



Global Distributed Storage Engineering
Citi Architecture & Technology Engineering

**Enginuity 5773 Beta Testing – Key
Findings and Benefits**

GDSE-08011-E

Evaluation

Version 1.0

CONTENTS

Legal Notice	5
Contact	5
Latest Versions	5
Acknowledgements	5
1 NEW SRDF FEATURES SUMMARY	6
2 EMC VIRTUAL (THIN) PROVISIONING SUMMARY	8
3 APPENDIX A - NEW SRDF FEATURES IN 5773	11
3.1 Cascaded SRDF & Cascaded Star	11
3.1.1 Description	11
3.1.2 Key findings	13
3.1.3 Recommendations	15
3.2 Moving dynamic RDF pair devices	16
3.2.1 Description	16
3.2.2 Key findings	16
3.2.3 Recommendations	16
3.3 SRDF/A R2 timestamp representation	16
3.3.1 Description	16
3.3.2 Key findings	17
3.3.3 Recommendation	17
3.4 Minimum cycle time for SRDF/A	17
3.4.1 Description	17
3.4.2 Key findings	17
3.4.3 Recommendation	17
4 APPENDIX B - HIGH-LEVEL COMPARISON OF SRDF 3-SITE SOLUTIONS	18
5 APPENDIX C - ENGINEITY 5773 VIRTUAL PROVISIONING	20
5.1 Introduction	20
5.2 I/O Functionality	20
5.3 Observations on Virtual Provisioning	21
5.4 Summary	21

6 APPENDIX D – EMC VIRTUAL PROVISIONING SUMMARY TEST RESULTS	22
6.1 Throughput Performance	22
6.1.1 Positives	22
6.1.2 Negatives	22
6.2 Drain Performance	22
6.2.1 Positives	22
6.2.2 Negative	23
6.3 Capacity Utilization	23
6.3.1 Positives	23
6.3.2 Negatives	23
6.4 Storage Pool Alerting	23
6.4.1 Positives	24
6.4.2 Negatives	24
6.5 Storage Pool DATADEV Removal	24
6.5.1 Positives	24
6.5.2 Negatives	24
6.6 Storage Pool Data Security	24
6.6.1 Positives	25
6.6.2 Negatives	25
6.7 Circumvention of Virtual Provisioning Objectives	25
6.7.1 Positives	25
6.7.2 Negatives	25

Legal Notice

Information contained herein is for Internal Use and may be used only for business purposes authorized by Citigroup.

CITI, CITIBANK, CITICORP and CITIGROUP are trademarks and service marks of Citicorp and are used and registered throughout the world. Citibank, Citicorp, Citigroup and their subsidiaries also claim rights in certain other trademarks and service marks contained in these pages.

Contact

Please address any comments or questions regarding the information contained in this document via email to the “*GT Global Distributed Storage Engineering” distribution list or alternatively to gdse@citigroup.com.

Latest Versions

It is recommended that the latest version of the document be used from the GDSE website

<http://engineering.citigroup.net/gdse/products/>

Newer versions may include expanded scope, or comments on issues discovered after deployment.

Acknowledgements

The authors would like to thanks Donal DeBurca (EMC) and other members of the EMC 5773 Beta team for their assistance with this work.

Some of the product description material is taken from various EMC documents.

1 New SRDF features summary

Enginuity 5773 combined with Solutions Enabler 6.5 offers several new SRDF features:

- 1) **Cascaded SRDF:** Cascaded SRDF is a new 3-site disaster recovery configuration where data from a primary site is synchronously replicated to a secondary site, and then asynchronously replicated to a tertiary site.

Benefits of Cascaded SRDF include it being the simplest (least sophisticated) 3-Site solution that offers the potential of zero data loss at Site C should Site A fail. It is also the most scalable SRDF-based 3-site solution. However, when used outside of SRDF/Star control it has a number of significant recoverability restrictions.

Cascaded SRDF may also be of value for temporary usage to aid with migrations.

It is recommended that Cascaded SRDF only initially be certified to be run under SRDF/Star. Use of “native” cascaded SRDF (not under Star control) should only be considered for certification if demand for 3-site solutions in Citi outstrips the scalability of SRDF/Star.

- 2) **Cascaded Star:** An enhancement to Cascaded SRDF that adds:

- Differential resynchronization capabilities of the surviving sites
- Automation of site switch tasks via the symstar command
- Dynamic reconfiguration to a concurrent configuration to deal with various failure scenarios

Hence, Cascaded Star addresses the recoverability limitations of “native” Cascaded SRDF.

It is recommended that certification of cascaded SRDF initially be limited to dealing with certain failure scenarios with Concurrent Star configurations, rather than full certification of Cascaded Star as a normal operating configuration.

- 3) **Moving RDF pair devices.** Provides the ability to move dynamic RDF pair devices between RDF groups without a full resynchronization.

It is strongly recommended that Citi Best Practice be updated to use this capability to reduce the overall time taken to allocation SRDF/A storage to hosts.

- 4) **SRDF/A R2 timestamp.** An actual timestamp, indicating the point-in-time that the SRDF/A R2 device represents can now be obtained both before and after an SRDF/A failure has occurred, and from either the source or target site. This can be provided to application owners after a disaster to help with quantifying data loss in a disaster scenario.

It is strongly recommended that Citi Best Practice be updated to use this new capability.

- 5) **Minimum Cycle time for SRDF/A.** This has now been reduced from 5 seconds to 1 second.

It is not recommended that this new minimum cycle time be adopted.

For further details, see the Appendices A & B of this document. Detailed test results can be found in the following document:

<http://engineering.citigroup.net/gdse/products/platforms/symmetrix/techinfo/08004-tr.pdf>

2 EMC Virtual (Thin) Provisioning summary

As a part of the EMC Symmetrix 5773 Beta evaluation, GDSE has tested the new virtual (thin) provisioning features. Overall, the EMC Virtual Provisioning (EVP) product (at this point a few days after GA release) is immature and not ready for production as delivered without either additional product development by EMC, or interim scripting and process development by Citi. The latter is recommended to avoid delaying unnecessarily use of this technology.

EMC's primary competitor for thin provisioning on enterprise storage is HDS. Citi has looked at other thin provisioning vendors (such as 3PAR), but were eliminated due to proprietary hardware or other critical aspects.

Conceptually, creating thin-provisioning LUNs on the EMC Symm is similar to HDS in that storage is assigned to pools and the thin LUNs are created and associated with a pool. Each time a thin LUN requires additional write space, more pages of the pool storage are allocated to the thin LUN. Both HDS and EMC have a significant application downtime penalty associated with migrations from standard to thin storage. This may be alleviated somewhat with new tools such as EMC's PPME. New applications using thin storage would not have a migration penalty.

ECC will not support 5773 features (including thin provisioning) until version 6.1 and Citi is not expected to support this version until 1H 2009. There may be a patch for 6.0 available sooner, but it will not support the thin features. However, although the current Citi released version of ECC 5.2(Sp5) identifies the HDS thin LUNs as the thin DP-Vols, it is not aware of utilization of the thin volumes. Citi may need to consider supporting ECC 6.1 sooner than its current adoption timetable.

There is a minor feature that EMC has which is not present in the HDS version. That feature is pre-allocation. This feature allows for pages within the thin provisioning volume to be pre-assigned to the LUN. In theory this would improve the initial write access to the thin LUN, but would go against the purpose of the thin LUN as there would be allocation without usage. Any advantage that this would provide would disappear as soon as the LUN is used. Citi has not found a good use-case for this feature.

EMC has significant work to do to make this a production Citi-ready product. The two main areas are:

1. With no reporting and limited alerting, there are significant operational risks associated with deployment. The thin pools could run out of storage unless significant scripting is developed. This would effect all applications sharing the same pool. There are a good number of system based thresholds (60,65,70,80%) that would provide warning, but these thresholds do not have array-based alerting. It is possible to set up host-based scripting that could possibly generate SNMP traps, but that scripting would be significant and may require additional software (SNMP) to run on a symcli server. In addition, a unique SNMP MiB may need to be created just for these alerts to come from a host. ECC normally generates SNMP traps for arrays and as discussed above is not ready for Thin Provisioning.
2. As there is no good way to report on the trending of the utilization of pools, it would be difficult to predict when a pool will run out of storage. This is important as it may take days or weeks to obtain additional storage to backfill usage. In addition, it would be harder to see a pattern of over-use by an unfriendly application. There are no new security risks with thin provisioning. EMC Thin provisioning does not yet support RAID5 storage at this time. RAID6 (supported) is more important to Citi long term, but the current Citi standard storage configuration is RAID5.

The following is a summary of the EMC vs. HDS functionality:

	Feature Description	EMC Symmetrix DMX 3,4	HDS TagmaStore USPV
1	Thresholds (system)	Yes (60,65,70,80%)	Yes (70%)
2	User-Definable Thresholds*	Yes**	Yes (2 Levels)
3	Alerts (based on thresholds)*	Yes**	Yes (SNMP, email)
4	Trending Reporting*	No	Yes (Tuning Mgr)
5	Chargeback Reporting (ECC)*	No	Yes (only DP-Vols, not % utilized)
6	Supports RAID5 (7+1)	No	Yes
7	Supports RAID6 (6+2 or 14+1)	Yes (both)	Yes (6+2 only)
8	Supports RAID10 (4+4)	Yes	Yes
9	Pre-Allocate Pages	Yes	No
10	Growth of pools	Manual	Manual
11	Restriping of pool data	No	Q4 2008
12	Thin-friendly analysis tool	No	No
13	Internal Array Replication Compatible	No	Yes
14	Array to Array Replication Compatible	No	Yes

* =key criteria

**=requires host-based monitoring/action scripting via symcli

The key items that EMC need to fix before it can be considered ready for Citi are:

- User Definable Array-Based Thresholds
- Threshold Array-Based Alerting (SNMP)
- Trending Reporting
- Chargeback Reporting (ECC)

If these items are corrected, then the product will be as ready as the HDS product was when Citi internal pilot began.

In order to speed up implementation of the EMC product withi Citi, the following could be developed by Citi staff as interim solutions:

- Scripted monitoring for User Definable Array-Based Thresholds
- Scripted monitoring for Threshold Array-Based Alerting (SNMP or log scrapping)
- Scripting for Trending Reporting

If this scripting is developed, then the current operational deficiencies of the product would be minimized. Chargeback would need to be based on the thin volumes, not what is utilized.

GDSE recommends the following next steps:

- Gain further understanding of which application and filesystem types are suitable for use with Virtual provisioning.
- Begin certification of EVP, but not until after a service pack release of Solutions Enabler 6.5 and the 73 code. This certification would need to include the development of scripting described above. The certification would need to include specific guidance on use cases for EVP as this technology is not yet suitable for ubiquitous usage.
- Investigate host-based thick-to-thin migration tools such as PPME for existing host migrations to thin storage.

For further details, see the Appendicies C & D of this document. A further detailed technical report on EMC Virtual Provisioning can be found at:

<http://engineering.citigroup.net/gdse/products/platforms/symmetrix/techinfo/08017-e.pdf>

3 Appendix A - New SRDF features in 5773

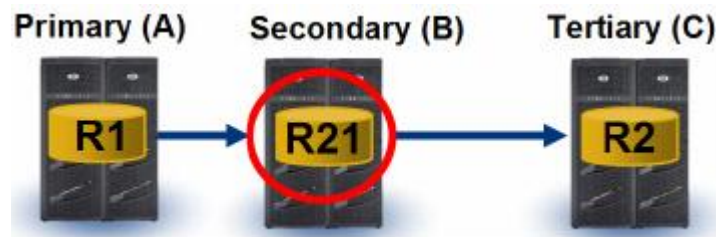
3.1 Cascaded SRDF & Cascaded Star

3.1.1 Description

Cascaded SRDF is a new three site disaster recovery configuration where data from a primary site is synchronously replicated to a secondary site, and then asynchronously replicated to a tertiary site.

Prior to Enginuity 5773, an SRDF device could be a primary device (R1 device) or a secondary device (R2 device); however it could not be in the dual role simultaneously. Cascaded SRDF introduces a new SRDF R21 device. The R21 device will assume dual roles of primary (R1) and secondary (R2) device types simultaneously. Data received by this device as a secondary can automatically be transferred by this device as a primary.

The basic cascaded SRDF configuration consists of a primary site (site A) replicating synchronously to a secondary site (site B) with SRDF/S, and then replicating the same data asynchronously to a tertiary site (site C) with SRDF/A.

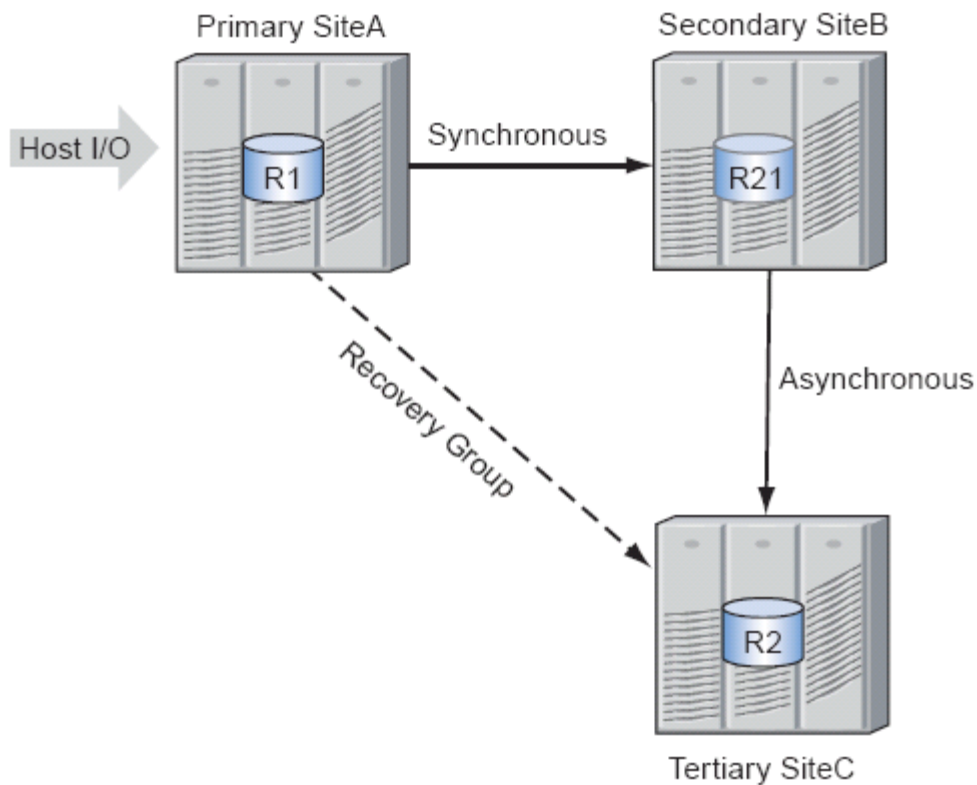


Cascaded SRDF devices are supported only on the Symmetrix DMX-3 and DMX-4 hardware platforms with Enginuity 5773 and above; however, only the Symmetrix that contains the R21 device is required to be Enginuity 5773. The Enginuity level for the Symmetrix containing the R1 and the R2 can be any level that EMC supports being connected to Enginuity 5773.

Cascaded Star provides the following significant benefits over “native” cascaded SRDF (i.e. a cascaded SRDF setup not controlled using Star):

1. Failover automation for Cascaded SRDF. Failover of Cascaded SRDF can be automated using home-grown scripts.
2. The ability to reconfigure a cascaded SRDF configuration into a concurrent SRDF (Star) configuration to allow replication directly from Site A to Site C should replication via Site B fail.
3. The ability to reconfigure a cascaded SRDF configuration into a concurrent SRDF (Star) configuration to allow on-going replication directly from Site C to Site A for a planned switch to Site C (e.g. after a server failure at Site A).
4. Differential resynchronization capabilities between the surviving sites after loss of a site.

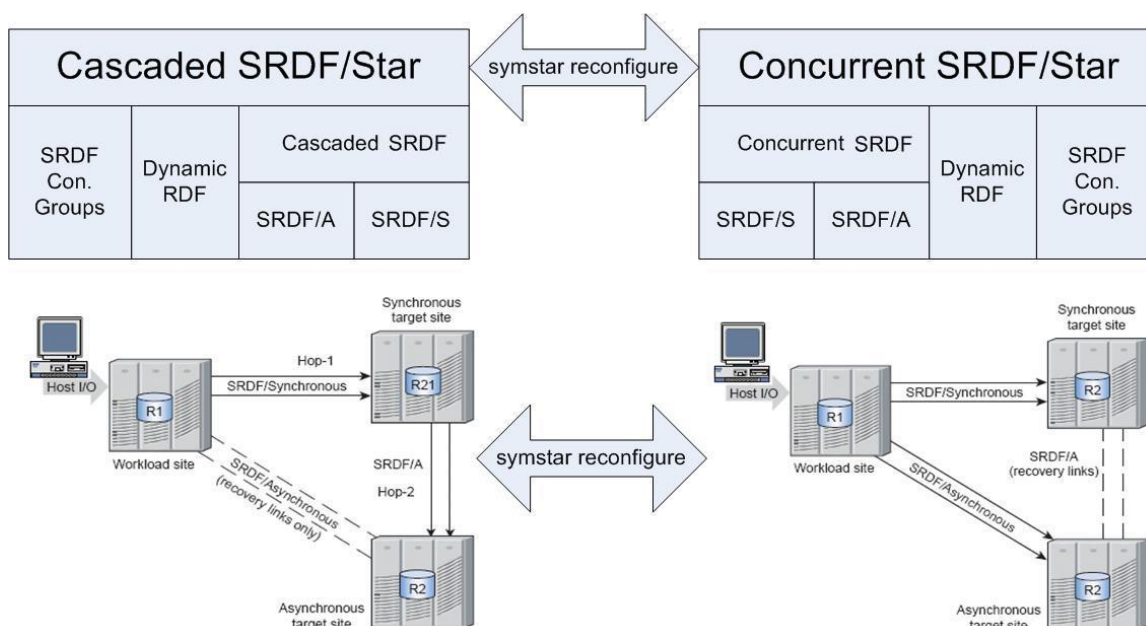
5. With Cascaded SRDF, asynchronous replication cache usage load is moved off of the production system.



Cascaded Star requires 5773 microcode at Sites A and B. “Native” Cascaded SRDF only needs 5773 at Site B.

The following diagram shows:

1. How Cascaded and Concurrent Star both are dependent on other SRDF technologies.
2. How the symstar reconfigure command can be used to move between the cascaded and concurrent configurations. This is vital in a Cascaded configuration as:
 - a. Loss of Site B (Synchronous Target), or links to or from Site B ($A \rightarrow B$ or $B \rightarrow C$) would lead to loss of DR protection. A reconfiguration to a Concurrent configuration to allow replication directly from A to C is necessary.
 - b. In a planned failover to Site C, replication back to Site A would require reconfiguration to a Concurrent configuration to allow direct SRDF/A replication from $C \rightarrow A$, as $C \rightarrow B \rightarrow A$ replication does not result in a consistent image at Site A.



3.1.2 Key findings

A table comparing four different SRDF-based 3-Site configurations can be found in Appendix B.

The key findings for the new configurations only ("Native" Cascaded SRDF and Cascaded Star) are discussed below.

3.1.2.1 Key findings - "Native" Cascaded SRDF (no Star control)

1. **Positive: Simplicity.** "Native" Cascaded SRDF is the simplest, entry-level 3-Site solution, that offers the potential of zero data loss should Site A fail. It is a replacement for the older "multi-hop" solution.
2. **Positive: Scalability.** "Native" Cascaded SRDF is more scalable than SRDF/Star (at time or writing). "Native" Cascaded SRDF scalability is limited by 2-Site SRDF/A limits, although Site B needs 2 RDF groups per cascaded configuration limiting the total number for cascaded configurations nominally to 64 replication groups per set of 3 arrays. Cascaded Star is limited to 32 replication groups per set of 3 arrays.
3. **Major Negative: Flexibility.** With "native" Cascaded SRDF, a failure of any of loss of Site B, the A – B link or the B – C link will stop replication to Site C. With "Native" Cascaded SRDF, this would cease replication to our COB/DR site (Site C).

3.1.2.2 Key findings - Cascaded Star

Cascaded Star refers to a cascaded SRDF configuration under Star control.

4. **Positive: Flexibility:** As indicated above, “Native” Cascaded SRDF has a number of undesirable restrictions. These can be addressed by running cascaded SRDF under Star control (i.e. in a Cascaded Star configuration).
 - i. The ability to reconfigure a cascaded SRDF configuration into a concurrent SRDF (Star) configuration to allow replication directly from Site A to Site C should replication via Site B fail is only available when a cascaded configuration is under Star control (via symstar command).
 - ii. The ability to reconfigure a cascaded SRDF configuration into a concurrent SRDF (Star) configuration to allow on-going replication directly from Site C to Site A for a planned switch to Site C (e.g. after a server failure at Site A) is only available when a cascaded configuration is under Star control (via symstar command).
 - iii. Differential resynchronization capabilities between of the surviving sites after loss of a site only available when a cascaded configuration is Star protected.
5. **Positive: Automation.** Failover automation for Cascaded SRDF is available via symstar when under Star control.
6. **Negative: Scalability.** Scalability of a cascaded SRDF configuration is reduced when running under Star control. This is primarily due to the scalability of the RDF daemons, as the number of Star sessions increases, and the number of SDDF sessions that a Symmetrix can support (SDDF is the Symmetrix Differential Data Facility that is used for incremental updates in SRDF).
7. **Negative: RDF daemon.** As with all SRDF/Star options, SRDF/A leg cycle switching is controlled by RDF daemons that run on external servers. These daemons must be running for replication to take place and physical servers for these daemons must be provided (an issue if a single pair of RDF daemons cannot deal with the cycle switching load of the SRDF/Star installed base). “Native” cascaded SRDF offers 3-site nominal “zero-data loss” without the use of RDF daemons.

3.1.2.3 Key findings - Generic Cascaded SRDF comments

The following points apply to both “native” cascaded SRDF and Cascaded Star.

8. **Positive: Loading.** With Cascaded SRDF, asynchronous replication cache usage load is moved off of the production system.
9. **Positive: Migrations.** Cascaded SRDF may be of use to help with migrations. Only the “middle” site would need to be running 5773. These scenarios were not explored as part of the Beta testing, but have been identified for future consideration.
10. **Positive: Cost (future).** Cascaded configuration is platform for the development of “diskless bunker” in the future. No official timeframes for development are available from EMC yet.
11. **Minor negative: Cost.** Cascaded SRDF needs an additional licence.
12. **Cascaded vs Concurrent:** When running in a Cascaded SRDF configuration, there are several scenarios that require one to reconfigure to a Concurrent configuration:

- I. Planned failover to Site C, with consistent replication back to Site A (presumed common)
- II. Loss of links between Site A and Site B
- III. Loss of links between Site B and Site C (assuming it cannot be routed at the network level via Site A)
- IV. Loss of Site B (Bunker)

Hence, if one adopts Cascaded SRDF, one will need to reconfigure to use Concurrent to work around the above issues and hence will need to run under Star control. There are fewer circumstances when one must change to a Cascaded SRDF configuration when running Concurrent SRDF, but they do exist:

- I. Loss of links between Site A and Site C (assuming it cannot be routed at the network level via Site B)

3.1.3 Recommendations

- “Native” Cascaded SRDF has a number of restrictions relating to dealing with failures and any Citi certification of cascaded SRDF configurations should be for use under Star control.
- There is a slight preference in Open Systems for an SRDF/Star configuration to be Concurrent rather than Cascaded for the following reasons:
 - Adding full support for Cascaded configurations to Citi’s existing certified Star solution that is based on Concurrent configurations will even further complicate the our SRDF/Star solution for Operations, with only modest benefit.
 - Planned switched to Site C, with consistent replication back to Site A (presumed common) require a concurrent SRDF configuration, or a reconfiguration from cascaded to concurrent. Hence, operating in Concurrent mode makes the normal configuration is “C → A ready”.
 - With the concurrent configuration, there are less failure scenarios where one would require reconfiguration
- The process for reconfiguring to cascaded configuration should be certified and added to the SRDF/Star Solutions Guide to deal with the failure scenario when A → C replication is not possible.
- If demand within Citi for 3-site solutions begins to outstrip the scalability of Star (limited by SDDF sessions and RDF daemon scalability), then certification of “native” Cascaded SRDF should be considered.

3.2 Moving dynamic RDF pair devices

3.2.1 Description

Previously, moving a dynamic RDF device pair from one RDF group to another required the deletion of the existing pair followed by the creation of the same pair in the new RDF group. While successful in moving the device pair, this process resulted in a full resynchronization of the data being performed.

Engenuity version 5773 allows this move to occur without the deletion of the existing dynamic RDF pair, thus avoiding a full resynchronization.

3.2.2 Key findings

- With SRDF/A, we usually have one RDF group per application. To add volumes to an application requires a full resync of the new volumes at the time they are added to the RDF group. This process causes the remote copy of the application data to appear not to be consistent throughout the entire period it takes for the new volumes to resynchronize. A point-in-time copy needs to be made prior to the resynchronization, which will age during the resync.
- The new movepair feature will allow us to pre-synchronize the volumes to be added, hugely reducing the time that the application is not being fully protected by SRDF/A.

3.2.3 Recommendations

- We will definitely want to certify the use of movepair for adding disk to SRDF/A groups and update Citi Best Practice documentation.

3.3 SRDF/A R2 timestamp representation

3.3.1 Description

When using SRDF/A, it is important to constantly monitor adherence to the recovery point objective (RPO). Previously, when querying an SRDF/A enabled RDF group, the “time that R2 is behind R1” for the R2 data was only displayed while the SRDF/A session was active. It was also only available when querying from a host connected to the production array.

With Engenuity version 5773, the microcode running on the Symmetrix DMX will maintain the time that the R2 is behind the R1 and return this information to Solutions Enabler SYMCLI commands. This value will then be available to the API whether the SRDF/A session is active or inactive. The information will be stored on both the R1 and R2 Symmetrix DMX and so will also be available when querying the SRDF/A devices from a host connected to the R2 Symmetrix DMX. SYMCLI will also calculate and display an actual timestamp, indicating the time at which the “apply” SRDF/A delta set transitioned from the “capture” delta set to the “transmit” delta set.

```
# symrdf -g RDFA query -rdfa
Device Group (DG) Name : RDFA
```

```

...
RDFA Session Status : Inactive
...
Time that R2 is behind R1 : 00:02:49
R2 Image Capture Time : Thu Dec 6 10:25:00 2007

```

3.3.2 Key findings

- Today it is not possible to determine after a failure the point-in-time that the application data represents. With 5773, we will be able to communicate to the business the capture time (to within ~1 second if NTP is used on the Solutions Enabler host) of the data on the R2 device.
- The value of this to any given application owner is uncertain, but it is an improvement on the situation today where we can give only a limited indication of the amount of data that might have been lost.

3.3.3 Recommendation

- We should add the use of this feature to Citi Best Practice documentation.

3.4 Minimum cycle time for SRDF/A

3.4.1 Description

In 5773, this is reduced from 5 seconds to 1 second (3 seconds is using Multi-Session Consistency).

3.4.2 Key findings

- Limited benefit to Citi. This is mostly a marketing benefit to show that SRDF/A can have an RPO of 1-2 seconds as HDS claim with HUR. In practice, there are disadvantages to setting a very short minimum cycle time for SRDF/A.
- You lose the write-folding benefits that result in 20-30% reduction in bandwidth requirements you get when the minimum cycle time is set to 30 sec.
- In testing, we found that the actual cycle time (not minimum time) extended to several seconds (depending on load) as it takes a finite amount of time to replicate the transmit delta to the remote site.

3.4.3 Recommendation

- No intention to recommend very low cycle times as Best Practice

4 Appendix B - High-level Comparision of SRDF 3-site Solutions

The following table compares the characteristics of each of four different 3-Site DR solutions now available from EMC, and based on Citi testing and evaluation.

- Items in **green** are positive or non-issue items.
- Items in **orange** are acceptable but not as good as the **green** items
- Items in **orange italics** are acceptable would require further Citi certification work
- Items in **bold orange** are considered weak (but not showstopping) features of the product.
- Items in **large bold red** are considered “showstoppers”.

	Concurrent		Cascaded (new)	
	“Native” Concurrent SRDF	Concurrent SRDF/Star	“Native” Cascaded SRDF	Cascaded SRDF/Star
Minimum microcode	5771	5771	5773 <i>(at Site B)</i>	5773 <i>(at Sites A & B)</i>
Certification status in Citi	Certified, but not strategic solution	Certified today	<i>Not Certified yet</i>	<i>Not Certified yet</i>
Requires use of RDF daemon	No	Yes	No	Yes
Scalabiity	High	<u>Low. Limited by RDF daemon and SDDF session scalability</u>	High	<u>Low. Limited by RDF daemon and SDDF session scalability</u>
Failover automation from EMC?	<u>None</u>	Yes, using symstar	<u>None</u>	Yes, using symstar
Speed of failover to Site C	<u>Slow</u>	Moderate with 5771/5772 but fast with 5773/SE6.5	Fast (5773/SE6.5)	Fast (5773/SE6.5)
Ability to provide Best Case (non-rolling disaster) zero-data loss at Site C in the event of the loss of Site A with minimal	<u>No</u>	Requires switch operation to Site C, keeping Site B data	Data automatically replicates to Site C. However, a	Requires switch operation to Site C, keeping Site B data.

intervention			symrdf failover would be needed to make it available to hosts.	Data does not automatically replicate from Site B to Site C as SRDF/A switching daemons run at Site A and would not be available. Data automatically replicates to Site C for <u>application</u> server failures at Site A.
Ability to deal with loss of Site B without compromising DR protection	Yes. Direct replication to Site C	Yes. Direct replication to Site C	No	Yes, via symstar reconfigure to a concurrent configuration.
Ease of planned switch to C with C → A replication (e.g. COB test)	<u>Requires Citi scripts</u>	OK	<u>Cannot provide a consistent image at Site A</u>	Yes, via symstar reconfigure to a concurrent configuration and symstar switch
Ease of unplanned switch to C with C → A (e.g. loss of server)	<u>Requires Citi scripts</u>	OK	<u>Cannot provide a consistent image at Site A</u>	Yes, via symstar reconfigure to a concurrent configuration and symstar switch
Ease of unplanned switch to C with C → B (e.g. Loss of Site A)	OK	OK	Requires Citi scripts	Yes via symstar switch
Ease of unplanned reconfiguration to A → C (e.g. Loss of Site B)	Not needed	Not needed	<u>Cannot be done</u>	Yes, via symstar reconfigure to a concurrent configuration
Platform for “dataless” bunker site?	No	No	Future development	Future development

5 Appendix C - Enginuity 5773 Virtual Provisioning

5.1 Introduction

EMC's Virtual Provisioning (EVP) uses thin devices presented to the host while consuming physical storage from a shared storage pool. EVP is a feature set introduced with Enginuity 5773 Microcode and Solution Enabler 6.5. EVP provides three distinct advantages over standard provisioning methods:

- **Improves storage capacity utilization** - applications are dynamically provided with sufficient storage, thereby allowing the storage allocation to grow with the application requirements.
- **Supports oversubscription** - the cumulatively advertised capacity exceeds the physical amount of storage available in the storage pool.
- **Increases throughput performance** – device-level wide striping of the data over all storage devices increases the throughput potential of the virtual devices.

EVP utilizes a storage pool to group virtual devices called Thin Devices (TDEV) with the actual storage volumes called Data Devices (DATADEV). The TDEV's function as a typical standard Symmetrix devices; however, the TDEV doesn't require any allocation of physical storage. The DATADEV's are created from physical storage that adheres to the same configuration requirements as standard Symmetrix devices (note: at the time of this evaluation EVP only supported mirror and Raid-6 protection types). The EVP storage pool manages the relationship between the TDEV's and DATADEV's, settings for oversubscription limits and monitors the allocation status.

5.2 I/O Functionality

EVP widely stripes the data across all available DATADEV within the storage pool, typically spanning a much greater number of drives than standard provisioning. This distribution of workload across the DATADEV's maintains a consistent level of performance by reducing the number and frequency of hot spots. Random read and write workloads will see the greatest benefit from EVP's device-level wide striping. Sequential workloads, which do not benefit greatly from wide striping, will see roughly comparable performance to standard devices, although performance will vary by workload and configuration.

In a write request, the Symmetrix allocates a minimum amount of physical storage from the thin pool. The pool's storage is allocated sequentially, regardless of whether data is written sequentially or randomly to the TDEV. This allocation is performed in small units of storage called "device extents" that are 12 tracks (768 KB) in size. A round-robin mechanism is used to balance the allocation of data device extents across all of the DATADEV's in the pool.

A read request retrieves the appropriate data from the DATADEV's utilizing pointers to locate the requested information in the storage pool. Reads request to an area that the thin device that has not been mapped will not trigger an allocation operation, resulting in a returned block where each byte is equal to zero.

5.3 Observations on Virtual Provisioning

- Target **no** higher than 60 to 80 percent capacity utilization per pool, leaving a buffer for unexpected growth or “runaway” application(s) that consumes more physical capacity than was originally planned. There should be sufficient free space in the storage pool greater in capacity than the largest unallocated thin device. A DATADEV maybe disabled and/or removed from a storage pool, requiring the allocated space to be “drained” to the remaining data devices in the storage pool before the device can be disabled or removed.
- Refrain from using file systems that do not reuse the space associated with deleted files, resulting in unused space.
- Avoid databases that pre-allocate space by writing zeros to the available capacity. This pre-allocated space cannot be shared in a thin pool, reducing or eliminating the capacity utilization benefits.
- EVP is a natural fit to ensure that premium Flash drive resources are highly utilized. Flash drives can offer both unparalleled response times as well as significant IOs/GB, and, unlike traditional drives, throughput and response time do not steadily decline as utilization of the drive increases.
- EVP is appropriate for applications for which some performance variability can be tolerated. Workloads with a high amount of random I/O will likely see performance improvements from wide striping.
- Document and software development/source code repositories, which can tolerate some level of performance variability, will benefit in the storage utilization efficiencies of thin provisioning.

5.4 Summary

EMC Virtual Provisioning (EVP) product introduced with Enginuity 5773 and Solution Enabler 6.5 provides the next logical step in array resource management. The core of EVP uses a storage pool to forge a relationship between Thin Devices (TDEV) and Data Devices (DATADEV), where TDEV's are presented to host applications like standard devices and DATADEV's that resemble standard devices are the actual storage volumes. The results of this lab evaluation indicate that EVP's core capabilities provide sufficient stability and performance. However, there are gaps in terms of monitoring, alerting and capacity planning that need to be addressed.

6 Appendix D – EMC Virtual Provisioning Summary Test Results

6.1 Throughput Performance

The fileserver/database simulations and the throughput performance test results are largely determined by the access times and the command queuing performance of the array, rather than bandwidth bottlenecks. As a means to closely simulated server applications, access patterns have been designed to step through several levels of queue depths and transaction sizes. Virtual devices uses device-level wide striping to distribute the data across all DATADEV's, to obtain a consistent high-level of through performance. According to EMC, virtual storage has a 10-15% performance advantage for OLTP type workloads over standard devices.

The performance tests compared an equivalent number of standard devices to DATADEV's, which will nullifies some of the performance potential of virtual devices. In actual deployments a much greater distribution of DATADEV's will be utilized to take advantage of EVP's device-level wide striping.

6.1.1 Positives

- Virtual devices showed highly consistent and predictable throughput at all queue depths tested.
- The virtual devices showed similar throughput performance as a transactions sizes approaches 32KB and larger.

6.1.2 Negatives

- The throughput test showed standard devices with a slight performance advantage over virtual devices at the small and medium transaction sizes (2-16KB).
- The fileserver/database simulation showed standard devices with a 7-20% throughput advantage over virtual devices.

6.2 Drain Performance

In the event that DATADEV(s) are removed from an existing storage pool, the DATADEV requires that the device be disabled and any allocated space on the device "drained" to the remaining DATADEV's and remapped from the TDEV's. Only on completion of the drain process can the DATADEV be removed from the storage pool.

6.2.1 Positives

- DATADEV's can be removed from the storage pool without any data discrepancies.

6.2.2 Negative

- During the draining process the storage pool showed a **98% reduction in throughput** as compared to normal operation. The draining process has an extreme negative affect on performance which points to the Symmetrix placing a high I/O priority on the draining process. A high priority on the draining process is understandable from the array perspective, since it impacts the integrity of the existing data within the thin pool; however, an equitable balance must be establish between the draining and typical I/O within the storage pool. Until EMC develops an equitable I/O sharing capability, disabling DATADEV's should be avoided.

6.3 Capacity Utilization

EMC bases the capacity alerting thresholds on the DATADEV percent full value; therefore, the allocated capacity within the storage pool must be measured accurately. EVP uses thin device extents of 12 tracks (786,432 bytes) each, which is the minimum amount of data that is allocated from the data pool when first host write occurs to a thin device. The unused portion of an extent maybe used for future data writes.

6.3.1 Positives

- EMC provides the capability to view both the DATADEV and TDEV utilization with a single CLI command.

6.3.2 Negatives

- The use of extents for the data writes makes it difficult to determine with any certainty the accuracy of the percent fill value. An initial host write will allocate a full extent of 786,432 bytes to the storage pool; therefore, in a worst case scenario (high number of random writes) the storage pool will depict a percent fill that is 11 times higher than actual. The probability of this capacity misrepresentation is greater with storage pools with a low amount of allocations, versus a highly allocated storage pool where much of the initial extents unused portion is written over with valid data.

6.4 Storage Pool Alerting

Of primary importance for thin provisioning is the ability to monitor the consumptions of pool storage space. EMC recommends that storage pools **not** surpass 80% of capacity allocation. A certain amount of free space is required to handle "runaway" host applications and/or draining of DATADEV(s). The Symmetrix places alert messages into the Symmetrix log at the following fixed consumption levels:

- 60% - issue a warning alert
- 65% - issue a warning alert
- 70% - issue a warning alert

- 80% - issue a critical alert
- 100% - issue a fatal alert

6.4.1 Positives

- The storage pool capacity alert messages are properly placed into the Symmetrix event log regarding capacity fill in the form of an informational, minor and major formatting messages.

6.4.2 Negatives

- Symmetrix event log requires the Symmetrix Event Daemon (symevntd) on the Solutions Enabler servers to be operational. Therefore, the collecting of alert messages into the Symmetrix log relies on the stability of the Solutions Enabler server and the "symevntd" being operational.

6.5 Storage Pool DATADEV Removal

EVP allows the storage pools capacity to be reduced by the removal of DATADEV(s) from the pool. To remove a DATADEV's from a storage pool it must first be disabled, which includes draining of any allocated space to the remaining DATADEV's in the storage pool. The draining procedure allows the TDEV's data to remain intact by changing the pointers to the new location in the remaining DATADEV's. Once drained, the DATADEV can be logically removed from the storage pool.

6.5.1 Positives

- EVP allows DATADEV(s) to be removed from the storage pool, in the event storage resources are misappropriated a need to be reallocated.

6.5.2 Negatives

- Removing DATADEV's from a storage pool has a detrimental affect on the distribution of drained data and newly written data to the remaining DATADEV's. The removal of a DATADEV seems to drain the majority of its data to closest neighbor in terms of logical device identifier. The unequal distribution of data within the pool may nullify EVP's performance attributes of the device-level wide striping. Therefore, DATADEV's should not be removed from the initial assigned pool, unless under serious resource constraints.

6.6 Storage Pool Data Security

The data written to the TDEV's are mapped into the DATADEV's and pointers are used to define the mapping locations. When a TDEV is unbound from the storage pool the mapping to the DATADEV's is also being removed rendering the information on the DATADEV's unintelligible.

6.6.1 Positives

- The unbinding process removes the pointer relationship between the TDEV's and DATADEV's, thereby dispersing or making the original data unintelligible. Dependant on the nature of the data, the unbinding of the TDEV's maybe sufficient to define the data as unintelligible, near equal to an erasure process.

6.6.2 Negatives

- If the unbinding process is not followed and the TDEV is re-used by masking to another host, the original data on written to the TDEV is compromised. Therefore, it is imperative that data on the TDEV be erased or be made unintelligible by the unbinding process.

6.7 Circumvention of Virtual Provisioning Objectives

The primary objective of EVP is to provide storage on a demand bases, thereby allowing an application to mature along with its storage needs. However, loopholes exist that allow users to circumvent EVP's primary objective by simulating storage demands, providing the illusion that the entire TDEV has been allotted.

6.7.1 Positives

The CLI provides the ability to view allocated status to an individual TDEV. This view allows micro-view of the storage pool to locate any TDEV that is inexplicably increasing allocation at a rapid rate.

6.7.2 Negatives

End-users may take advantage EVP's capability of providing storage on demand by using null or substitution bytes to reserve the entire TDEV storage capacity. This miss appropriation of virtual storage maybe accomplished by low-level formatting, "dd" (data definition) a disk or creating dummy substitution files to fill-in the file system. If the majority of the end-users utilize these types of reservation techniques, the storage pool would be quickly allocated.