



Global Distributed Storage Engineering
Citi Architecture & Technology Engineering

WAN throughput optimization for SRDF/A with Cisco MDS

GDSE-07026-SG

Solutions Guide

Version 1.0

VERSION CONTROL

Version and Date of Issue	Author	Comment
V1.0 13th August 2007	Chris Clayton	First draft, reviewed by Scott Torgerson and Colin Bradley

CONTENTS

Legal Notice	5
Contact	5
Latest Versions	5
Acknowledgements	5
1 INTRODUCTION	6
1.1 Cisco MDS 9000 FCIP	6
1.2 References	6
2 PROBLEMS TO BE ADDRESSED	7
2.1 Latency	7
2.1.1 Latency and SCSI write operations	8
2.2 Packet Drop	8
3 BEST PRACTICE SUMMARY	9
3.1 Best Practice Overview	9
3.2 Example of improvement via Best Practice	10
4 DETAIL OF BEST PRACTICE RECOMMENDATIONS	11
4.1 Use “clean” networks	11
4.2 Use Jumbo Frames, if supported on your network	12
4.3 Enable Write Acceleration	12
4.4 FCIP TCP Optimization.	13
4.5 Use Port Channeling	13
4.6 Enable Hardware Compression	14
4.7 Configure multiple RDF groups	14
4.8 Impact of block size on throughput	15
4.9 Scaling the FCIP infrastructure for SRDF/A	16

4.9.1	How many MPS cards do you need?	17
5	APPENDIX A - FURTHER DETAILS OF FCIP TCP OPTIMIZATION	19
5.1.1	Dynamic TCP Windowing and buffer space	19
5.1.2	Storage-optimized Congestion Recovery & Packet shaping	19
5.1.3	TCP Selective Acknowledgement (SACK)	20
6	APPENDIX B – GLOSSARY OF SELECTED TCP TERMINOLOGY	21
7	APPENDIX C – RELEVANT SRDF/A INFORMATION	22
7.1	Some key facts about SRDF/A	22

Legal Notice

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

CITI, CITIBANK, CITICORP, CITIGROUP and the Umbrella Device 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/>. Newer versions may include expanded scope, or comments on issues discovered after deployment.

Acknowledgements

The author would like to thank the following people for assistance with the contents of this document:

Scott Torgerson (Citigroup – Global Distributed Storage Engineering) for the Cisco MDS 9000 throughput testing work.

Bob Smith (EMC) for advice on configuring SRDF for long distance replication

1 Introduction

The Global Strategic Data Centre team has determined that Citigroup will be adopting an Out-of-Region (OOR) Continuation of Business (COB) and Disaster Recovery (DR) model going forward. As part of that initiative, there is a need for long-distance data replication.

Global Distributed Storage Engineering has already certified EMC's DMX-based SRDF/A Asynchronous Replication Solution for long distance replication over IP networks. However, SRDF/A suffers from network throughput reduction over long distances due to latency and packet drop. This equates to a reduction in utilization of expensive long-distance bandwidth.

The goal of this Solutions Guide is to provide Best Practice for minimizing SRDF/A network throughput reduction when using Cisco MDS 9000 FCIP (the current standard for new SRDF/A deployments). A sister document covering the legacy McData IPS product is listed in the References table below.

1.1 Cisco MDS 9000 FCIP

The Cisco MDS 9000 Multi-protocol Services Modules (MPS) extend the reach of Fibre Channel SANs by using Fibre Channel over IP (FCIP). FCIP is a mechanism that allows SAN islands to be interconnected over IP networks. The connection is transparent to Fibre Channel, and the result of an FCIP link between two fabrics is a single fully merged Fibre Channel fabric. TCP/IP is used as the underlying transport to provide congestion control and in-order delivery of error-free data.

The GDSE Cisco solution has the following documented relevant features:

- Gigabit Ethernet line-rate performance for FCIP
- Support for Jumbo frames
- Port Channeling functionality
- FCIP TCP Optimization
- FC Write Acceleration
- Hardware-based compression

The Cisco MDS 9000 FCIP solution is certified by GDSE within Citigroup and the MDS Product Page is <http://engineering.citigroup.net/gdse/products/platforms/ciscomds/>

1.2 References

Document Name	URL
Cisco MDS 9000 FCIP V1.0 Solutions Guide	http://engineering.citigroup.net/gdse/products/platforms/ciscomds/stdsdocs/07015-SG.pdf
Improving Network Performance for SRDF/A Position Paper	http://engineering.citigroup.net/gdse/products/platforms/ciscomds/stdsdocs/06034-pp.pdf
WAN Throughput optimization for SRDF/A with McData IPS	http://engineering.citigroup.net/gdse/products/platforms/symmetrix/stdsdocs/06034-sg.pdf

2 Problems to be addressed

2.1 Latency

Network latency is the time it takes a packet to travel from source to destination and back. In an SRDF configuration, the latency is the elapsed time between transmitting information from the source Symmetrix and receiving a response back from the remote Symmetrix. This is also known as the round trip time (RTT) and is measured in milliseconds (ms).

There are two main factors that contribute to the amount of latency in a network: propagation delay and processing delay. Propagation delay is the delay due to the distance the packet travels multiplied by the speed at which it is traveling. Processing delays are latency impacts due to passing through network devices such as routers. As a rule of thumb, latency increases by 1ms per 200km of circuit distance.

For long distance replication, latency can become a limiting factor on SRDF throughput rate, as the following test results shown in Figure 1 illustrate.

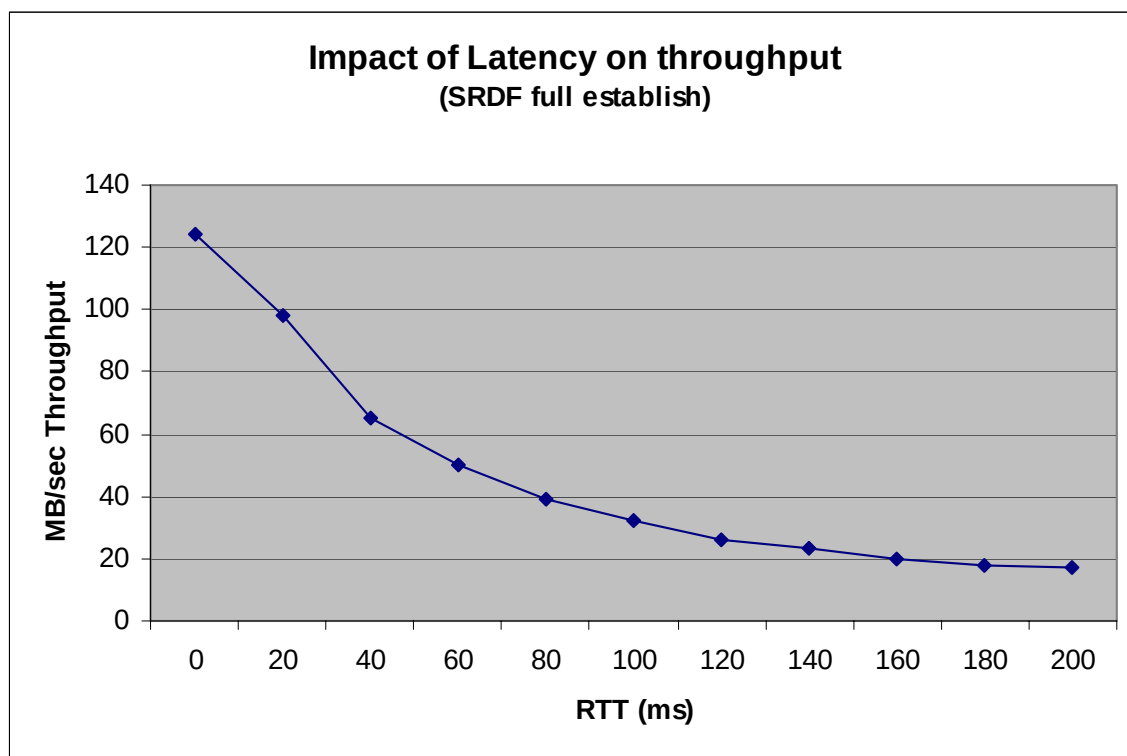


Figure 1 – Impact of latency on DMX3 full establish (Cisco MDS 9000, 1 RDF group with 30 LUNs, 0% packet loss & Jumbo frames)

2.1.1 Latency and SCSI write operations

SCSI was originally designed with the assumption that the SCSI client and SCSI server were in close proximity to each other. Where SCSI is used over Fibre Channel (FC), and FC over IP to transfer storage data over thousands of miles (latencies measured in many milliseconds), the performance of SCSI write operations can be significantly impacted. This is because SCSI write operations typically require more multiple handshakes between SCSI client and SCSI server (a write using SRDF/FC requires 2 round trips per write, although SRDF using native GigE only requires 1).

2.2 Packet Drop

Packet Drop refers to packets transmitted from source Symmetrix that do not arrive at the target Symmetrix, usually due to congestion. Packet drop is specified as a percentage (e.g. 1 packet dropped out of 10,000 is referred to as 0.01% packet drop. When packet loss occurs, packets must be retransmitted to the target Symmetrix. In addition, as part of the standard TCP/IP flow control behavior; the transmitting IP port will throttle down its rate of transfer temporarily to prevent further dropped packets, which results in reduced SRDF/A throughput.

Testing by EMC and Citigroup has found that whilst network latency can impact SRDF/A throughput, packet drop in conjunction with network latency can have a very profound impact on SRDF/A throughput and stability.

Note that Packet drop is also commonly referred to as packet loss. However, the latter phrase is being generally avoided in this document as in network parlance packet loss is a different effect relating to hardware issues, which generally no longer occurs in modern networks.

3 Best Practice Summary

The following Best Practice for overcoming the impact of latency and packet drop is based on testing in the GDSE lab and discussions with EMC and Cisco on the capabilities and behavior of both SRDF and the FCIP product. Each recommendation is discussed in more detail later in this document.

3.1 Best Practice Overview

1. **Use “clean” networks.** SRDF/A should be deployed over networks with low packet loss/drop characteristics. If possible, use dedicated bandwidth. See section 4.1 for further details.
2. **Use Jumbo frames, if supported on your network.** Use of Jumbo frames significantly increases throughput on the link. See section 4.2 for further details.
3. **Enable Write Acceleration** This will help minimize the impact of latency on SCSI write operations. See section 4.3 for further details.
4. **FCIP TCP Optimization.** Various TCP parameters and functions have been tuned for use with FCIP to be optimized for long distance data replication. Many are on by default. Some require certain parameters to be set appropriately. See section 4.4 for further details.
5. **Use Port Channeling.** Port channeling simplifies zoning across the WAN, eliminating the need to manually spread zones across the GigE connections for load balancing and performance. See section 4.5 for further details.
6. **Enable Hardware Compression*.** Compression does not address latency and packet drop issues directly, but is relevant to reducing bandwidth requirements and hence is highly important. See section 4.6 for further details.
7. **Configure multiple RDF groups.** This is typical practice for SRDF/A where each application will usually have its own RDF group. See section 4.7 for further details.

* At the time of writing, hardware compression is not yet certified because of a known Cisco bug, but it will be best practice at the next release if the bug is fixed.

3.2 Example of improvement via Best Practice

An example of the improvement possible with a combination of the above optimization steps are shown in the Figure 2. The example data from Citigroup testing is for a DMX3 full establish, and shows throughput as a function of latency.

The lower curve shows throughput for a poor configuration, whereas the higher curve shows the additional benefit of spreading the same load across multiple RDF groups, enabling Write Acceleration and enabling Jumbo frames.

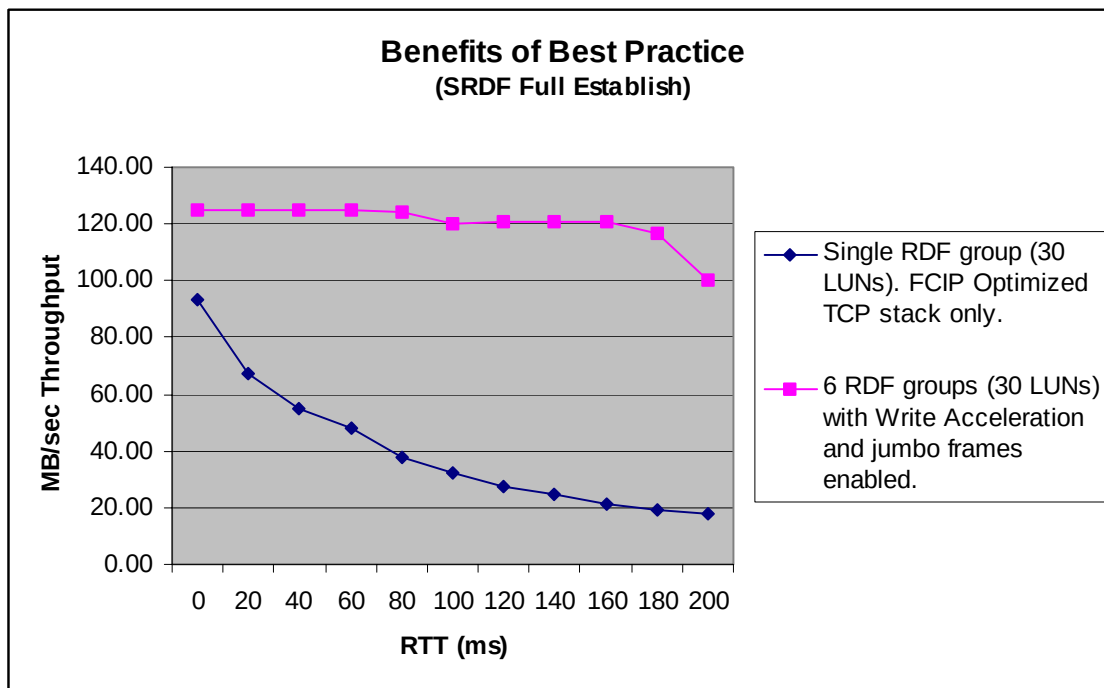


Figure 2 – Example of the throughput difference between partially optimized (FCIP Optimized TCP stack and 0% packet loss) and fully optimized SRDF configurations

4 Detail of Best Practice recommendations

4.1 Use “clean” networks

SRDF/A should be deployed over networks with low packet loss/drop characteristics to avoid the throughput degradation shown in Figure 2. . You must work closely with your network engineering when planning an SRDF/A implementation to ensure that the network component meets your requirements. This can be achieved in part by observing the following principles:

- Do not use public networks. Private circuits should be provided.
- Ensure with network engineering that there are no bit errors on the line (genuine hardware packet loss rather than packet drop caused by congestion)
- Telco should not be allowed to change the routing and hence latency and quality of the links.
- Use dedicated bandwidth for storage applications. **Use of dedicated networks give the potential to use Jumbo frames**, if the network can support it.
- Connect the storage applications into the network so that there is minimal congestion with other Citigroup traffic before it leaves site. For example, connecting to a server farm is not recommended as the replication traffic will need to compete with the server traffic before it even leaves site.

Figure 3 shows the benefit of using a network with no packet loss.

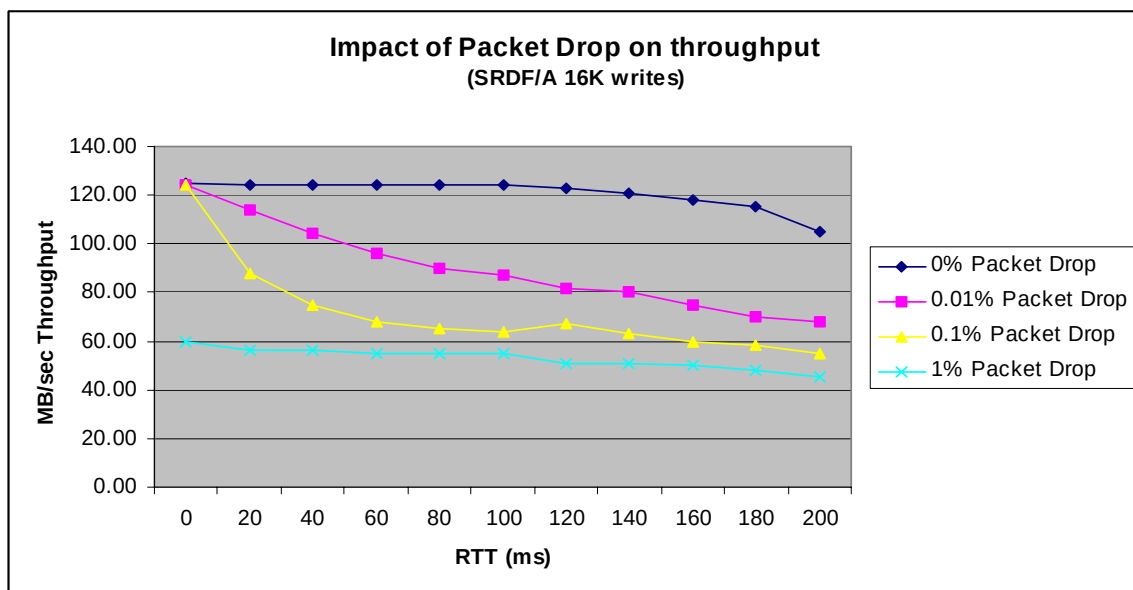


Figure 3 – Impact of packet drop on DMX3 SRDF/A (Cisco MDS, 3 RDF groups, 30 LUNs and Jumbo frames)

4.2 Use Jumbo Frames, if supported on your network

A maximum transmission unit (MTU) is the largest size packet or frame that can be sent in an IP network. TCP uses the MTU to determine the maximum size of each packet. Too small an MTU can result in greater header overhead and an increased number of acknowledgements. The maximum FC frame size is 2148 whereas the standard Ethernet MTU is 1518. As the FC frame approaches 1500 bytes, fragmentation of the frame occurs, leading to degraded throughput.

Using jumbo frames (MTU = 2300) the FCIP payload can accommodate a full-size FC frame and thereby avoid fragmenting the frame into two TCP packets.

To use jumbo frames the entire IP infrastructure between the local and remote MPS cards, and a dedicated network is required (with the modified MTU). Hence jumbo frames may not be supported in all locations.

Figure 4 shows the significant throughput benefit of using Jumbo frames.

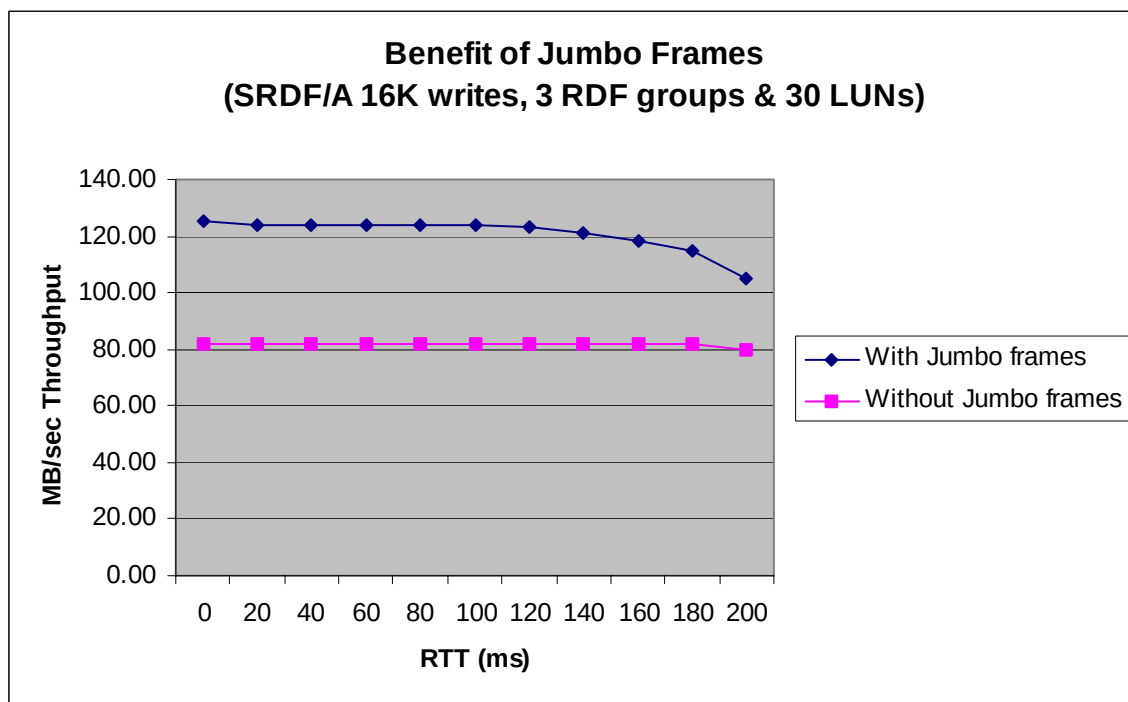


Figure 4 – Benefit of using Jumbo Frames

4.3 Enable Write Acceleration

Write Acceleration is a SCSI protocol-spoofing mechanism that improves replication throughput by reducing the overall service time for SCSI write I/Os and replicated write I/Os over distance. Write Acceleration reduces each write I/O to One Round Trip over WAN (from two or more) as the local FCIP end-point “proxies” SCSI XFER_RDY.

Write Acceleration is not enabled by default. See the Cisco MDS 9000 FCIP Solution Guide for details on how to enable Write Acceleration.

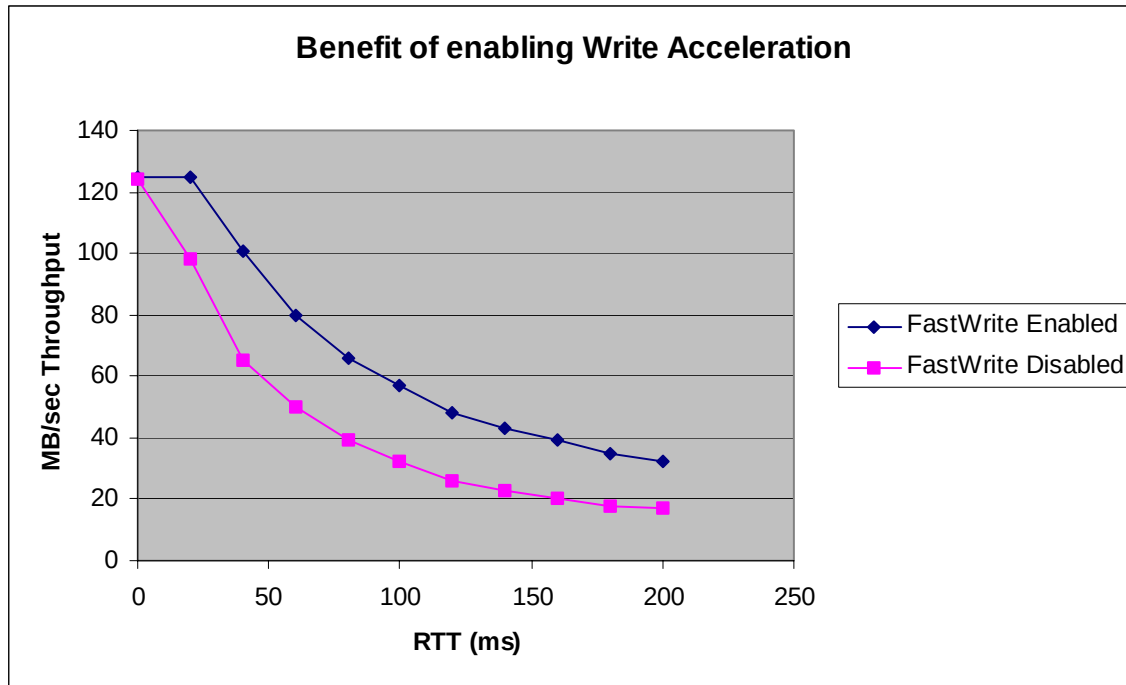


Figure 5 – Benefit of enabling Write Acceleration (SRDF Full establish with 1 RDF group, 30 LUNs and Jumbo frames)

4.4 FCIP TCP Optimization.

Storage traffic has different characteristics to most other traffic (large volume of write traffic, often bursty in nature) to which “Traditional” TCP is not well suited. “Traditional” TCP has a tendency to overreact to packet drops by halving the TCP window size, resulting in a reduction in throughput. MDS 9000 employs storage optimized TCP for FCIP which:

- Minimizes congestion with per flow packet shaping
- Responds more appropriately to congestion and drops
- Controls ramp up and max network usage, with tunable settings for Minimum and Maximum bandwidth
- Supports TCP writes of up to 32MB window per FCIP link to combat latency droop.

See Appendix A for further details of FCIP TCP Optimization. The Cisco MDS 9000 FCIP Solution Guide gives details on the parameter settings currently recommended for FCIP TCP optimization.

4.5 Use Port Channeling

The original SAN WAN extension product certified by GDSE (McData IPS SAN Routers) has no way to trunk the iFCP GigE cards, which means cross-site zones have to be *manually* spread across the iFCP GigE cards for optimal performance and load balancing. In contrast, the Cisco MDS FCIP Solution provides Port Channeling which allows implementers to bundle FCIP GigE

interfaces into one large pipe which removes the need to manually lay out cross-site zones for load balancing and performance.

Furthermore, with Port Channeling all RDF groups (possibly from multiple Symmetrix) will now be sharing the same set of GigE ports, and hence one no longer has to worry about having sufficient RDF groups per RA (as with the McData IPS) to ensure adequate outstanding I/Os and hence acceptable throughput.

See the Cisco MDS 9000 FCIP Solution Guide for details on how to configure Port Channeling.

4.6 Enable Hardware Compression

At the time of writing, hardware compression is not yet certified because of a known Cisco bug, but it will be best practice at the next release if the bug is fixed.

Check the Cisco MDS Solutions Guide to determine if your hardware supported compression. If supported, it is best practice to use this feature. Compression does not address latency and packet drop issues directly, but is relevant to reducing bandwidth requirements and hence is highly important. Hardware compression is preferred over software compression, where the compression engine can become the throughput bottleneck.

See the Cisco MDS 9000 FCIP Solution Guide for details on how to enable hardware compression.

4.7 Configure multiple RDF groups

All SRDF modes, including SRDF/A, have a limit of 32 IOs per hypers on the link. Hence, if you are only replicating a small number of hypers (< 16 per RDF group), it may be difficult to provide sufficient I/O to fill the network pipe. One method to maximize the number of outstanding I/Os to fill the network pipe is to increase the number of RDF groups (for the same set of LUNs).

Cisco testing has shown that the improvement is dramatic for SRDF/A, as the following figure shows the increase in throughput as a function of latency for a full establish by simply spreading the same set of devices between an increasing number of RDF groups.

With the McData IPS, it was necessary to ensure that there were sufficient RDF groups per RA as each RA was mapped to a specific iFCP port. With the Cisco solution, all replication is load-balanced across all GigE ports (each with its own FCIP tunnel) that are port channelled together. Hence, it is the total number of RDF groups on the port channel that matters, rather than per iFCP port (and hence RA) as with the McData solution.

Hence, with Cisco MDS, this is now a slightly academic problem as in general one is very unlikely to be running just one or two RDF groups in an Open Systems environment (where each application would usually have its own RDF group). However, the situation of very low numbers of RDF groups could arise for a small number of applications or in situations where multiple applications share the same RDF group.

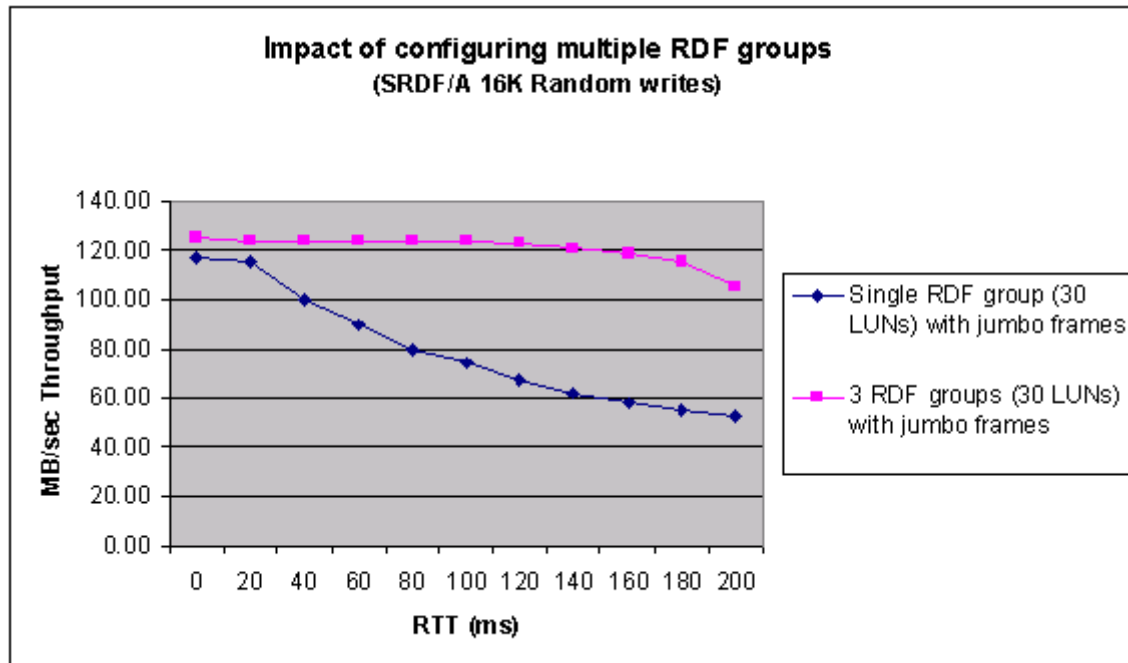


Figure 6 – Impact of configuring multiple RDF groups

4.8 Impact of block size on throughput

Previous testing of SRDF/A with McData IPS and DMX2 showed that the IOPS limit of a DMX RF can be a significant limiting factor when host I/O is composed of small random I/Os. The follow graph (Figure 7) with 0ms latency illustrates this point.

For the McData IPS data, a single RA on a DMX2 (track size 32K) had many RDF groups and hypers configured to use it. The maximum capacity of the network link to which it is connected was 90MB/s. The throughput of SRDF/A host write traffic varies as a function of the size of the host traffic. Throughput for smaller I/Os is limited by the peak IOPS limit on the RF (~3700 IOPS for DMX2).

SRDF/A random I/Os 32K or larger (not shown) also are degraded, as they will be broken up into 32K or smaller chunks. This is because I/Os are not aligned to cache slots / tracks, an occurrence also known as track crossing. The impact of track crossing only becomes negligible when the host I/O size reaches 128K. For sequential I/Os however, entire cache slots are full and hence all SRDF/A transfer I/Os are 32K (for DMX2). 3700 32K IOPS is equivalent to a throughput of > 100MB/s and hence the network link is filled, in this example.

For DMX3, the track size has increased to 64K and the peak IOPS limit on the RF has risen to nearly 5000. With these changes, the new measurements made during MDS FCIP testing (see Figure 7) show that the impact of small block sizes is now negligible, with the throughput rising to near maximum (120MB/sec with Jumbo frames) for writes as small as 8K.

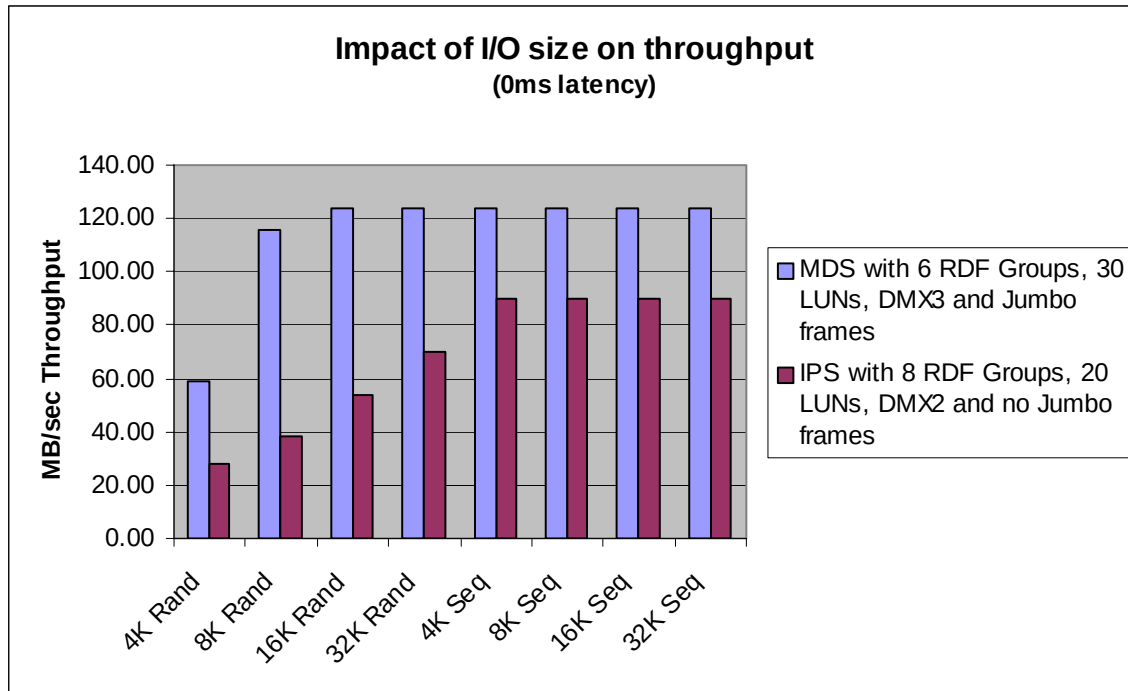


Figure 7 – Impact of SRDF/A I/O size on throughput

Hence, with the newer DMX3 and Cisco MDS 9000, there is no longer a need to optimize for small random I/Os, as was potentially necessary with the McData IPS and DMX2 combination.

4.9 Scaling the FCIP infrastructure for SRDF/A

With the McData IPS it was necessary to manually spread zones across the GigE connections for load balancing and performance. As a consequence, detailed Best Practice was developed to map RDF groups to RAs, and the RAs in turn to specific iFCP ports and shares. With the Cisco MDS Port Channeling functionality, one simply combines all the links for a given fabric (red or blue in the diagram below) to create a single virtual “fat” pipe (per fabric) that is shared between multiple DMX frames, and which provides automatic load balancing between the underlying individual GigE links from which it is created.

The diagram below taken from the MDS 9000 FCIP Solutions Guide illustrates this point. Two DMX frames are sharing the same Port Channeled pipes between sites. In each fabric, the GigE ports from one or two MPS 14+2 cards are combined into a single large pipe. There is no mapping between FCIP links and SAN ports; the FCIP tunnel establishes a virtual ISL which can carry traffic between sites. The main limit is the total line bandwidth.

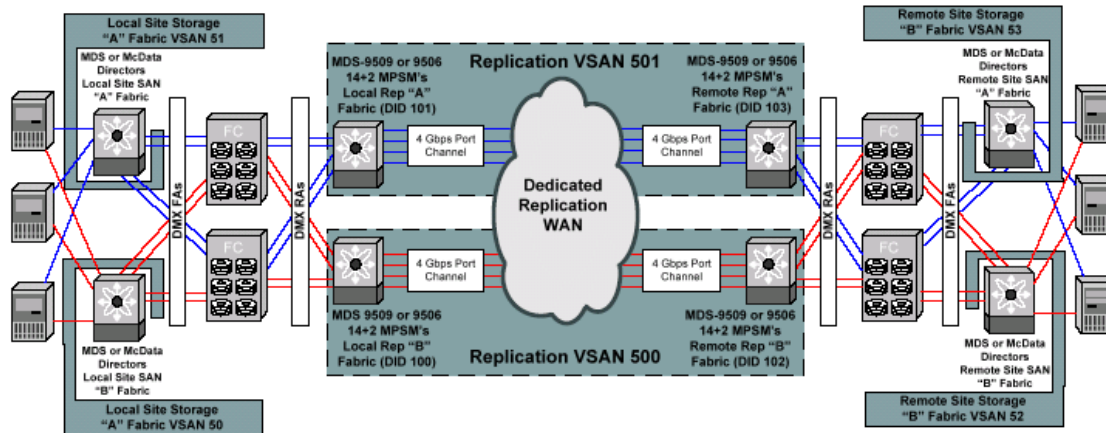


Figure 8 – Example Cross-Site Data Replication Topology

As additional MPS cards are added, the total bandwidth of the port channelled pipe can increase, providing the ability to add the RAs from additional DMX units to the replication fabrics, with self-load balancing across all connections for all the DMX units.

Note that with this configuration it is not possible to throttle the bandwidth available to a particular DMX. To achieve that, one would either have to:

1. use smaller port channel groupings, one per DMX, which would overall be less efficient as spare capacity on one port channel grouping would not be available to other DMX units. Hence, this is not a recommended configuration.
2. use QoS at the individual zone level (RA to RA). However, limiting is only achieved by preferring one DMX over another rather than setting a cap on the bandwidth available to a given array. Hence, there is not a recommended solution.

4.9.1 How many MPS cards do you need?

The factors that determine the number of MPS cards that are needed are:

1) Required throughput for SRDF/A workload (Mbps), (determined from SRDF/A Solutions Qualification analysis)

- Workload determines the minimum number of GigE ports needed to support that throughput. Workload also determines the size of the WAN link to support that throughput.
- Strictly this also determines the number of RAs, but as there are more RAs than GigE ports per card, the GigE ports will be the limiting factor from a throughput perspective.
- If GigE ports are throttled, then a given workload would require a greater number of GigE ports. For example, GigE ports may be throttled if additional MPS cards has been installed to provide redundancy against MPS card failure.

2) FC connectivity to support number of RDF groups required

Microcode	RDF group Limit per RA pair	RDF group Limit per DMX
5x71	8	64
5772	32	128

- For 5x71 code, 1 pair of RAs is needed for every 7 application RDF groups. Hence, with MPS 18+4 cards (once certified), and lower I/O applications, it is possible that the number of FC connections on the MPS cards could be exhausted before the GigE port bandwidth.

Example: A workload with a total of 126 application RDF groups, with an aggregate replication bandwidth of < 4Gbps would be limited by the number of FC ports on the MPS 18+4 cards (once certified), and not the number for GigE ports. In this case, it would be necessary to purchase additional MPS cards to provide the required FC connectivity.

- With 5772 code, the number of supported RDF groups per RA increases to 32 and hence, it will become very unlikely that the number of RDF groups and hence number of RAs/FC port required will dictate the number of MPS cards required.

Summary

- Details on determining the exact number of MPS cards required can be found in the Cisco MDS 9000 FCIP Solutions Guide
- In general, the number of MPS cards required will be dictated by the number of GigE ports needed to provide the bandwidth into the WAN, and the level of redundancy required to protect against card failure
- In exceptional cases, the number of MPS cards needed may be driven by the number of FC connections needed. The most likely case is when large numbers of RA connections are needed to support the number of RDF groups needed.

5 Appendix A - Further details of FCIP TCP Optimization

5.1.1 Dynamic TCP Windowing and buffer space

In TCP, each Maximum Window Size (MWS) worth of data must be acknowledged by the other end as having been received intact. Therefore, if the MWS is too small relative to the network latency, the FCIP connection will effectively be throttled due to the latency impact of waiting for TCP acknowledgments from the remote end.

Each MPS GigE port has a total of 180MB of memory for packet buffers (out of a total of 256MB). Each TCP connection can have a maximum window size of 32MB (minimum is 14KB). This value is automatically calculated from the product of the maximum bandwidth and RTT (both set in the FCIP profile). The RTT will dynamically adjust up to eight times the configured value in the FCIP profile. You can display the FCIP interface buffers with the command

```
show interface fcip #
```

The following fragment of output from the above command shows the current window size is ~22MB.

```
TCP Parameters
  Path MTU 2300 bytes
  Current retransmission timeout is 230 ms
  Round trip time: Smoothed 202 ms, Variance: 1 Jitter: 176 us
  Advertized window: Current: 22926 KB, Maximum: 22926 KB, Scale: 13
  Peer receive window: Current: 22920 KB, Maximum: 22952 KB, Scale: 13
  Congestion window: Current: 20656 KB, Slow start threshold: 23395 KB
  Current Send Buffer Size: 22926 KB, Requested Send Buffer Size: 0 KB
  CWM Burst Size: 50 KB
```

Cisco claims there is no limit on the number of outstanding I/Os that a GigE/FCIP tunnel can support (the McData IPS has a limit at the time of writing of 512 per TCP session).

5.1.2 Storage-optimized Congestion Recovery & Packet shaping

Traditional TCP has a tendency to overreact to packet drops by halving the TCP window size, leading to a reduction in speed that is not acceptable for storage replication. Packet shaping paces the flow of data to allow the TCP window to ramp up aggressively without overrunning downstream routers and causing congestion. This reduces the probability of drops and increases the resilience to drops if and when they occur. When you configure a GigE port, you specify the minimum available bandwidth. TCP can then use this as the slow start threshold. The packet shaper ramps up to this threshold almost immediately.

When congestion occurs, the FCIP TCP implementation is more aggressive during recovery than Traditional TCP. Hence, if Traditional TCP shares the same link with FCIP, the Traditional TCP flows recover more slowly. The bandwidth allocation then strongly favours the FCIP traffic. This is another reason for using dedicated bandwidth for storage replication.

5.1.3 TCP Selective Acknowledgement (SACK)

TCP SACK improves performance on lossy networks.

When multiple packets are lost from one TCP window of data, without SACK, a TCP sender could learn about only one lost packet per round trip time. With SACK, the receiving TCP host returns selective acknowledgement packets to the sender i.e. the receiver can acknowledge packets received out of order. The sender can then retransmit only the missing data segments, instead of everything since the first missing packet).

SACK is automatically set on with the Cisco MDS FCIP.

6 Appendix B – Glossary of selected TCP terminology

Retransmission Timeout (aka Retransmit Timer)

TCP uses a retransmit timer to ensure data delivery in the absence of any feedback from the remote data receiver. The duration of this timer is referred to as RTO (retransmission timeout).

Fast Retransmit & Fast Recovery

In TCP/IP, fast retransmit and recovery (FRR) is a congestion control algorithm that makes it possible to quickly recover lost data packets. Without FRR, the TCP uses a timer that requires a retransmission timeout if a packet is lost. No new or duplicate packets can be sent during the timeout period. With FRR, if a receiver receives a data segment that is out of order, it immediately sends a duplicate acknowledgement to the sender. If the sender receives three duplicate acknowledgements, it assumes that the data segment indicated by the acknowledgements is lost and immediately retransmits the lost segment. With FRR, time is not lost waiting for a timeout in order for retransmission to begin.

Congestion window

TCP's congestion window controls the number of packets a TCP flow may have in the network at any time.

Slow Start

Slow-start is part of the congestion control strategy used by TCP. Slow-start is used in conjunction with other algorithms to avoid sending more data than the network is capable of transmitting, that is, network congestion. Slow-start works by increasing the TCP congestion window size until acknowledgements are not received for some segments. TCP assumes this loss is due to network congestion, and takes steps to reduce the offered load on the network. A slow-start algorithm consists of two distinct phases: the exponential growth phase, and the linear growth phase. Although the strategy is referred to as "slow-start", its congestion window growth is quite aggressive.

Fast Recovery

There is a variation to the slow-start algorithm known as fast recovery. In the fast recovery algorithm, when packets are not received (detected through 3 duplicate ACKs) the congestion window size is reduced to the slow-start threshold, rather than the smaller initial value.

7 Appendix C – Relevant SRDF/A information

7.1 Some key facts about SRDF/A

1. All SRDF modes, including SRDF/A, have a limit of 32 I/Os per hypervisor on the link. Hence, if you are only replicating a small number of hypervisors (< 16 per RDF group), you may have trouble providing sufficient I/O to fill the network pipe. GDSE standards dictate that RAID1 LUNs be formed as Meta Devices (multiple hypervisors, typically 4 or more) and hence, this threshold of 16 hypervisors will quickly be exceeded in many cases.
2. SRDF/A host writes have no limit or queuing, and they write directly to cache.
3. SRDF/A transfers directly from cache to the remote Symmetrix and bursts at the link throughput limit. Therefore, there are no issues with Disk Adapter copy limits, which can sometimes limit other SRDF modes, such as adaptive copy.
4. The only SRDF/A limit is a maximum on 30,000 tracks/cycle on the link when establishing. This is not likely to be a practical limitation.
5. SRDF/A throughput depends on average write size. DMX2 can do about 3500 IOPS peak per RF. For DMX3, this increases to 5000 IOPS per RF.
6. For sequential SRDF/A host I/O, the I/Os for SRDF/A Transmit → Receive buffer copy will generally be the same as the DMX track size (32K for DMX2; 64K for DMX3), regardless of the size of the original host I/Os.
7. For random I/O, the I/Os will often cross track boundaries and in general will be split into two separate, smaller I/Os. This will result in a much lower SRDF/A Transmit → Receive buffer copy I/O rate compared with sequential host I/O.