with FlashCopy can be very useful and powerful. For example multiple point-in-time versions of a database may be saved for backing out changes. However as the number of targets increase, more and more of the disk system's available bandwidth may be consumed by the copy process. This is especially true with FlashCopy SE which has less performance capability than classic FlashCopy. For example, with two targets the write throughput potential of the source volume will be roughly cut in half compared to a single target. With four targets, the throughput would be approximately one-quarter of what it would be with a single target. With the availability of RAID-6 function in September, 2008, more available bandwidth may be consumed when RAID-6 ranks are used compared to RAID-5 ranks, especially when used in FlashCopy SE relationships. The throughput potential of FlashCopy source volumes may be reduced up to 33% when the target volumes are on RAID-6 arrays instead of RAID-5

Figure 3 below illustrates a hypothetical example of how FlashCopy SE might be used. In this example FlashCopy SE is used to backup one volume to tape, a second and third volume are used to make periodic online backups and a fourth is copied for data mining or test purposes. All four targets are provisioned by the same repository.

 © Copyright IBM Corporation, 2016

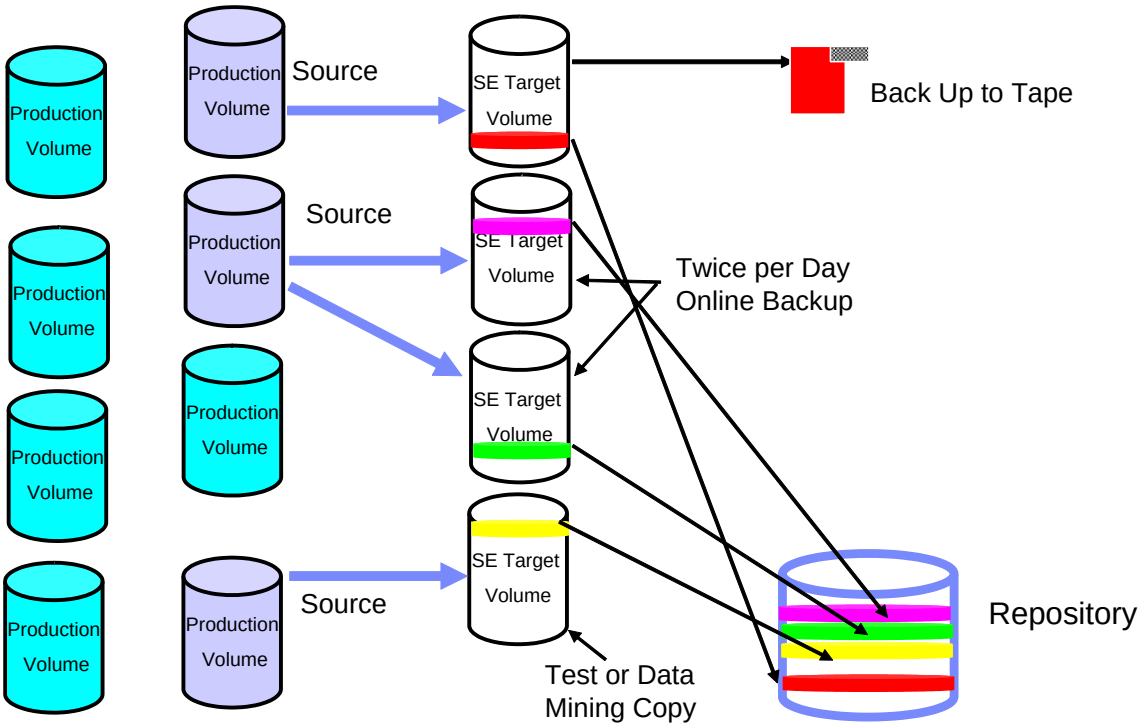

Figure 3 - FlashCopy SE Use Case Example, Backup/Test/Data Mining

Using FlashCopy SE with GM or MGM

Global Mirror (GM) and Metro/Global Mirror (MGM) are increasingly popular long distance asynchronous remote copy methods for the DS8000 platform. This section will discuss recommendations and best practices for using FlashCopy SE in conjunction with these advanced functions.

First let us review some frequently used terminology related to Global Mirror and Metro Global Mirror. In Global Mirror, the source volumes in the copy relationship are called “A” volumes. These typically hold production data that we wish to back up remotely and reside on the primary disk system. The “A” volumes are remotely mirrored to “B” volumes which reside on the secondary disk subsystem. The “B” volumes are in a FlashCopy relationship with “C” volumes which also reside on the remote secondary disk system. The FlashCopy is reinitiated each time a consistency group is formed. Thus the “C” volumes contain the consistent copy which may be used for disaster recovery. Optionally, some users create a “D” copy at the remote secondary. The “D” volumes, sometimes referred to as a “practice copy”, are a second consistent copy of the “A” volumes created via FlashCopy. The “D” volumes are normally used for testing or application development.

For Metro Global Mirror, the nomenclature is slightly different. There is a Metro Mirror copy that serves as the source volumes in the GM session. So the “A” volumes are still the production source, the “B” volumes are a synchronous copy of the “A” volumes. The “C” volumes are the GM targets of the “B” volumes. The “D” volumes are the FlashCopy targets of the “C” volumes. If a “practice copy” is used, it resides on the “E” volumes.

GM and MGM volumes may be good usage candidates for FlashCopy SE. This helps make these solutions more cost effective. However, using FlashCopy SE may introduce some risk to the solution. For example, if the GM relationship is suspended for some reason, changes at the primary will build up over time. If the amount of changed data were to exceed the repository size at the secondary, it may not be possible to form a new consistency group without removing the current consistency group. Thus there may be an exposure to disaster during the resynchronization.

 © Copyright IBM Corporation, 2016

The following guidelines should be considered when planning and implementing FlashCopy SE in a GM or MGM environment. Note that these guidelines are written using GM nomenclature. Replace the “C” with “D” and the “D” with “E” when applying these guidelines to MGM.

- Only the "C" (journal) volumes may be created with FlashCopy SE. It is not recommended that the “D” volumes (practice volumes) be created with FlashCopy SE since it is possible that during a long D/R test or actual recover event could be prolonged and use all allocated thin provisioned space.
- No additional FlashCopies (SE or otherwise) of the “B” volumes should be made besides the "C" or "D" volumes if either "C" or "D" are created with FlashCopy SE. Additional FlashCopies may increase the time between consistency groups and result in a longer Recovery Point Objective (RPO).
- When using FlashCopy SE with Storage Pool Striping on a GM secondary disk system, the SE "C" volume should be placed in a different extent pool than it's corresponding source "B" volume.. You may mix "B" volumes with different "C" volumes in the same extent pool. Additionally, if “D” volumes are used, they should be placed in a different extent pool than their corresponding “B” and “C” volumes. You may mix different “B”, “C and ”D” volumes within the same extent pool. This guideline should be considered a configuration goal but not a firm rule. It may not always be possible to achieve this for every SE relationship depending on the overall DS8000 rank configuration.
- When planning to use FlashCopy SE with GM, the following rules of thumb may be applied to ensure that GM performance is not negatively affected:
    - For a single RAID array, the maximum write throughput should not exceed 20 MB/sec or 300 IO/sec.
    - For a single extent pool of four or more RAID arrays, the maximum write throughput should not exceed 80 MB/sec or 1200 IO/sec
    - For an entire secondary DS8300 the maximum total box write throughput should not exceed 200 MB/sec or 5000 IO/sec. For a DS8100 use half these maximum recommended full box figures. (Refer to the tables below)

- During a resynchronization (after network failure, etc.), repository space for “C” volumes will be consumed until the resynchronization is complete. Therefore, the space configured in the repository should be enough to cover any anticipated network outage duration in order to maintain GM operation. Note: the consistent copy will be preserved even if the repository fills, but it will not be possible to complete the resynchronization if the repository is full without first removing the GM FlashCopy or releasing space in the repository in some other way.
- A mixture of FlashCopy SE and classic FlashCopy may be used for different GM "C" volumes within the same GM session or on the same secondary DS8000. This may be a good approach for environments with a few "hot" volumes where a fully provisioned target may be more efficient and perform better.
- If a long Consistency Group Interval (CGI) is set (i.e. several minutes or more) FlashCopy SE may not be a good choice for use with GM “C” or “D” volumes
- There is a procedure for converting GM “C” volumes from FlashCopy SE to classic FlashCopy or vice-versa. A complete procedure is beyond the scope of this paper.

Recommendations and Guidelines for Implementation

The following are other current guidelines and recommendations for FlashCopy SE.

- Both the source and target volumes of any FlashCopy SE relationship should always be on the same server cluster in the DS8000.
- FlashCopy SE is supported for use with Global Mirror (GM) or Metro Global Mirror (MGM) relations. For GM, only the C (journal) copy should be defined as a SE volume.
- For MGM, only the D (journal) copy should be defined as a SE volume.
- Virtual Capacity must be at least two times the size of the associated Repository size.

 © Copyright IBM Corporation, 2016

- You may want to allow the virtual capacity to be automatically assigned rather than specifying the virtual capacity size while initially defining the repository.
- A maximum of twelve FlashCopy SE multi-target relationships per are allowed to be used at any one time – when not in a GM/MGM relationship. Performance may be reduced when running twelve targets, depending on the configuration and the workload and system recovery may take longer. When a rank is configured with FlashCopy SE multi-target relationships, the performance for all volumes that share the rank may be affected – even those not involved in any FlashCopy relation**.** An isolation strategy may help avoid this situation when FlashCopy SE multi-target is used
- There are no restrictions on the amount of virtual space or the number of SE volumes that can be defined for either CKD or FB system storage. The following table lists recommendations based on the typical environments used in IBM's internal testing:

Storage type	Number of volumes	Number of SE targets per volume	Repository size	Number of ranks in extent pools containing repositories
CKD	8000	1	5 to 10 TB	4 to 8
CKD	500	1 to 12	5 to 10 TB	4 to 8
FB	1000	1 to 12	1 to 10 TB	4 to 8
GM/MGM CKD	2500	1 to 2	1 to 5 TB	4 to 8
GM/MGM FB	2500	1 to 2	1 to 5	4 to 8

The following lists some performance guidance at the machine level for the DS8000 models which support Track Space Efficient FlashCopy:

DS8000 Model	Maximum Throughput Limit for SE FLC
DS8100	100 MB/second or 2500 I/O per second
DS8300	200 MB/second or 5000 I/O per second
DS8700	350 MB/second or 9000 I/O per second
DS8800	350 MB/second or 9000 I/O per second
DS8870	550 MB/second or 12000 I/O per second

- Although not the typical test environment, IBM has tested:
  - A single 32 rank repository
  - 32 single rank repositories
  - A 25 TB repository
  - 2000 CKD volumes and 2000 FB volumes with 12 targets each

 © Copyright IBM Corporation, 2016