Modeling and Analysis of Storage Area Network Extension Solutions

Radha Telikepalli¹, Tadeusz Drwiega² and James Yan³

Nortel Networks, 3500 Carling Avenue, Ottawa, CANADA, K2H 8E9
¹ radhat@nortel.com, ² drwiega@nortel.com, ³ jim.yan@sympatico.ca

Abstract: This article describes analytical models developed for analyzing the impact of distance, packet loss and transport protocols on the application throughput of Storage Area Network (SAN) extensions. The models have been used to compare the throughput and performance of asynchronous data replication of four SAN extension solutions offered over optical and IP-based transport networks. The models include the effects of TCP implementations on data replication performance.

Keywords: Storage Area Network, SAN extension, Analytical models

1.0 INTRODUCTION

1.1 Motivation

Traditionally, storage was based on a centralized architecture in the computing industry. With the introduction of client-server based distributed computing, development of SAN became essential. The initial focus of coverage area was limited to short distances such as campuses, where the effect of natural calamities (earthquake, flood, and fire), and man-made disruptions (power outage, physical/cyber attacks, terrorist attacks) would be severe making disaster recovery an important part of storage planning. In disaster recovery, a secondary site located away from the primary site is maintained to provide access with the same performance as the primary site. Synchronization between data stored in both sites is achieved through synchronous and asynchronous data replication.

Several protocols have been proposed in the recent past in support of SAN extensions over SONET/DWDM/Optical Ethernet using Fibre Channel (FC) [1], over Internet Protocol (IP) using Internet SCSI (iSCSI) [2], Internet Fibre Channel Protocol (iFCP) [3] and Fibre Channel over TCP/IP (FCIP) [4]. SCSI is the application layer protocol that is used in the storage end devices. Depending on the Open Systems Interface (OSI) layer at which these protocols were defined, SCSI commands will be either converted or encapsulated.

Many of the existing references [5-9] on this topic address measurement aspects of different solutions, but none have information on the effect of network variables – delay, packet loss etc. on SAN performance. The authors had an earlier publication [10] that discusses protocols available to extend storage area networks and gives the analysis of application throughput performance as a function of network variables. However, there was limited information about the models developed for the analysis. The purpose of this paper is to address the details behind the analytical models. The paper describes models developed to analyze the performance of asynchronous data replication, which usually consists of sending large volumes of data from one SAN island to another over potentially long distances.

1.2 Structure of the paper

The remainder of the paper is organized into two major sections with Section 2 describing the models developed for calculation of application throughput. Section 2 begins with short descriptions of protocols used in SAN and their extensions. The remainder of the section describes how those protocols and network characteristics have been modeled. In Section 3, the models are used to analyze the impact of protocol parameters and network characteristics on throughput and performance of several solutions to SAN extensions.
2.0 THE MODEL

A typical configuration that is used for SANs is shown in Figure 1, where two SAN islands are connected using an IP/SONET network with servers on the left. Storage devices situated in short distances from servers are known as local storage, in this case, the island that is closer to servers hosts local storage devices and the farther island hosts remote storage devices. Both local and remote devices store identical information. The servers use local storage for information retrieval under normal conditions and use remote storage in case of local site failures or disasters.

2.1 Protocol Description

2.1.1 SCSI

The SCSI protocol is used by data storage devices and servers to control the process of sending data from one device to another. SCSI is the top layer protocol in SAN extensions. SCSI command set is based on request and response model. A SCSI command is not said to be completed until a response is received. The response includes a status that indicates the final disposition of the command. The commands are exchanged between an initiator and a target. An initiator is a SCSI device that contains application clients which originate service requests and a target is a SCSI device that receives and executes commands received from an initiator. A command is communicated by sending a Command Descriptor Block (CDB). Some of the parameters of CDB are transfer length that specifies the data to be transferred, parameter list length that specifies the number of bytes sent from the data-out buffer and allocation length that specifies the number of bytes that an application client has allocated for returned data. SCSI commands can be transported as such, or over other transport protocols – Fibre Channel Protocol (FCP) [11], and TCP/IP. Commands and responses involved in a typical SCSI write operation is shown in Figure 2 (a).

2.1.2 Fibre Channel

Fibre Channel is an ANSI standard that has several layers 0 to 4 with FC-0 defining the physical layer and FC-4 defining mapping with upper layer protocols. The other layers can be mapped onto data link and network layers of OSI stack. SCSI commands are transported over Fibre Channel using FCP. FCP utilizes Information Units (IUs) that are mapped from SCSI commands, and uses link layer based flow control as defined for FC, where, data reception is guaranteed based on class of service chosen. Command sequence for an FCP based SCSI command write is as shown in Figure 2 (b), where the flow control is assumed to be that of Class 3, where acknowledgments are from the nearest node. SCSI command request is sent from one FCP port to the other FCP port as an FCP_CMND_IU that starts an I/O operation. Once the command is processed, the target sends FCP_XFER_RDY IU indicating that the target is ready to receive data and which portion of data it is ready to receive. At any time, the amount of data transmitted on the outgoing link depends on the number of Buffer credits or data buffer available in FC device. The initiator FCP_port transmits an FCP_Data IU, combination of XFER_RDY and Data IUs continue until the data transfer requested by the SCSI command is complete. Error recovery mechanism combines existing link layer error detection and controls with SCSI level methods and timers. Some of the error recovery methods are sequence level recovery in case of corrupted frames, command retransmissions with lost commands or acknowledgements, exchange level recovery involves...
aborting of the failed exchange, restart the exchange with reissue of the command. The methods are different for acknowledged and unacknowledged classes of service at the link layer.

2.1.2 iSCSI

iSCSI solution is for carrying SCSI commands encapsulated in IP data grams. This is the only solution that utilizes an end-to-end TCP/IP session and related flow control. Commands and responses of iSCSI are the same as that of SCSI and will be acknowledged and maintained by TCP layer.

2.1.3 FCIP and iFCP

One of the protocols available for IP-based storage extension is FCIP that connects existing Fibre Channel islands using LAN, MAN or WAN IP networks. FCIP entities/gateways are located at the edges of IP networks and are responsible for encapsulation and de-capsulation of FC frames. It uses two sessions – an end to end FC session and a TCP session between the gateways. FC based flow control is used in addition to TCP based control in IP network in terms of time outs and retransmissions. Command sequence for data write is as shown in Figure 3 (a). Since TCP is ignorant of the SCSI command sequence, gateways can be designed to send command and the data in the same data window thus reducing the data transfer latency by one round trip time.

iFCP is a gateway to gateway protocol and connects existing FC islands at wire speeds, across an IP network. Data delivery between the gateways is guaranteed by TCP/IP and by FC in FC islands. Since presence of IP network is not visible to the FC side of a gateway, dependency of buffer credits can be mitigated by using larger memory in the gateways. Also the gateways can be designed to handle both commands and data in the same window thus reducing data transfer latency as shown in Figure 3 (b).

2.2 Development of Model

2.2.1 Performance Metrics

Storage applications that typically utilize extension solutions are backup, recovery and replication. Data replication involves a local disk and a remote disk that carry the same information. Depending on the requirement whether the remote disk has to carry the same information at the same instant as the local disk or the remote disk represents a replica of the local disk within an allowable time window, the application can be planned. In the former type, the local server expects to receive an acknowledgement from the remote disk before proceeding to the next step, limiting the distance of separation between two SAN islands that contributes to latency. The process involved in the latter type, known as asynchronous replication, requires the data transfer to the remote disk to be completed within a specified time window, thus making data latency one of the performance metrics. Since asynchronous replication involves huge chunks of data within a specified period of time, typically a few hours, another required performance metric for SAN extension is application level throughput.
The following sub sections describe the mean-value model developed for calculation of application throughput for asynchronous replication.

End-to-end data latency calculations need to include path latencies, IO device write latencies and TCP processing latencies in end devices, if any. Application throughput is calculated as a ratio of total number of bytes transferred to end-to-end data latency between an initiator and target pair based on the assumption that the path is the same during data transfer.

2.2.2 FC-based extension

End-to-end latency of a data transfer across an FC-based SAN extension solution is given by (1)

$$D_{FC} = D_{RTT} + n_{trip} \times [ D_{RTT} + D_{IO} ]$$  \hspace{1cm} (1)

where,

- first $D_{RTT}$ accounts for the latency incurred by Command and response pair;
- $n_{trip}$ = number of trips needed across the network to transfer a particular chunk of data. It is determined from the available number of buffer credits or the amount of data that can be placed on a link at a time and the number of IO devices;
- $D_{IO}$ = delay experienced by the data in end device. E.g., in case of a disk, it is the sum of seek, rotation and IO transfer times;

$$D_{RTT} = 2 \times (D_{prop} + D_{NE})$$  \hspace{1cm} (2)

where, $D_{prop}$ = propagation delay;
- $D_{NE}$ = delay encountered in network elements

Packet loss in a SONET network that utilizes Generic Framing Procedure (GFP) for error detection and correction is negligible. In networks without GFP, packet loss may result in re-transmissions of frames, and thus increase the time taken by the transfer of the total data chunk. Equation (3) gives data transfer latencies in the presence of packet loss by including the additional time spent on re-transmissions of lost frames in (1).

$$D_{FC} = D_{RTT} + n_{trip} \times [ D_{RTT} + q \times D_{RTT} + D_{IO} ]$$  \hspace{1cm} (3)

where, $q$ = fraction of windows with lost packets when packet loss probability is $p$ and is given by (4).
\[ q = p \times \left[ \frac{\text{number of buffer credits} \times \text{size of each buffer}}{\text{frame size}} \right], \]

\[ \text{if average number of packets lost per transmission} < 1 \]

\[ = 1, \text{otherwise} \quad (4) \]

The average application throughput is the total number of transferred bytes divided by the end-to-end latency, \( D_{\text{FC}} \).

### 2.2.3 IP-based extension

IP-based extensions utilize a combination of FC and TCP/IP networks with flow control depending on the protocol used for the extension. The TCP model is based on the method given in [12]. Successful packets are acknowledged with an ACK message, the number of which is implementation dependent. Packet loss is detected with the reception of four ACKs or triple-duplicate ACKs, with the same sequence number at the sender. Without any loss of packets, the congestion window increases by \( \frac{1}{b} \) every round up to a maximum of \( W_{\text{max}} \), with a packet loss detected by triple-duplicate ACKs, the window is reduced by a factor of two. The resulting throughput is given by equation (5).

\[
\text{Throughput (Mb/s)} = \min \left\{ \frac{W_{\text{max}}}{D_{\text{TCP}}}, \frac{MSS \times \sqrt{\frac{3}{2b}p}}{D_{\text{TCP}}} \right\} \quad (5)
\]

where, \( W_{\text{max}} \) = Maximum TCP window;

\( MSS \) = Maximum Segment Length or Packet size;

\( b \) = Number of packets acknowledged by an ACK;

\( p \) = Packet loss probability in the IP network;

\( D_{\text{TCP}} \) = Round trip delay in IP network

Calculation of \( D_{\text{TCP}} \) is dependent on the type of protocol used for storage extension, due to differences in data processing in gateway elements, end devices and is as discussed in the following equations.

With iSCSI, \( D_{\text{TCP}} = 2 \times (D_{\text{prop}} + D_{\text{NE}}) + TCP_{\text{processing}} + D_{\text{IO}} \) \quad (6)

where, TCP\_processing delay includes TCP header and checksum processing delays that is experienced by the data in the end device.

With FCIP, \( D_{\text{TCP}} = 2 \times (D_{\text{prop}} + D_{\text{NE}}) + D_{\text{FC}} \) \quad (7)

With iFCP, \( D_{\text{TCP}} = 2 \times (D_{\text{prop}} + D_{\text{NE}}) \) \quad (8)

For an iFCP solution, resultant throughput from Equation (5) can be used to calculate end-to-end data transfer latency by including latencies encountered in FC islands and IO devices. TCP processing occurs in gateways in parallel with IO processing in end devices for FCIP and iFCP solutions and thus is not taken into account in Equations (7) and (8). Application layer throughput calculation for IP-based solutions is the same as in FC-based extension.

### 2.2.4 Advances in TCP implementations

Currently there are several references available on the topic of maximizing TCP throughput performance on high speed transmission links including IETF RFC 1323 [13], on increasing the maximum available TCP window based on the bandwidth*delay product.
Other proposals for achieving high TCP throughput are

- Usage of jumbo Ethernet frames that increases throughput up to six times;
- Parallel TCP sessions that can result in a proportionately larger throughput;
- Modified TCP response function that can remove the constraint of very low packet loss rates that are difficult to be achieved. This feature is not included in the models presented in this document.

The last one is also called High Speed TCP (HSTCP) [14]. Effect of increased maximum TCP window and use of jumbo frames on application throughput performance of asynchronous replication are described in the following subsections.

2.2.4.1 Increased Maximum TCP window

With increased TCP window or by increased buffer space in the sender and the receiver, amount of data that can be sent before receiving an acknowledgement from the far-end increases, thus filling bandwidth pipe more efficiently than in conventional TCP implementation. However, TCP with larger congestion window is more sensitive to packet loss, sustaining high TCP throughputs at low packet losses. Sometimes, this implementation may require unrealistically low packet loss rates to achieve the high per-connection throughput. An optimal TCP window is derived from the product of available bandwidth and one way delay as given in Equation (9) -

\[ \text{Optimal TCP Window} = \text{Available Bandwidth} \times \text{One way Delay} \]  

Calculation of one way delay in storage environment involves propagation delay, network element delays and IO processing delays that depend on availability of parallel writes in the target disk.

2.2.4.2 Increased Frame sizes

Depending on the underlying technology, frame sizes can be increased resulting in reduced number of data transfers. In case of iSCSI, if interconnecting Ethernet network supports, the frame size can be set at 9 kB instead of 1.5 kB. From equation (5) of TCP based throughput, achievable gain in throughput can be as much as 6 times. So is the case with the other IP-based extensions. However, the gain may be camouflaged as the TCP/IP network interconnects two FC islands and the application throughput performance depends on a combination of the variables mentioned in Section 3.1 for FC islands and IP network.

3.0 PERFORMANCE ANALYSIS

3.1 Performance affecting factors

- Distance between SAN islands: Throughput of a SAN extension depends on the time that a data window and its acknowledgement take to traverse from one island to the other. Propagation delay is the main component of that traversal time and it is directly proportional to separation distance between SAN islands;
- Packet loss: Packet loss is second major network impairment next to latency. It increases with network congestion and results in retransmissions. Packet loss in SONET is caused by BER, which is around 1e-12, and is 1e-15 with the GFP correcting ability. Present day IP networks are reported to have packet losses in range of 1e-04 to 1e-03;
- Frame sizes: With larger frame sizes, the number of frames to transfer a particular chunk of data will be smaller, hence resulting in reduced data transfer latency. Also under the same packet loss probability, fewer of them will be lost and hence reduced re-transmissions. Frame sizes depend on the underlying technology used at layer 1 to layer 3. E.g., max frame size in FC is 2148 B with a low overhead of 44 B in networks with GFP framing. Maximum frame size in IP-based networks is 1518 B where jumbo Ethernet frames are not supported, otherwise 9 kB with a overhead of 134 B;
- Extension bandwidth: Bandwidth determines the rate, with which the data can be sent;
- Data window size: It determines amount of data that can be transferred before receiving acknowledgement for reliable data delivery. It gets translated into the number of trips in Equations (1) and (3). In case of Fibre Channel, the data window is based on negotiated buffer credits. For IP-based protocols, the data window is based on maximum TCP window;
### Table 1 Protocol-specific variables used in modeling

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Fibre Channel based</th>
<th>iSCSI based</th>
<th>FCIP &amp; iFCP based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maximum number of buffer credits available: 32</td>
<td>Single TCP connection with a window of 64 kB</td>
<td>Single TCP connection with a window of 64 kB</td>
</tr>
<tr>
<td>Frame size (Bytes)</td>
<td>2104 payload + 44 OH (GFP)</td>
<td>8192 or 1404 payload + 114 OH</td>
<td>1384 payload + 134 OH</td>
</tr>
<tr>
<td>TCP processing delays</td>
<td>Not Applicable</td>
<td>TCP processing OH: protocol related (packet based) + checksum (byte based)</td>
<td>TCP processing occurs in gateways in parallel to the processing in end devices</td>
</tr>
<tr>
<td>IO devices in the target disk &amp; IO block</td>
<td>= 1, with SONET and standard TCP implementations; 64 kB</td>
<td>= 1, with SONET and standard TCP implementations; 64 kB</td>
<td>= 8, with advanced TCP implementations</td>
</tr>
</tbody>
</table>

Note: OH – Overhead

### 3.2 Performance

In this section, performance analysis with varying performance affecting factors listed in the previous section is discussed. Except in section 3.2.1 and 3.2.4.1, the distance between SAN islands is fixed at 1000 km and the other design parameters are as given in Tables 1 and 2. Equations (1)-(4) are used for FC-based extensions and Equations (5)-(8) are used for IP-based extensions.

#### 3.2.1 Distance between SAN islands

Application throughput performance with increasing distance between SAN islands (i.e., increasing $D_{prop}$ and $D_{NE}$) is shown in Figure 4, where the throughput decreases with distance due to a rise in data transfer latency. At distances above 200 km, IP solutions using FCIP and iFCP have higher throughputs than FC, as they involve one lesser round trip in data transfer latency. A solution can be extended over any distance where it has satisfactory application throughput. In the figure, if 5 MB/s is taken as the target throughput, FC-based solution can be extended up to 200 km, FCIP and IFCP solutions can be extended up to 500 km. Application throughput of iSCSI is below this requirement due to TCP processing involved in the end device. These distances are calculated based on the assumptions in Table 1. The reach of the extension for FC-based solution is much longer for numbers of credits higher than 32. For the IP-based solutions the reach can also be increased with larger

### Table 2 Network-specific variables used in modeling

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAN separation distance</td>
<td>0 to 5000 km; all the equipment collocated in the same office for 0 km</td>
</tr>
<tr>
<td>Available bandwidth</td>
<td>1 Gbps, unless specified otherwise</td>
</tr>
<tr>
<td>Packet loss</td>
<td>Point to point over optical with GFP: negligible &lt; 1e-06; IP routed network: 1e-06 to 1e-03; Typical range in commercial IP networks: 1e-04 to 1e-03</td>
</tr>
</tbody>
</table>
TCP-windows, but not as much because the TCP throughput becomes very sensitive to packet losses. This is discussed in the following subsections.

3.2.2 Packet loss

Effect of packet loss on application throughput is demonstrated in Figure 5. With changes in packet loss, number of re-transmissions required for a successful data transfer will be different. Equations (3) and (4) are used for FC-based extensions and Equation (5) is used for IP-based extensions. Effect of packet loss below 1e-04 is negligible on application throughputs of IP-based solutions. Application throughput decreases by approximately 10% in 1e-04 to 1e-03 packet loss range that can occur in present day IP networks. When the packet loss exceeds 1e-03 the throughput decreases dramatically. This means that performance of IP-based extensions is very sensitive to congestion in IP networks. Effect of packet loss on FC without GFP is negligible as data transmission was considered to be successful before FC timeouts.

3.2.3 Extension bandwidth

Purpose of this variable is to specify optimal bandwidth required for data transfer across an IP-based storage extension as bandwidth is one of the valuable resources. Application throughput varies with network
bandwidth for an iFCP based solution as shown in Figure 6. Effect of bandwidth is negligible on iFCP solution due to data sending and acknowledgement receiving process involved in any TCP process. Size of TCP window (64 kB, in this case) places an upper limit on the amount of data to be sent and the effect of reduced bandwidth is much smaller than the propagation delays involved.

3.2.4 Increased data window size

An increase in the data window allows the sender to transmit increased amount of data before expecting an acknowledgement from the far end. Depending on the data to be transferred, data can be transferred in parallel onto the disks in the target. The maximum number of parallel writes is assumed to be eight with each IO block size of 64 kB, based on the existing product information. Effect of increased data window size—with more number of buffer credits in an FC-based extension and larger maximum TCP window in an iFCP-based extension is discussed in this sub-section.

3.2.4.1 Buffer credits and FC-based extension

In an FC-based extension, data window size can be increased with availability of larger buffer space or buffer credits in an FC end device that result in reduced number of trips in Equation (1). Variation of application throughput with the number of buffer credits is shown in Figure 7, where an extension solution can offer higher throughputs over large distances with increasing buffer credits. For e.g., for a target throughput of 10 MB/s, an FC-based SAN extension can be offered to cover 200 km with 60 buffer credits, 850 km with 125 buffer credits and approximately 2000 km with 255 buffer credits. Application throughput is below 10 MB/s for a system that has only 25 buffer credits.

3.2.4.2 Increased maximum TCP window

If a large memory space is available in iFCP gateways, application throughput performance will be determined based on the efficiency at which available parallel writes are being utilized. With a data window smaller than 512 kB, all available parallel writes can not be used efficiently. For larger windows, the data transfer will be faster due to reduced disk access latencies. For this analysis, four different sizes were considered for the maximum TCP windows as shown in Figure 8. From equation (9), the optimum window was calculated to be 1224 kB at a SAN separation distance of 1000 km. However, due to IO processing delays involved, the optimum window was found to be at 6650 kB with application throughputs of the order of 56 MB/s for packet losses below 1e-08 that are hard to be achieved in IP networks. As the window size increases, effect of packet loss becomes significant. For realistic packet losses above 1e-04, the application throughputs are almost the same irrespective of the maximum TCP window.

![Figure 8 Application throughput of iFCP with increased maximum TCP window](image1)

![Figure 9 Application throughput of iSCSI with different frame sizes](image2)
3.2.5 Frame size

Effect of frame size on application throughput is two fold – magnitude and variation of throughput with packet loss. Theoretically, the throughput increases by approximately 6 times with a frame size changing from 1.384 kB to 8.192 kB as given by equation (5). Application throughput performance of iSCSI based solution with different frame sizes is shown in Figure 9, where the gain in the throughput is limited to 1.3 due to TCP processing in the end device. Also the throughput is constant at 2 MB/s for packet losses below 1e-02 for larger frame sizes.

CONCLUSIONS

Analytical mean-value models were formulated to assess the application throughput performance of storage area networks and their extension solutions. From these models, it was found that network variables (delay, packet loss, and bandwidth) and implementation technology dependent variables (data window sizes and frame sizes) can affect the application throughput performance. The analysis of the application throughput was presented as a function of packet loss with the different variables mentioned above as design parameters.

REFERENCES

1) ANSI INCITS standards, Sub committee: T11, //www.t11.org/.