İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

FIBER CHANNEL VS. INTERNET SCSI ON STORAGE AREA NETWORKS FOR DISASTER RECOVERY OPERATIONS

M.Sc. Thesis by

Hakan ARIBAŞ, B.Sc.

(504971641)

Date of submission : 29 September 2006

Date of defence examination: 3 November 2006

Supervisor (Chairman): Assist. Prof. Dr. Osman Kaan EROL

Members of the Examining Committee Assoc. Prof. Dr. Feza BUZLUCA

Assoc. Prof. Dr. Ece Olcay GÜNEŞ

NOVEMBER 2006

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

FIBRE CHANNEL VE INTERNET SCSI TEKNOLOJİLERİNİN VERİ DEPOLAMA AĞININ FELAKET KURTARIMI İŞLEMLERİNDE KULLANILMASI

YÜKSEK LİSANS TEZİ

Müh. Hakan ARIBAŞ
(504971641)

Tezin Enstitüye Verildiği Tarih : 29 Eylül 2006
Tezin Savunulduğu Tarih : 3 Kasım 2006

Tez Danışmanı :	**Yrd. Doç. Dr. Osman Kaan EROL**
Diğer Jüri Üyeleri	**Doç. Dr. Feza BUZLUCA**
	Doç. Dr. Ece Olcay GÜNEŞ

KASIM 2006

PREFACE

I wish to thank my supervisor Assist.Prof.Dr. Osman Kaan EROL for all his help and guidance during the preparation of this thesis. I would also like to thank to all my friends for their valuable support and for their friendship. My wife, my son and my parents deserve special thanks for their support, motivation, and patience.

November, 2006 **Hakan ARIBAŞ**

TABLE OF CONTENTS

ABBREVIATIONS

ACP	: Association of Contingency Planners
AHS	: Additional Header Segment
APA	: HP Auto-Port Aggregation
BCM	: Business Continuity Management
BCP	: Business Continuity Plan
BHS	: Basic Header Segment
BIA	: Business Impact Analysis
CA	: HP StorageWorks Continuous Access EVA Software
CDB	: Command Descriptor Block
CEO	: Chief Executive Officer
CNT	: Computer Network Technology
CRA	: Central Registry Agency Inc. of Turkey
CRC	: Cyclic Redundant Check
DAS	: Directly Attached Storage
DNS	: Domain Name System
DRP	: Disaster Recovery Plan
EOF	: End-Of-Frame
EVA	: HP StorageWorks Enterprise Virtual Array
FC	: Fibre Channel
FCIP	: Fibre Channel over Internet Protocol
FCP	: Fibre Channel Protocol
FFIEC	: Federal Financial Institutions Examination Council
Gbps	: Giga-bit per second
HBA	: Host Bus Adapter
I/O	: Input Output
IFTF	: Internet Engineering Task Force
IP	: Internet Protocol
iFCP	: Internet Fibre Channel Protocol
iSCSI	: Internet SCSI
LAN	: Local Area Network
LUN	: Logical Unit Number
MKK	: Merkezi Kayıt Kuruluşu
MTU	: Maximum Transmission Unit
NIC	: Network Interface Card
OSHA	: Occupational Safety and Health Administration
OUI	: Organizationally Unique Identifier
PDU	: Protocol Data Unit
PLOGI	: Fibre Channel Port Login
PRM	: Process Resource Manager
QoS	: Quality of Service
R2T	: Ready to Transfer
RAID	: Redundant Array of Independent Disks
RPO	: Recovery Point Objective

RTO	: Recovery Time Objective
SAN	: Storage Area Network
SCSI	: Small Computer System Interface
SOF	: Start-Of-Frame
SRDF	: Symmetrix Replicated Data Facility
TCML	: Turkish Capital Market Law
TCO	: Total Cost of Ownership
TCP	: Transmission Control Protocol
WWNs	: World Wide Names

LIST OF TABLES

LIST OF FIGURES

FIBRE CHANNEL VE INTERNET SCSI TEKNOLOJİLERİNİN VERİ DEPOLAMA AĞININ FELAKET KURTARIMI İŞLEMLERİNDE KULLANILMASI

ÖZET

Kuruluşların bir felaketle karşılaştıkları zaman hayatlarını devam ettirebilmeleri için bir iş sürekliliği planına sahip olmaları önemli bir kriterdir. Felaket kurtarımı ise iş süreklilik planlarının teknoloji bacağı olduğu için aynı şekilde önemlidir. Felaket kurtarımının temeli ise verinin korunmasına dayanmaktadır. Verinin korunabilmesi için ise uzak bir noktaya kopyalanması gereklidir. Günümüzde en yaygın olarak kullanılan veri depolama teknolojisi Veri Depolama Ağı olduğundan, veri depolama ağının uzak bir noktaya kopyalanması ve yedeklenmesi felaket kurtarımının en önemli adımlarıdır. Veri depolama ağlarında yaygın olarak kullanılan protokol Fiber Kanal protokolüdür. Fiber Kanal protokolünün pek çok avantajı olmasına rağmen yüksek maliyet ve bunu yönetecek özel yetiştirilmiş uzman gereksinimi gibi dezavantajları vardır. Ayrıca Fiber Kanal teoride 10km, pratikte ise sadece 300m mesafede etkin olarak kullanılabilmektedir. Bu mesafe sınırı felaket kurtarımı için tavsiye edilen mesafelerin oldukça altındadır. Internet SCSI (iSCSI) protokolü aynen Fiber Kanal gibi veri ağlarına yönelik bir protokoldür ancak Fiber Kanal gibi özel ve pahalı donanımlar gerektirmek yerine mevcut Ethernet ağını ve cihazlarını kullanarak kolayca oluşturulabilmektedir. Fiber Kanal veri taşıma mekanizması olarak kendine özel mekanizmaları kullanıyorken, iSCSI TCP/IP ağını kullanmaktadır.

iSCSI tabanlı veri depolama sistemleri taşıma mekanizmasındaki bu önemli farklılık nedeniyle klasik Fiber Kanal tabanlı veri depolama sistemlere göre farklı performans karakteristikleri gösterebilmektedir.

Bu tez çalışmasında iSCSI tabanlı veri depolama ağlarının performansının iyileştirilmesi için iSCSI ve TCP katmalarının birbiriyle etkileşimi incelenmektedir. Bu inceleme neticesinde en uygun iSCSI ve TCP parametre değerleri belirlenmeye çalışılmıştır. Uygun parametre değerleri kullanılarak optimize edilmiş bir iSCSI veri depolama çözümünün Fiber Kanal tabanlı veri depolama çözümlerine alternatif olabileceği gösterilmeye çalışılmıştır.

FIBER CHANNEL VS. INTERNET SCSI ON STORAGE AREA NETWORKS FOR DISASTER RECOVERY OPERATIONS

SUMMARY

Business continuity planning prepares an organization to recover after some interruption of its business operations. Disaster recovery planning is the technological aspect of business continuity planning. Data protection is a key component of disaster recovery planning. Data replication is the copying of data from one system and its disks to another system and its completely independent and redundant set of disks, and the key component of data protection. Storage area network (SAN) that allows a dedicated network to provide access to centralized storage resources are widely used nowadays. Therefore, remote data replication and remote data backup of storage are network are very important subjects. While Fibre Channel (FC) work well in providing high speed storage area networks, there are some drawbacks associated with this technology such as high total cost of ownership, limited operating distance. Fibre Channel requires a separate fibre-optic network for the SAN, while iSCSI uses the existing TCP/IP network.

Due to the significant shift in its transport mechanism, iSCSI-based storage systems may possess different performance characteristics from the traditional FC-based storage systems.

This thesis examines the interactions between the iSCSI and TCP layer in order to improve the performance of iSCSI-based storage system. As a result of this study, the most proper iSCSI and TCP parameter values were supposed to be determined. By using these proper parameter values, it was tried to be shown that an optimized iSCSI-based storage solution with suitable parameters can be an alternative to FC-based storage solutions.

1. INTRODUCTION

The last decade has been a time of enormous increases in both the amount and importance of information stored on computers and distributed across networks. Internet access has accelerated the demand for e-commerce services, media and educational applications, and critical operations involving health care, finance, transportation and law enforcement. As the amount of data has grown exponentially, its importance has also increased significantly. In these environments a massive volume of data must be moved for processing and analysis and then safely stored so that it can be accessed if required. Data protection is therefore dependent on seamless disaster recovery in the event that a primary storage location is damaged or destroyed in a man-made or natural disaster.

There are some drawbacks associated with FC-based SANs such as high total cost of ownership, limited operating distance, and step learning curve. The technology that has evolved to solve the problems of the historical FC-based SANs is iSCSI-SANs that are based on Internet Protocol (IP) technology. Although, the FC-based SAN is required a significant investment in new FC hardware, many days of installation, and weeks of training, iSCSI-based SANs rely on the use of industry standard Ethernet components - components widely used in business LANs, and familiar to every system manager. In many businesses no additional investment in the underlying network infrastructure is needed. In the cases where the network is upgraded or a switch is added, the hardware is one-third to one-quarter the cost of FC equivalents, and no additional training is required. Network Interface Cards for servers are familiar commodity parts.

iSCSI is an emerging standard for encapsulating storage I/O in TCP/IP. Due to the ubiquity and maturity of TCP/IP networks, iSCSI has gained a lot of momentum since its inception. On the other hand, the iSCSI-based storage is quite different from a traditional one. A traditional storage system is often physically restricted to a limited environment, e.g. in a data center. It also adopts a transport protocol specially tailored to this environment, e.g. parallel SCSI bus, Fibre Channel, etc. These characteristics make the storage system tend to be more robust, and achieve more predictable performance. It is much easier to estimate the performance and potential bottleneck by observing the workload. While in an iSCSI storage, the transport is no longer restricted to a small area. The initiator and the target can be far apart. The networking technology in between can be diverse and heterogeneous, e.g. ATM, Optical DWDM, Ethernet, Wireless, satellite, etc. The network condition can be congested and dynamically changing. Packets may experience long delay or even loss and retransmission, etc. Thus, the situation facing the iSCSI storage is quite different from the traditional one.

To take advantage of iSCSI protocol to build iSCSI storage systems, it is needed to better understand the iSCSI characteristics, such as the performance characteristics in various networking situations, the relationship between iSCSI and underlying TCP/IP protocol, etc.

The sensitivity of iSCSI to the underlying network is examined in this thesis. It is tried to answer how the existent IP infrastructure that is originally built for networking applications can be adapted for carrying storage and to what extent characteristics of the network such as link delay, MTU size will effect the end-to-end performance of a storage application. iSCSI and TCP combined have a large space in which they can be configured including parameters such as TCP window size which sets a limit on the achievable TCP performance, and the iSCSI maximum burst size which limits the maximum amount of data transferred in any single SCSI command. This thesis also investigates how these parameters and the other TCP and iSCSI parameters relate to one another in order to successfully deploy the SCSI protocol over TCP/IP. The mail goal of this thesis is to examine the interactions between the iSCSI and TCP layer in order to improve the performance of iSCSI-based storage system, and to show that an optimized iSCSI-based storage solution with suitable parameters can be an alternative to FC-based storage solutions.

The next chapter gives an overview of business continuity, disaster recovery planning and related objectives such as Recovery time objective (RTO) and recovery point objective (RPO) for a better understanding of their importance.

Chapter 3 investigates remote data replication which can be carried out several ways such as synchronous, asynchronous, and periodic.

Chapter 4 examines storage area networks (SANs) with different technologies, e.g. Fiber Channel (FC), Internet SCSI (iSCSI), and Fibre Channel over Internet Protocol (FCIP), Internet Fibre Channel Protocol (iFCP). iSCSI and FC are compared in view of disaster recovery operations.

Chapter 5 examines the performance characteristics of iSCSI for different iSCSI and TCP parameters to build high performance iSCSI-based SAN. The results of the effect of iSCSI parameters in iSCSI layer and TCP parameters in TCP layer to iSCSI data access performance are analyzed. It is also examined both how the CPU usage affects the iSCSI performance and how PDU burst size affects the CPU utilization.

Chapter 6 provides three samples of real world solutions for remote data replication and remote data backup. While first of them is a cost effective remote backup solution, two of them are disaster recovery solutions.

Chapter 7 is the last chapter that concludes and summarizes the work in this thesis.

2. BUSINESS CONTINUITY AND DISASTER RECOVERY

2.1 Business Continuity Planning: Expect the unexpected

Business continuity planning can be described as:

“The Process of developing advance arrangements and procedures that enable an organization to respond to an event in such a manner that critical business functions continue with planned levels of interruption or essential change.” [1]

Business continuity management can best be defined as:

“A holistic management process that identifies potential impacts that threaten an organization and provides a framework for building resilience with the capability for an effective response that safeguards the interests of its key stakeholders, reputation, brand and value-creating activities.” [2]

Building-in business continuity, making it part of the way that you run your business, rather than having to ‘firefight’ any emergency, helps prepare you to offer ‘business as usual’ in the quickest possible time. Planned business continuity management, so that your staff, customers and suppliers are reassured that you have an effective policy and practice for managing the unexpected, helps build confidence in your business.

Business continuity or business resumption planning prepares an organization to recover after some interruption of its business operations. A business continuity plan should comprehensively address all issues, from losing the physical location and key personnel to restoring the ability of the computer system to process information.

Business continuity and disaster recovery can be confused. Disaster recovery planning is the technological aspect of business continuity planning. The advance planning and preparations those are necessary to minimize loss and ensure continuity of the critical business functions of an organization in the event of disaster.

According to market researcher Gartner, Inc., eighty-five percent of large organizations have some sort of disaster recovery plan, but only twenty-five to thirty percent of small organizations have anticipated the costs of a disruption of their business [3]. The Occupational Safety and Health Administration (OSHA) requires that all firms with more than ten employees have a written disaster plan [4].

2.2 Business Continuity Planning Process

Without business continuity, a natural or man-made disaster could result in:

- Loss of work to competitors
- Failures within your supply chain
- Loss of reputation
- Human Resources issues
- Health and Safety liabilities
- Higher insurance premiums.

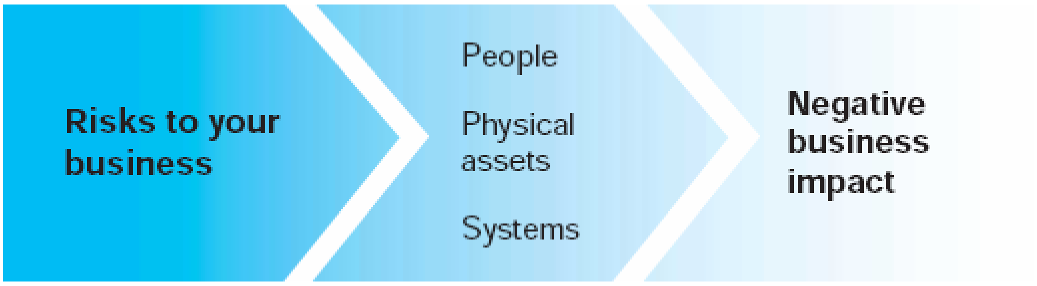


Figure 2.1: Risks of business

And, as every organization knows, when setbacks arrive in combination, the worst case scenario can eventually be business failure.

Financial institutions should conduct business continuity planning on an enterprise-wide basis. In enterprise-wide business continuity planning an institution considers every critical aspect of its business in creating a plan for how it will respond to disruptions. It is not limited to the restoration of information technology systems and services, or data maintained in electronic form, since such actions, by themselves, cannot always put an institution back in business. Without a BCP that considers every critical business unit, including personnel, physical workspace, and similar issues, an institution may not be able to resume serving its customers at acceptable levels. Institutions that outsource the majority of their data processing, core processing, or other information technology systems or services are still expected to implement an appropriate BCP addressing the equipment and processes that remain under their control.

Financial institutions should also recognize their role in supporting systemic financial market business processes (e.g., inter-bank payment systems, and key market clearance and settlement activities) and that service disruptions at their institution may significantly affect the integrity of key financial markets.

In the United States, The FFIEC agencies expect financial institutions that play a major role in critical financial markets to have robust planning and coordinated testing with other industry participants. Critical markets include, but may not be limited to, the markets for federal funds; foreign exchange; commercial paper; and government, corporate, and mortgage-backed securities. The best example of theses firms for Turkey is Merkezi Kayıt Kuruluşu that is a private company established in 2001 as a legal entity under the provision of the Turkish Capital Market Law (TCML). The registration of securities and the rights related to them are being registered electronically in a book entry form with respect to issuers, intermediary institutions and owner of rights by MKK.

The FFIEC agencies encourage financial institutions to adopt a process-oriented approach to business continuity planning that involves:

- Business impact analysis (BIA)
- Risk assessment
- Risk management
- Risk monitoring.

This framework is usable regardless of the size of the institution. Business continuity planning should focus on all critical business functions that need to be recovered to Personnel responsible for this phase should consider developing uniform interview and inventory questions that can be used on an enterprise-wide basis.

Key steps in developing business continuity management can be summarized as:

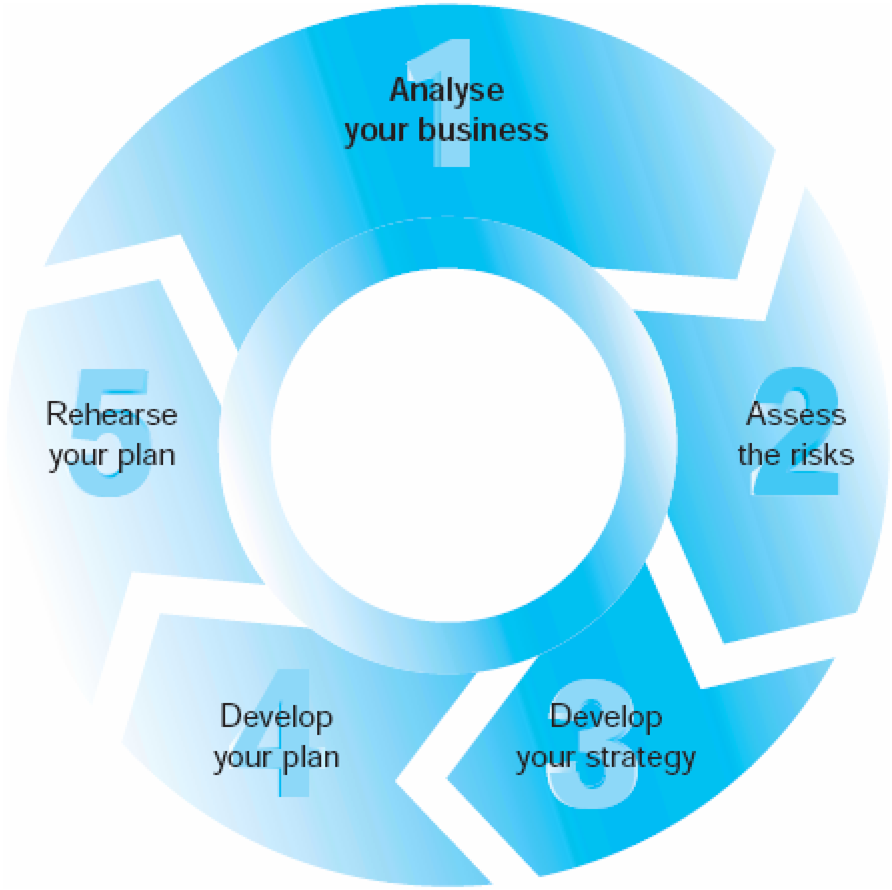


Figure 2.2: Key steps in developing business continuity management

Analyze different approaches to business continuity to decide which strategies would best apply to your needs [5]. An organization would want to review the following types of alternate facility strategies: mirror sites, hot sites, cold sites, warm sites, reciprocal arrangements with other organizations and alternate manual processes.

Data synchronization can become a challenge when dealing with an active/back-up environment. The larger and more complex an institution is (i.e., shorter acceptable operational outage period, greater volume of data, greater distance between primary and back-up location), the more difficult synchronization can become. If back-up copies are produced as of the close of a business day and a disruption occurs relatively late the next business day, all the transactions that took place after the back-up copies were made would have to be recreated, perhaps manually, in order to synchronize the recovery site with the primary site.

Management and testing of contingency arrangements are critical to ensure the recovery environment is synchronized with the primary work environment. This testing includes ensuring software versions are current, interfaces exist and are tested, and communication equipment is compatible. If the two locations, underlying systems, and interdependent business units are not synchronized, there is the likely possibility that recovery at the back-up location could encounter significant problems. Proper change control, information back up, and adequate testing can help avoid this situation. In addition, management should ensure the back-up facility has adequate capacity to process transactions in a timely manner in the event of a disruption at the primary location.

Many organizations duplicate or “mirror” their systems at remote sites. This allows them to immediately “fail over” to the alternate site when the primary site encounters a problem. This can also be used for load balancing in normal processing. For example in Merkezi Kayıt Kuruluşu (MKK), there is a mirror storage box in the same building but different floor. These two storage boxes work synchronous replication, replication direction is from primary box to mirror box. There is two 2Gbps FC connection between these boxes.

A “Hot” site has the equipment and resources necessary to restore business functionality after a disaster. These sites can be owned by the organization or reserved with an outside company in case of emergency. For example in MKK, there is hot site in the same city, Istanbul.

A “Cold” site is a facility that does not have the necessary hardware. An organization would need to ensure that they could quickly find the equipment they needed. Some vendors offer mobile cold sites, which are facilities in a trailer that can be moved to the site of a disaster to restore computing capabilities. For example in MKK, there is a cold site in the different city, Ankara.

An organization can also plan to replace the computer system with manual processing. However, the processes must be well documented [6].

2.3 Disaster Recovery Planning: The technological aspect of business continuity planning

Traditionally, disaster recovery planning has focused on the restoration of computer processing ability after a system failure. Information Technology divisions often draft disaster recovery plans. Therefore, disaster recovery planning can center on restoration of data and processing power and overlook broader issues such as the cost of downtime and replacement of the physical plant. The disaster may be a force of nature (flood or fire), or the work of man (computer theft or security violations), or a hardware or software failure that leaves the current computer system useless or the data corrupt. The disaster recovery plan must cover restoration of software and potential replacement of hardware and the physical location of hardware [6].

Part of the risk process is to review the types of disruptive events that can affect the normal running of the organization. There are many potential disruptive events and the impact and probability level must be assessed to give a sound basis for progress. To assist with this process the following list of potential events has been produced:

- **Environmental Disasters**: Hurricane, flood, snowstorm, earthquake, electrical, storms, fire, freezing conditions, contamination and environmental hazards, epidemic.

- **Organized and / or Deliberate Disruption**: Act of terrorism, act of sabotage, act of war, theft, arson, labor disputes / industrial action.

- **Loss of utilities and services**: Electrical power failure, loss of gas supply, loss of water supply, petroleum and oil shortage, communications services breakdown, loss of drainage / waste removal.

- **Equipment or System Failure**: Internal power failure, air conditioning failure, production line failure, cooling plant failure, equipment failure (excluding IT hardware).

- **Serious Information Security Incidents**: Cyber crime, loss of records or data, disclosure of sensitive information, IT system failure.

- **Other Emergency Situations**: Workplace violence, public transportation disruption, health and safety regulations, employee morale, mergers and acquisitions, negative publicity, legal problems

2.4 Recovery Time & Recovery Point Objectives (RTO & RPO)

Recovery Point Objective (RPO) is the point in time that the restarted infrastructure will begin to become evident. Basically, RPO is the rollback that will occur as a result of the recovery. To reduce a RPO it is necessary to increase the synchronicity of data replication.

The RPO indicates how up-to-date the recovered data must be in relation to when the disaster occurred. The RPO will tend to decrease as data becomes more time critical. For example, an organization may keep updated personnel files on all employees. These files are rarely updated and may need to be duplicated once a month. If a disaster strikes, the recovery point may be structured such that the recovered data is no more than a month old. This recovery point would not be sufficient for financial data that may be updated thousands of times an hour. In this case, the RPO would likely be set such that the recovered data is only a few minutes old so that critical financial information would not be lost. In some cases, the RPO may be so low that any loss of data is intolerable and thus the backup copies would have to be up-to-the-second duplicates of the original data at the primary site.

Recovery Time Objective (RTO) is the time that will pass before an infrastructure is available. In order to reduce the RTO, it is required for data to be online and available at another site.

The solution to safeguarding mission critical data is making sure that a replica copy exists at a location other than the primary site. This aims to ensure that if the primary site is damaged or destroyed the data will still be available once functionality is restored. In the past, a company would generally make weekly or nightly tape backups of important information and physically ship the tapes to a remote location where they were stored. Unfortunately, as data has grown in volume and importance, two factors continue to make shipping tape off-site less effective. The first is the decreasing RTO.

The RTO is a measure of how fast a company can recover from a disaster. The metric is somewhat ambiguous since it does not specify what constitutes a disaster. For example, the recovery time for a failed disk drive may be a matter of minutes or even seconds. But a site-wide event such as a fire or natural disaster may take many days to recover from. The recovery time objective has decreased to a matter of minutes or hours for some applications regardless of the scope of the disaster. In the most mission critical environments, the RTO may even be a matter of seconds for site-wide disasters. For instance, financial institutions may be able to recover from a site-wide disruption in just a few hours using remote data centers and high-speed networks. However, military and defense applications may require that recovery time be only a few seconds if a primary site is damaged or restored.

This can be further illustrated by categorizing RTOs and RPOs into different classes. These classes may vary from one organization to another but the general principle still applies. Computer Network Technology (CNT) defines these classes as follows

[7]. Class 1 is the lowest level, where acceptable recovery times range from 72 hours to 1 week and the most up-to-date data can be from a weekly backup (up to a week old). A Class 4 recovery environment contains the most stringent requirements. Here, the recovery time must be immediate and the data recovered must be less than one second old. The following table illustrates these different classes. Any high performance, high assurance system will require a class 3 or class 4 recovery categorization.

Table 2.1: Classification of RTO and RPO

	Class 1	**Class 2**	**Class 3**	**Class 4**
RTO	72 hours – 1 week	8 hours – 72 hours	Less than 8 hours	0 minutes
RPO	Last full backup – less than a week	Last backup – less than 24 hours	Less than 15 minutes before event	0 minutes

Class 4 environments require the most expensive solutions so as to keep up with the high demand of traffic and the high degree of availability required. These demands have traditionally required continuous availability (even in the event of a disaster) as well as fault tolerant hardware to guard against as many failures as possible. Furthermore, the RPOs in these environments tend to be so high that any loss of data is intolerable. Possible solutions for different classes are outlined in the following table.

Table 2.2: Classes of Recovery

Classes of Recovery			
Class 1	**Class 2**	**Class 3**	**Class 4**
Off-site Tape Warehousing Cold Server Recovery	Remote Tape Vaulting Warm Server Recovery Application replication	Storage controller data Replication Storage virtualization data replication	Continuous Availability w/ real-time replication Fault Tolerant Hardware

A disaster recovery solution must be defined by making a trade-off among implementation costs, maintenance costs, and the financial impact of a disaster, resulting from performing a business impact analysis of your business. In the following figure [8], cost of outage versus cost of solution can be seen.

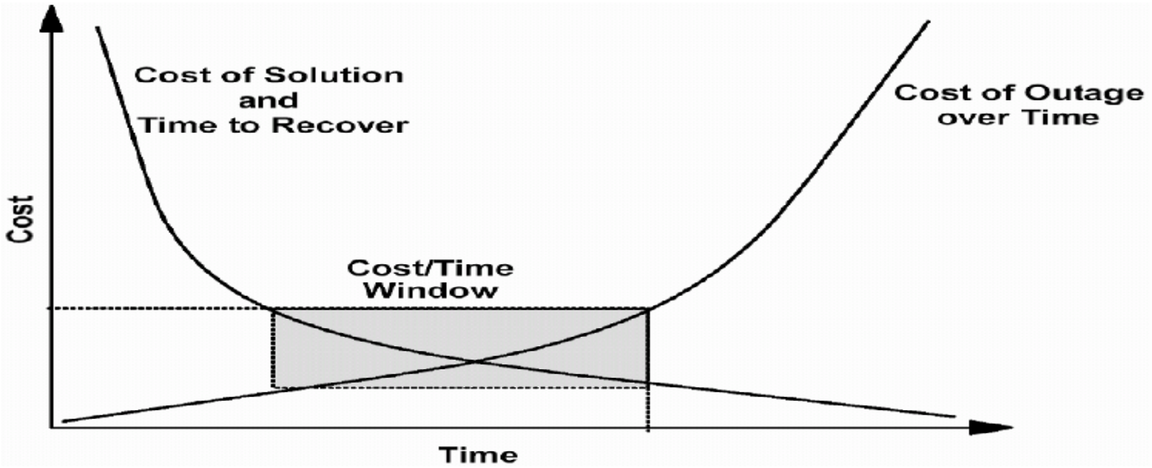


Figure 2.3: Cost of outage versus cost of solution

It will be shown that iSCSI is appropriate for all of the recovery classes listed above except for Class 4. For Class 4, a FC solution will be more convenient way.

2.5 Increasing Assurance using Remote Backup

In 2002, the US Securities and Exchange Commission issued a draft report in which it was suggested that remote backup sites be located between 200 and 300 miles from the primary facility [9]. It was suggested that increasing the distance between a primary site and backup site would greatly enhance the assurance characteristics of the system. The primary reason for this increase is due to the fact that further separation of primary and backup sites reduces the probability that both sites will be adversely affected by the same disaster. In 2003, the U.S. Chapter of the Association of Contingency Planners (ACP) participated in a survey conducted by The Disaster Recovery Journal to determine the best distance separation between primary and backup sites. According to the ACP, the average minimum distance was determined to be approximately 105 miles. This is shown in the following figure [10].

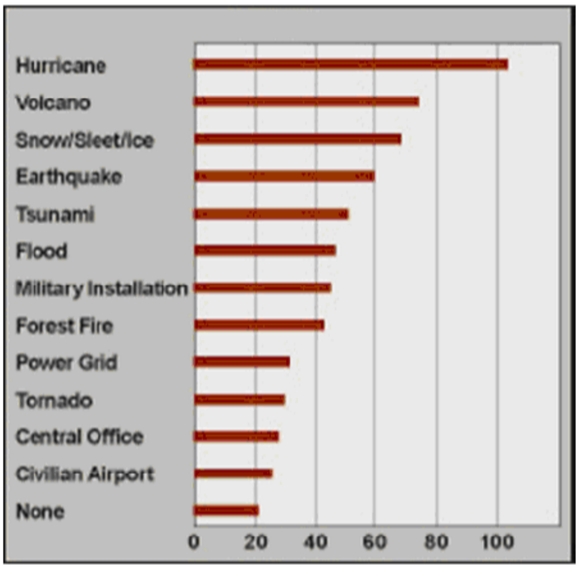


Figure 2.4: Distance recommendation

It was clear from the survey that this distance recommendation was made in large part to protect against a large natural disaster such as a hurricane. This provides a baseline number so as to understand how high-assurance networks should be formed. There is a performance and assurance tradeoff depending on the separation distance of the primary and backup sites. As the distance increases, the survivability of the system increases at the cost of performance (since latency increasingly degrades performance). On the other hand, as the distance is decreased, the chance of both sites becoming unavailable due to the same event increases, thus decreasing assurance.

3. REMOTE DATA REPLICATION

3.1 Data Protection: A key component of disaster recovery

Data protection is a key component of disaster recovery planning. Measures to classify and safeguard information assets (including records, microfilm, and electronic and optical data) need to be formalized and articulated.

3.2 Data replication

Data replication can be defined in the following way [11].

Replication is the copying of data from one system and its disks to another system and its completely independent and redundant set of disks. Replication is not the same as disk mirroring, because mirroring treats both sets of disks as a single, logical volume with enhanced availability, while replication treats the disk sets as two completely independent items. Mirroring is con- fined to a single computer system, while replication moves data from one system to another. The end result is two consistent and equally viable data sets ideally in distinct physical locations.

Most popular use of replication is for disaster recovery. Data replication can be classified in the following manner:

- synchronous
- asynchronous
- periodic

3.2.1. Synchronous

Synchronous replication can be seen as the only real time replication and provides the best data integrity. As data is written on the local node, the remote node receives the same data from the local node and has to acknowledge at least the receipt of the data to the local node. This acknowledgment is often called network acknowledgment. Often the local node also requires that the remote node acknowledges the successful write of the data, not just the successfully receipt. This acknowledgment is often called data acknowledgment. Which type of acknowledgment is used for synchronous replication depends on the solution used. Once the acknowledgment is received the local node reports to the application that the write has completed. Waiting for the acknowledgment causes a delay. This delay is called latency. The latency is caused by various components between local and the remote nodes. Synchronous replication is often used for solutions where local and remote nodes are positioned close together, because then the performance impact will be still acceptable. Synchronous replication ensures that the local node can not generate its next write until the current write has been acknowledged. Some critical applications will require that the data is always completely up-to date and therefore synchronous replication will be the only possible choice. Another advantage of synchronous replication is that data is always written to the remote node in the same order that it is written on the local node. For other types of replication (e.g. semi-synchronous, periodic) this is not always the case. Maintaining write ordering is important to guarantee that the data is always in a consistent state on both nodes during replication. If this is not maintained, data can be unusable on the remote node and therefore this defeats the entire purpose of implementing a replication solution.

3.2.2. Asynchronous

When using asynchronous replication, data is queued on the local node and sent to the remote node when network bandwidth or system performance permits. Writes are acknowledged to the application as soon as they are queued on the local node. As a result the impact on the application's performance will be minimal. The disadvantage of asynchronous replication is that data might not be completely up-to-date on the remote-node, however it is always in a consistent state. The logic of asynchronous replication is, that a failure resulting in a switchover to the remote node is an extremely rare event, and therefore only in case of a real disaster where the local node is completely destroyed, it is acceptable that some writes may be lost.

3.2.3. Periodic

Periodic replication is also often called "batch-style" replication. When using periodic replication data is saved up, and from time to time replicated in a batch, at once. The problem is, that write ordering normally is not maintained, meaning that data that has reached the remote node is not consistent during the replication. The problem deriving is, that the data on the remote node is not usable if the local node fails during replication. An advantage is that it can be selected quite flexible when the additional load on the network introduced by replication should occur. A disadvantage is that data can be very much out-of-date on the remote node, depending on the replication interval. When using periodic replication for disaster recovery, it is critical to maintain additional copies of the data. On the remote node some form of mirroring or snapshotting should be used. A mirror should be split off or a snapshot should be taken on the remote node before replicating to the local node. When the replication has finished the mirror can be reestablished or the snapshot can be removed. This maintains a consistent copy of the data during the replication process. Often it will also be highly desirable to split off a mirror or take a snapshot on the local node before replicating. This ensures that the data that is replicated is in a consistent state and does not change during replication.

3.3 Hardware-based and Application-based data replications

Replication is initiated by the disk system itself in the hardware-based replication. The disk system is also often called storage box or disk subsystem. Hardware-based solutions replicate to another disk system that has to be from the same vendor normally. The most suitable network type for this kind of replication is storage area network (SAN). For instance, HP StorageWorks Continuous Access EVA is used as a hardware initiated replication in Merkezi Kayıt Kuruluşu. HP StorageWorks Continuous Access is an array based application to create, manage and configure remote replication between two storage boxes that is connected each other via fibre channel protocol (FCP).

The most commonly used application-based replication probably is database replication. There are mainly two ways for database replication:

- log replay
- database replication managers

The simplest way to implement database replication is the use of database logs. A database normally is able to log all changes into files. This log files can then be copied to the remote node and in case of a failure they can be reapplied to the database on the remote node. This process is normally done manually. One disadvantage of log replay is that some transactions can be lost, since log files are normally transferred in a periodic way.

A database replication manager automates the process of copying the log files to the remote node and reapplies them automatically. Most modern databases offer replication managers. However most databases only support master/slave configurations, where writes are only accepted on the master node. For instance, Oracle Data Guard with Redo Apply option that used for physical standby databases is used as database replication managers in Merkezi Kayıt Kuruluşu.

4. STORAGE AREA NETWORKS (SAN)

4.1 Limits of directly-attached storage

In the great majority of computing environments, servers are at the center of the information system. They store data on their disks and access them to satisfy the users' requests that arrive on the LAN. Each server is connected to its disks by means of fixed, dedicated channels, such as the Small Computer System Interface (SCSI) parallel bus. Storage resources connected to a server are exclusively accessed and managed by that server. This paradigm is called "Directly-attached Storage" (DAS). As servers grow in number and request additional capacity, several different problems arise. The most important are summarized below.

- Scalability
- Performance
- Distance limitations
- Availability
- Data protection
- Efficiency

4.2 Requirements for next-generation storage subsystems

The list of limitations of DAS can be used as a starting point to quantitatively state the requirements of next-generation storage subsystems. Such a system must have the following features:

- offer almost unlimited scalability, allowing the interconnection of thousands of devices
- provide high, dedicated bandwidth

- allow large movement of data between storage devices (for example, for backup or replication purposes) without involving neither the servers nor the LAN

- allow the reconfiguration of the system and almost any other maintenance operation without downtime offer advanced and centralized management capabilities

While not built to fulfill only disaster recovery obligations, storage area networks have been introduced to allow a dedicated network to provide access to centralized storage resources. In traditional configurations, the SAN is a separate network and isolated from the LAN. This SAN allows for storage expansion without impacting the server resources on the LAN. Similarly, any server upgrades or expansions will not impact the storage resources on the SAN. Since the SAN is isolated from the LAN, it can allocate full bandwidth to storage needs while the LAN only has to be responsible for the communication between servers and other users. In other words, storage access will not impact the performance of the LAN and heavy LAN traffic will not degrade the performance of storage access. It is this characteristic that makes storage area networks extremely valuable for backup and recovery of mission critical data. In the past, tape backups had to be carefully planned so that the LAN was not congested with typical production traffic. As a result, tape backups had to be executed during off-peak hours – generally on the weekends or late evenings when the majority of the workforce had left for the day. This was the only way to avoid network resource contention [12]. There are a number of problems with this approach. The first is that backups can only be performed, at best, every night. This may not be sufficient in those environments where the RPO is very low. It also means that any recovery operations will impact the production LAN and that recovery operations will be significantly slowed because of traffic on the LAN during peak hours [13]. Furthermore, there are some applications in which there are no off-peak hours. In these cases, business continuity must be maintained 24 hours a day, 7 days a week and traditional backup methods can not be used. In short, the traditional approach will not be sufficient in low RPO, low RTO settings (class 3 and 4 environments).

4.3 Fundamentals of SAN

SAN is a network established in a similar manner to a Local Area Network (LAN). Like the LAN, the SAN is a group of computers in relatively close proximity sharing a common communication network, and one or more servers. However, unlike the

LAN, the SAN is designed for the sole purpose of establishing a direct connection between a host server and storage devices, such as RAIDs, tapes or hard drives. The SAN is useful because it creates a means by which the storage bus can be extended beyond the physical limitations of the host bus itself, allowing data to be transferred in blocks. Additionally, SANs allow storage devices to be shared by multiple hosts, regardless of platform. Fibre Channel is one of the fastest SAN technologies in existence, and SCSI is one of the most efficient protocols used for the storage of data [14]. Together, Fibre Channel and SCSI provide a very effective storage solution.

The introduction of networking concepts and technologies as a replacement of a single, direct connection allows completely rethinking the server/storage relationship and building new storage infrastructures. SANs are networks in all respects and present all the features typical of networking technologies. The most important characteristics inherited from the networking world are:

- serial transport, that allows shipping of data over long distances at high rates
- packetizing of data, to achieve high efficiency and fair sharing of network resources
- addressing schemes that support very large device population
- routing capabilities, to provide multiple, redundant paths between source and destination devices
- a layered architecture: by dividing the architecture in multiple, independent layers, different protocols can be transported at the upper layers and different interfaces can be used at the lower ones.

In the following figure [15] there is a basic SAN layout:

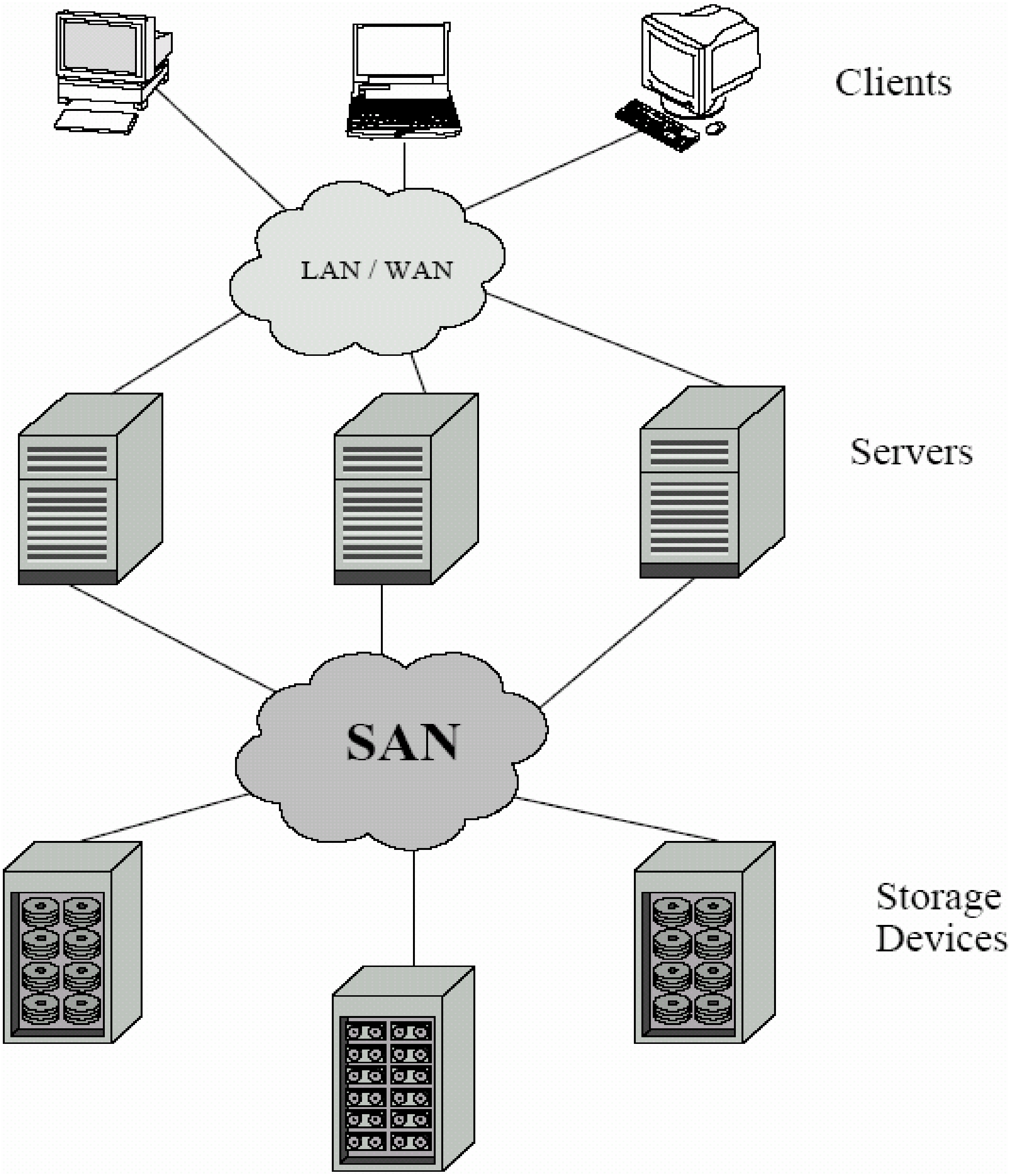


Figure 4.1: Basic SAN layout

4.4 Network Protocols for SAN

SANs can be built using different networking technologies, however storage traffic is very demanding and not all the existing technologies are able to satisfy its requirements.

- Fibre Channel
- iSCSI
- iFCP

- FCIP

4.5 Fibre Channel (FC)

The Fibre Channel (FC) protocol is an ANSI standard protocol built for storage area networking. It tries to combine the performance and reliability characteristics of an I/O interface with the flexibility and connectivity of a network. Fibre Channel provides a means by which storage takes place across a serial network. SCSI command descriptor blocks originate from the host but the SCSI bus used for transport has been replaced with Fibre Channel connections. Unlike the SCSI bus, the FC resources can be connected using switches and hubs. Fibre Channel networks can currently achieve throughputs as high as 4 Gbps and it is predicted that 10 Gbps capacity is possible in the near future. Because of its layered architecture, it supports multiple serial transmission media (optical fiber and copper) at gigabit speed and is capable of transporting multiple protocols (SCSI, IP and others).

Fibre Channel allows consolidated storage to both disk and tape. Storage pools can be shared, management of storage is centralized and backup/recovery operations are made significantly easier. The most important characteristic is that it provides for LAN-free backup and is thus very well suited for disaster recovery and replication needs. The serial interface used for Fibre Channel can span a distance of nearly 10 km while a typical SCSI bus only allows separations of a few meters. Fibre Channel also allows far more devices to be connected. A SCSI implementation is limited to 16 devices but an FC implementation can have 126 nodes per loop and 16 million addresses per fabric [16].

4.5.1. Origins of FC

Fibre Channel was specifically designed for computing environments and is based on the assumption that the transport media is a reliable media which standards require bit error rate to be less than 10^{-12}, therefore error recovery mechanisms are reduced to a minimum and are mostly left to upper layer protocols. All the components of the network have been specifically designed to avoid frame dropping due to congestion. A simple, credit-based mechanism is used for flow and congestion control. Data is fragmented and encapsulated in network protocols with minimum overhead in order to achieve high efficiency. These characteristics of the Fibre Channel data-path make an almost fully hardware-based implementation feasible, with minimum software intervention.

In order to facilitate the communication that takes place between Fibre Channel devices, some devices act as initiators, and some act as targets. Additionally, accessory devices, such as switches and bridges, may be present in any given Fibre Channel network. The initiators and targets are known as end devices, and they are usually either servers or storage devices such as disks, RAIDs, and tape drives. The servers are the initiators, since they contain the applications that originate service and task management requests to be processed by the targets. A target is a storage device that receives these requests from an initiator and guides them to the appropriate locations, where they are then executed. The servers must therefore initiate the actual processing mechanism whereby information is stored and retrieved from the storage devices. The storage devices must simply wait for this to take place, and respond appropriately to commands as the initiators issue them.

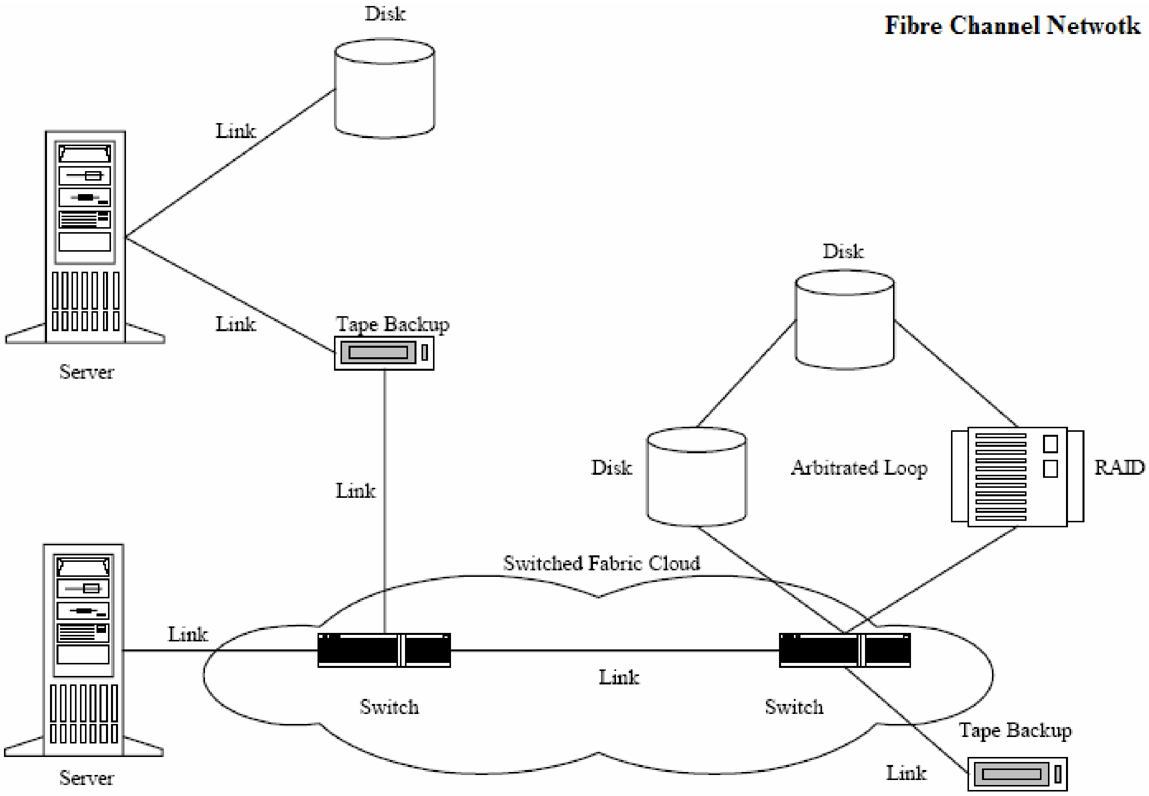


Figure 4.2: Sample Fibre Channel network

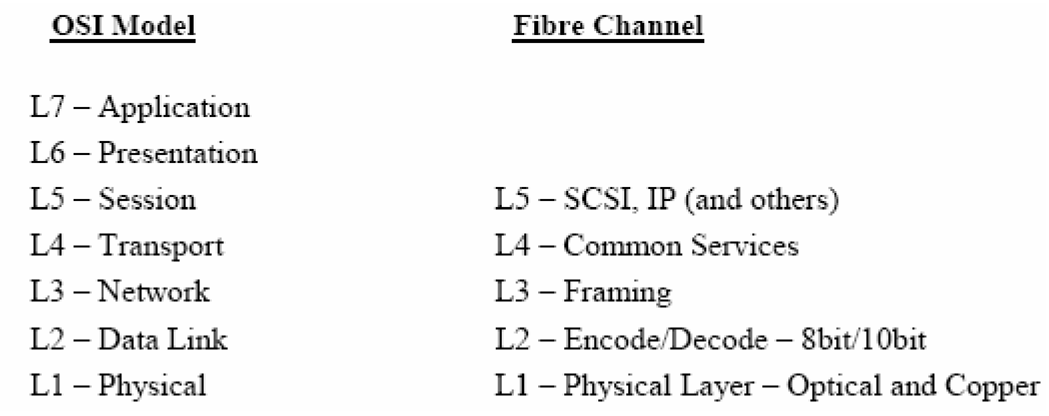
Once any combination of Fibre Channel devices are physically attached, they must initialize in order to gain access to one another. In this manner, the devices create either a link or an Arbitrated Loop. In any link, only two devices are present, and a generic addressing scheme is used. On the other hand, a loop among Fibre Channel devices may involve 2 to 126 end devices with or without the addition of one switch. The Arbitrated Loop has been defined in order to allow multiple end devices to exchange information without depending on a switch, thereby reducing costs and other maintenance issues introduced when a network is excessively reliant upon accessory devices. For this reason, a specific addressing scheme must be used on a loop in which each device is identified in a unique manner. Any number of loops and links may also be interconnected via switches, and a switch, or fabric, topology may consist of one or more links between switches and other switches or end devices.

After two Fibre Channel devices have initialized, data may be transferred between them in either direction via frames. This data transfer may or may not be connection based, depending on the class of service. Analogous to a packet in IP, a Fibre Channel frame is the unit, ranging from 256 to 2112 bytes in length, which carries a segment of data over the Fibre Channel bus. Frames contain fields which provide the Fibre Channel protocol with its own addressing scheme, sequence identifiers, exchange identifiers, Cyclical Redundancy Checks for error detection purposes, and other elements. Additionally, the Fibre Channel protocol guarantees that these frames are delivered in order between initiators and targets, if they are directly connected. However, in-order delivery between an end device and a switch is negotiable for certain classes of service. A frame sequence consists of one or more Fibre Channel frames, and any given sequence may take place in only one direction. An exchange may take place either in one direction or both, and it is comprised of one or more consecutive frame sequences. The Login frames, which are exchanged between devices directly after initialization, allow the devices to convey various operational parameters to one another, such as protocol version, buffer space, and class or classes of service supported. This exchange eventually leads to those involving the actual I/O requests and data transmission.

4.5.2. Fundamentals of FC

Fibre Channel is a Storage Area Networking protocol comprised of five layers, FC-0 through FC-4, which correspond to layers one through five of the OSI model. The following chart [17] attempts to draw parallels between the Fibre Channel layering and the theoretical OSI model, for purposes of comparison:

Table 4.1: Comparison of Fibre Channel layering and OSI model



OSI Model	Fibre Channel
L7 – Application	
L6 – Presentation	
L5 – Session	L5 – SCSI, IP (and others)
L4 – Transport	L4 – Common Services
L3 – Network	L3 – Framing
L2 – Data Link	L2 – Encode/Decode – 8bit/10bit
L1 – Physical	L1 – Physical Layer – Optical and Copper

Fibre Channel protocols are organized in a layered architecture as mentioned above and shown in the following figure. If a node has multiple ports, it has a separate instance of the levels FC-0 to FC-2 for each of them and a unique instance of FC-3 and FC-4.

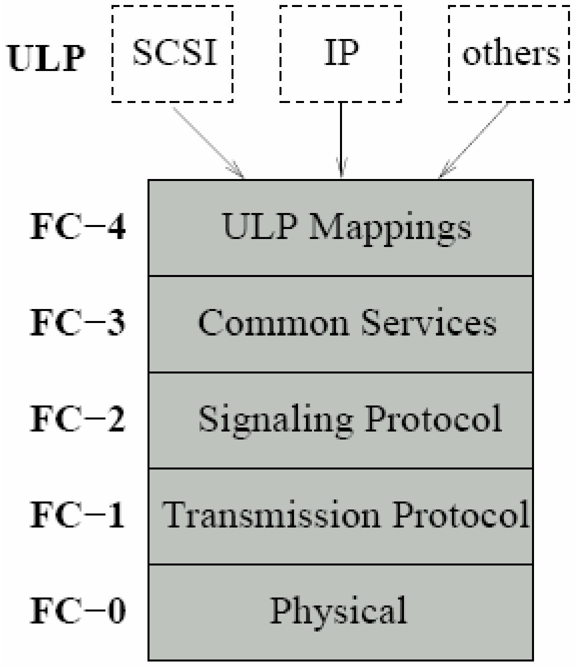


Figure 4.3: Fibre Channel layers

FC-0: This layer contains the description of the physical interface, including supported media types (multi- and mono-mode optical fiber, twisted pair, copper coax, etc.), cabling requirements, connectors, transmitters and receivers specifications (wavelengths, data rates, signal levels, etc.). Supported data rates, according to the standard [18], are 1.06 Gbps, 2.12 Gbps (commercially available) and 4.25 Gbps. 10 Gbps specifications are under development.

FC-1: The transmission protocol layer defines the encoding scheme used on the link and low level link control facilities (initialization, reset, etc.). Data are encoded before transmission on the physical media and decoded upon reception. The encoding scheme that Fibre Channel uses is called “8B/10B” and was originally developed by IBM. Every 8-bit sequence is encoded in a 10-bit one that has important characteristics that increase the robustness of the transmission. The encoding guarantees that in every sequence there is a minimum number of transitions from the ”low” to the ”high” level (or vice-versa) and that an overall balance of high and low bits is maintained on the link. This prevents receiver synchronization problems and aids in error detection.

FC-2: The signaling protocol defined at this layer specifies the rules to transfer data blocks end-to-end. It defines the frame format, addressing, information unit segmentation and reassembly, flow control, error detection and recovery.

FC-3: The FC-3 layers deals with common services in a node. As no such services have been defined, this layer can be considered a placeholder for future facilities. Functions that might fit at this level are encryption and compression.

FC-4: Fibre Channel is capable of transporting multiple upper-layer protocols (ULP). This layer defines how each of these protocols must use the services offered by the lower layers to transport its information units across the Fibre Channel network. For example, FCP is the mapping for SCSI-3 and specifies how SCSI commands, data, parameters and status information are to be packed into FC-2 entities. The IP mapping deals with the encapsulation of IP packets in Fibre Channel frames and the resolution of IP addresses.

The format of a Fibre Channel frame is shown in figure 4.2. The maximum size of a frame is 2148 bytes and the size must be evenly divisible by 4, as the minimum unit that the transmission layer accepts is a 4-byte word.

Table 4.2: Frame format (sizes are in bytes)

SOF	FRAME HEADER	OPT HDR	DATA FIELD	PAD	CRC	EOF
(4)	(24)		(0 to 2112)		(4)	(4)

SOF: Each frame starts with a delimiter, called “Start-of-Frame” (SOF). This preamble marks the beginning of the frame, serves multiple purposes and can take different forms. It declares the class of service of the transported frame and may request the allocation of resources in case of connection-oriented services. It also informs the receiver whether the frame is the first of its sequence or belongs to an already active one.

Frame Header: The next field is the frame header, which contains source and destination addresses, exchange and sequence identifiers, numbering, transported protocol and other control information.

Data Field: The data field contains the payload of the frame and is of variable length. Optional headers containing additional control fields can take up to 112 bytes of the data field. The length of the data field must be evenly divisible by 4, so up to 3 bytes of padding can be added to the actual payload.

CRC: The Cyclic Redundant Check is a 4-byte value used to verify the data integrity of the frame. It is calculated with a well-known algorithm on the frame header and the data field prior to encoding for transmission and after decoding upon reception.

EOF: The “End-Of-Frame” (EOF) delimiter marks the end of the frame. A particular value of this field is used to notify the receiver that the frame is the last of its sequence. In case of connection-oriented classes of service, specific values of the EOF indicate that the dedicated connection can be removed.

4.6 Internet SCSI (iSCSI)

iSCSI protocol [19,20] is an official standard ratified on February 11, 2003 by the Internet Engineering Task Force that allows the use of the SCSI protocol over TCP/IP networks. The iSCSI protocol uses TCP/IP for its data transfer. Unlike other network storage protocols, such as Fibre Channel, it requires only the simple and ubiquitous Ethernet interface or any other TCP/IP-capable network to operate. This enables low-cost centralization of storage without all of the usual expense normally associated with Fibre Channel storage area networks.

iSCSI is designed to transport Protocol Data Units, or PDUs, over TCP/IP connections between the iSCSI Ports which reside on end devices. Unlike FCIP and iFCP, the iSCSI protocol does not incorporate Fibre Channel frames, and may thus be used by a device as a replacement for Fibre Channel altogether. In this protocol, multiple initiators may connect with a given target, and a single initiator may connect with multiple targets. Additionally, multiple connections can exist between any given pair of devices. Every connection, or set of connections, which exists between an initiator and a target is known as a session. In order to avoid confusion, initiators and targets, sessions and connections must all be uniquely identified. Furthermore, iSCSI initiators and targets may negotiate a large number of parameters, dictating which authentication method, if any, shall be used, PDU sizes, PDU burst sizes, whether CRCs are used, and many other features. Some of the negotiated features are applicable only to single connections, while others apply to entire sessions. After the parameter negotiation phase has taken place, the iSCSI nodes exchange encapsulated SCSI Commands in order to execute actual I/O operations.

In the figure 4-4, each server, workstation and storage device support the Ethernet interface and a stack of the iSCSI protocol. IP routers and Ethernet switches are used for network connections.

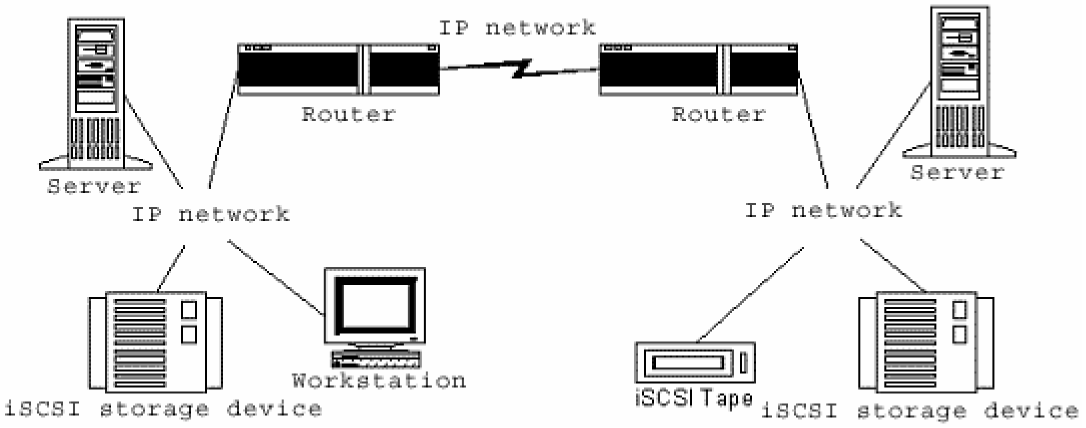


Figure 4.4: Sample iSCSI network

Perhaps the most obvious disadvantage associated with Fibre Channel is its cost. The cause of this dramatic cost increase is two-fold. First, Fibre Channel is only realizable in hardware, with special host bus adapters (HBAs) necessary for each node and FC switches required to interconnect FC components. Essentially, this means that in addition to having a production LAN, a totally separate network must be purchased and maintained in order to use FC. The second cost component is incurred with the management of the FC network. Fibre Channel is not an ordinary networking protocol and requires people with specialized training to operate it. Therefore, it is necessary to hire additional system administrators to oversee the correct operation of the FC network while others may be required to oversee the company's production LAN. This makes the total cost of ownership (TCO) of FC so high that only government agencies and medium/large corporations could ever afford it [21]. There is no small business or home office market for FC.

Another disadvantage of Fibre Channel is its inherent distance limitation. Typically, the range of FC is limited to a few hundred meters and if larger distances are required, significant money must be invested in expensive FC extenders. In the past, this was not seen as a critical flaw in Fibre Channel but in the wake of the attacks on September 11, 2001, a renewed interest in disaster recovery has arisen. This interest is especially strong in remote disaster recovery operations where sites are separated by hundreds or even thousands of miles. Of course, a SAN implemented using FC can not be distributed over remote sites because of its inherent distance limitation. It should be noted that Fibre Channel vendors are attempting to remedy this with Fibre Channel over IP (FCIP) and the Internet Fibre Channel Protocol (iFCP) [17]. Again, these extensions to the FC protocol are still expensive to implement and assume that the user has already invested in base FC components such as switches and HBAs.

The high prices and difficultly in extending Fibre Channel over large distances have made it not only prohibitively expensive for smaller companies, but even large organizations are reluctant to spend extra money to use it over long distances. Any data that must be protected should be replicated at another site over 100 miles away. Fibre Channel does not allow for remote disaster recovery at a reasonable cost. The need for a high-speed solution, with no inherent distance limitation, at a lower cost is critical in helping businesses (large and small) realize a strong disaster recovery and business continuity infrastructure.

4.6.1. Origins of iSCSI

The iSCSI protocol was designed as a more cost effective solution to traditional SAN demands. At the same time, a great effort was made to add more functionality without affecting performance. Until recently, any high-performance SAN had to be implemented using Fibre Channel. Fibre Channel was viewed as the best performing protocol for SAN deployment since it was among the first protocols to offer transfer rates of 1 Gbps and can now support 4 Gbps transactions. Also, Fibre Channel is a very “elegant” protocol in that it was designed from the beginning for Storage Area Networking. In the table 4-3, there is an illustration how iSCSI compares to FCIP and iFCP [22].

Table 4.3: Comparison of IP-based Protocols

Protocol Attributes	**iFCP**	**iSCSI**	**FCIP**
Implementation	Native IP Transport	Native IP Transport	Encapsulation, Tunneling
SCSI Encapsulation	FCP	iSCSI Layer	FCP
End Device Interface	FC/FCP	IP/iSCSI	FC/FCP
FC Device Support	Yes	No	Yes
Relative Cost	$$	$	$$$

Figure 4.5 [23] shows diagram of networks based on them.

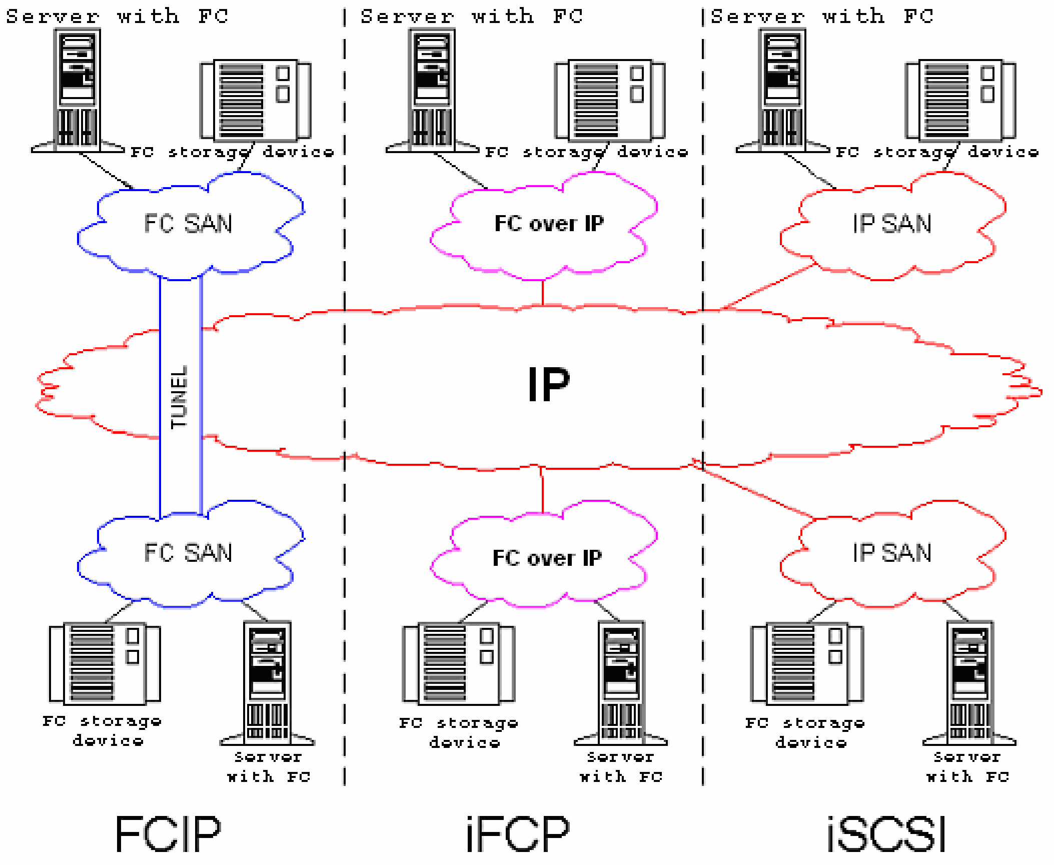


Figure 4.5: Diagram of IP Storage networks

Both iFCP and iSCSI using native TCP/IP for transport while FCIP uses the proprietary encapsulation provided with Fibre Channel. Thus, iFCP and iSCSI are both equal in terms of distance limitations (of which there are none). Device Integration with Internet Storage Name Service (iSNS) is used for fast discovery of target nodes. This is a completely separate IETF standard but both iSCSI and iFCP can take advantage of it. Finally, FC device support is obviously provided with iFCP and FCIP. However, iSCSI does not provide for this within its protocol. There are switches that can be used for iSCSI to FC bridging so that organizations with existing FC components can deploy iSCSI solutions that will function with their expensive FC infrastructure.

The Internet Engineering Task Force (IETF) was keenly aware of the cost and distance drawbacks of FC-based SANs and decided to undertake a project to develop a new SAN protocol that could be affordable while operating at long-distances at high-speeds. Their solution was to develop and adopt the iSCSI protocol. iSCSI is a true SAN protocol that was designed from the beginning to do everything Fibre Channel could do, but also do things that FC could not. In addition, careful steps were taken in order to ensure iSCSI remained an affordable option for smaller businesses.

iSCSI provides the same service that Fibre Channel provides, that is, transporting SCSI information over a storage area network. iSCSI differs in that it provides this service using already existing networks by encapsulating SCSI data in TCP/IP packets (as shown in the following figure). This issue was first introduced in paper [24] where application level framing is presented as an architectural consideration for a next-generation transport protocol. It is this difference that makes iSCSI so versatile and cost-effective. iSCSI generally takes a performance hit because of the extra work required to strip off the TCP/IP layer data. Fibre Channel does not suffer from this problem but has a number of other problems that may make iSCSI a better fit for many users.

Table 4.4: iSCSI Encapsulation

Ethernet Header	IP Header	TCP Header	iSCSI Header	iSCSI Data	Ethernet Trailer

Consider the cost associated with iSCSI and Fibre Channel. As illustrated above [21], FC is not only very expensive from a hardware perspective, but it also requires additional investment in personnel that can maintain the complex network. iSCSI on the other hand can be implemented using significantly cheaper hardware HBAs. Furthermore, iSCSI can operate over standard TCP/IP switches, which means that no additional investment is needed for specialized iSCSI switching equipment. iSCSI switches do exist however, and provide increased performance and features that standard Ethernet switches cannot provide. Since iSCSI utilizes a standard, well-known networking protocol, the average system administrator has enough expertise to maintain the iSCSI SAN – although a small amount of additional reading may be necessary to fully understand iSCSI. Nevertheless, it is no longer necessary to seek out system administrators with specialized FC training and certification. iSCSI also allows for a full software implementation. This is extremely useful for organizations that can not afford iSCSI HBAs or high performance switches. The software is simply a driver that can be downloaded and allow for iSCSI transmissions over the existing LAN. In effect, this type of implementation requires no financial investment whatsoever.

iSCSI, unlike FC, has no inherent distance restrictions. Because iSCSI transmits information using TCP/IP infrastructure, it is bounded by the same distance (and performance) limitations of TCP/IP. This means that iSCSI can be used to distribute a Storage Area Network over 100s of miles or interconnect disparate SANs separated by 100s of miles to provide for backup and recovery operations at great distances. Figure 4-6 shows an example of this.

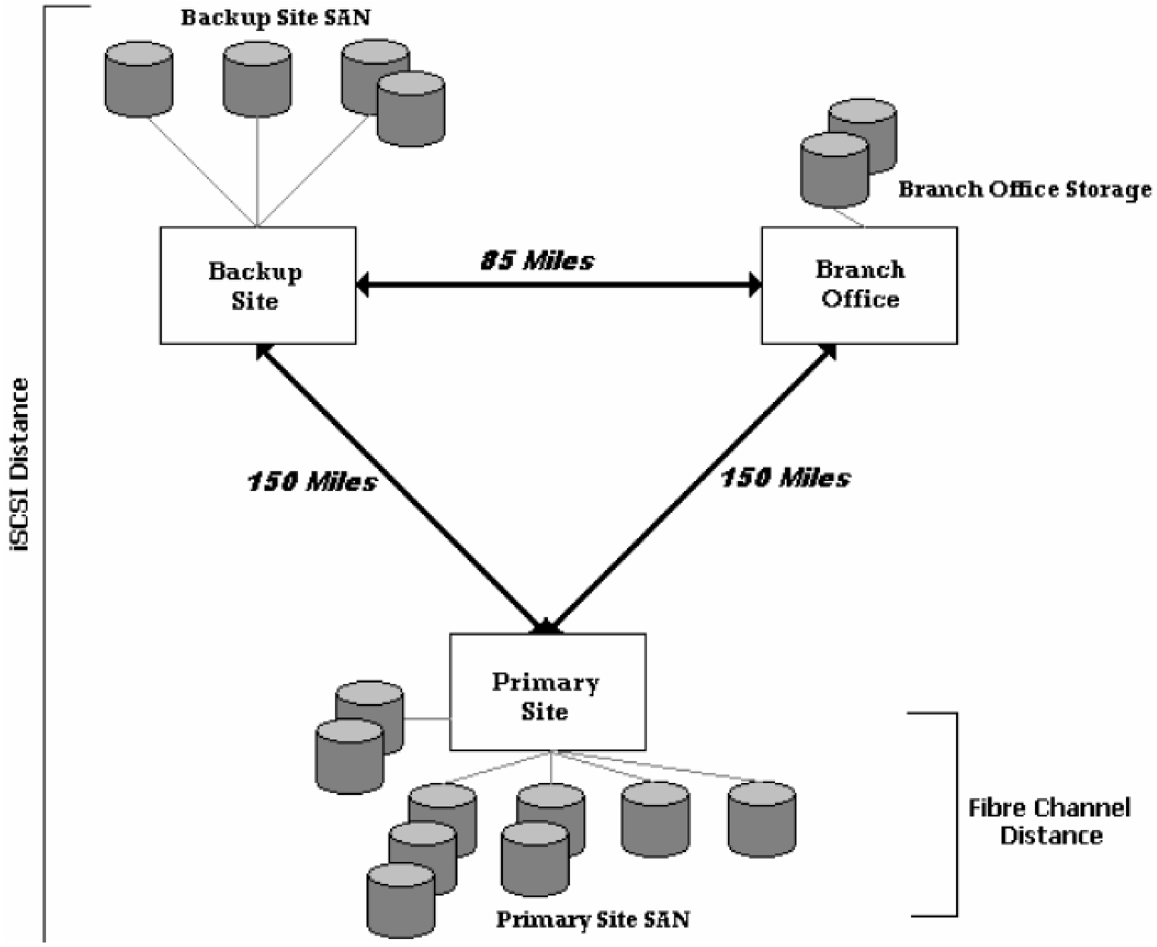


Figure 4.6: iSCSI and Fibre Channel in different distances

This increases the assurance metric of the SAN solution by decreasing the chance that a disaster could affect the primary and backup site simultaneously which would result in loss of data. Finally, the performance characteristics of iSCSI are very important if the protocol is to be successful. It has been a long held belief that iSCSI could never match the performance delivered by a Fibre Channel network. In fact, in its early form, iSCSI was associated with more affordable, albeit poor performing networks that only had a market for those who could never afford a Fibre Channel SAN. Whenever high-performance was needed, iSCSI was overlooked and FC became the network of choice. The dependence on TCP/IP for iSCSI transport was originally considered a weakness because iSCSI could never perform better than the TCP/IP fabric in which it was implemented. FC was not bound by the performance of TCP/IP and thus it was thought that FC would always be able to maintain a performance advantage. But what was once considered a flaw in the iSCSI design has turned out to be a blessing in disguise. Certainly, few people thought that Ethernet (or TCP/IP) performance would ever reach Gigabit and 10 Gigabit speeds. As such, the iSCSI protocol has been given the underlying support structure it needs to also perform at multi-gigabit per second throughputs and has now been shown to be just as capable as FC from a performance standpoint. With future projections of 40 Gbps and 100 Gbps Ethernet becoming more widespread, it is reasonable to assume that iSCSI will be able to operate at these speeds as well.

4.6.2. Fundamentals of iSCSI

Architecture of a pure SCSI is based on the client/server model. A client, for example, server or workstation, initiates requests for data reading or recording from a target - server, for example, a data storage system. Commands which are sent by the client and processed by the server are put into the Command Descriptor Block (CDB). The server executes a command which completion is indicated by a special signal alert. Encapsulation and reliable delivery of CDB transactions between initiators and targets through the TCP/IP network is the main function of the iSCSI, which is due to be implemented in the medium untypical of SCSI, potentially unreliable medium of IP networks.

Figure 4-7 is a model of the iSCSI protocol levels which allows us to get an idea of an encapsulation order of SCSI commands for their delivery through a physical carrier [25].

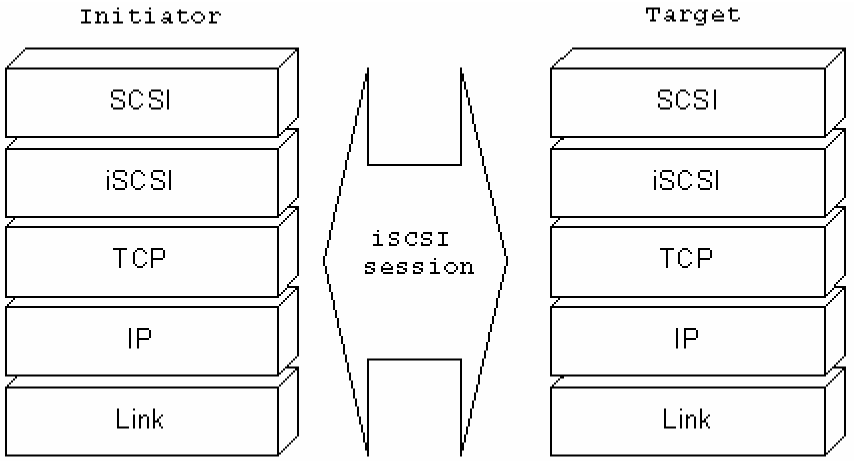


Figure 4.7: Model of lower levels of the iSCSI protocol

An iSCSI PDU is made up of several segments, the most important of which is the Basic Header Segment (BHS). The BHS is a 48-byte field that contains vital information such as the Logic Unit Number (LUN), the CDB, and other operational information. The other PDU fields, such as the Additional Header Segment (AHS), Header Digest, Data Segment, and Data Digest, are all optional. In fact, the majority of PDUs contain only a BHS. The Data Segment contains the data that is being sent to the target from the initiator. The header digest is an error check for the header of the PDU and the data digest is an error check word responsible for the data.

The iSCSI protocol controls data block transfer and confirms that I/O operations are truly completed. In its turn, it is provided via one or several TCP connections. The BHS contains iSCSI Opcodes, information about the length of the corresponding data segments, LUN information, SCSI CDBs, and other information necessary for session establishment and management. The structure of the BHS depends on the opcode contained in its first byte. This opcode will correspond to a certain PDU type such as a “Login_Request_PDU” or “Task_Management_Function_PDU”.

The iSCSI has four components [25]:

- iSCSI Address and Naming Conventions.
- iSCSI Session Management.
- iSCSI Error Handling.
- iSCSI Security.

4.6.2.1 Address and Naming Conventions

As the iSCSI devices are participants of an IP network they have individual Network Entities. Such Network Entity can have one or several iSCSI nodes.

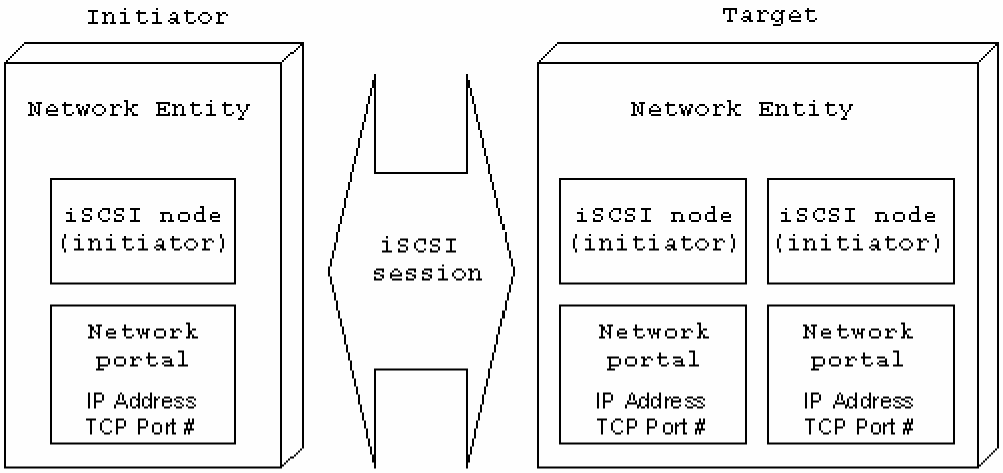


Figure 4.8: Model of Network Entities

An iSCSI node is an identifier of SCSI devices (in a network entity) available through the network. Each iSCSI node has a unique iSCSI name (up to 255 bytes) which is formed according to the rules adopted for Internet nodes. For example, fqn.com.ustar.storage.itdepartment.161. Such name has an easy-to-perceive form and can be processed by the Domain Name System (DNS). An iSCSI name provides a correct identification of an iSCSI device irrespective of its physical location. At the same time in course of handling data transfer between devices it's more convenient to use a combination of an IP address and a TCP port which are provided by a Network Portal. The iSCSI protocol together with iSCSI names provides a support for aliases which are reflected in the administration systems for better identification and management by system administrators.

4.6.2.2 Session Management

The iSCSI session consists of a Login Phase and a Full Feature Phase which is completed with a special command [25]. The Login Phase of the iSCSI is identical to the Fibre Channel Port Login process (PLOGI). It is used to adjust various parameters between two network entities and confirm an access right of an initiator. If the iSCSI Login Phase is completed successfully the target confirms the login for the initiator; otherwise, the login is not confirmed and a TCP connection breaks. As soon as the login is confirmed the iSCSI session turns to the FULL Feature Phase. If more than one TCP connection was established the iSCSI requires that each command/response pair goes through one TCP connection. Thus, each separate read or write command will be carried out without a necessity to trace each request for passing different flows. However, different transactions can be delivered through different TCP connections within one session.

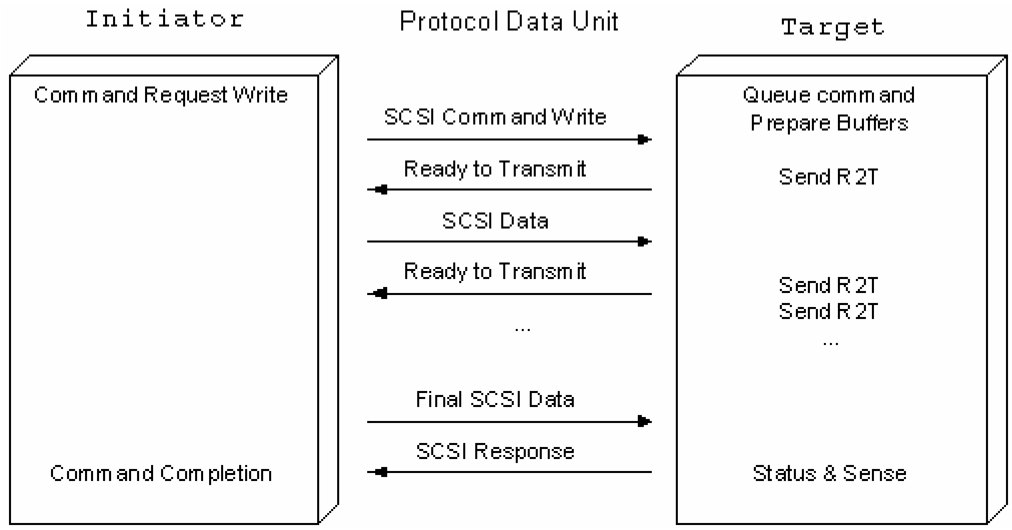


Figure 4.9: iSCSI write example

At the end of a transaction the initiator sends/receives last data and the target sends a response which confirms that data are transferred successfully. The iSCSI logout command is used to complete a session - it delivers information on reasons of its completion. It can also send information on what connection should be interrupted in case of a connection error, in order to close troublesome TCP connections.

4.6.2.3 Error Handling

Because of a high probability of errors in data delivery in some IP networks, especially WAN, where the iSCSI can work, the protocol provides a great deal of measures for handling errors. So that error handling and recovery can work correctly both the initiator and the target must be able to buffer commands before they are confirmed. Each terminal must have a possibility to recover selectively a lost or damaged PDU within a transaction for recovery of data transfer.

Here is the hierarchy of the error handling and recovery after failures in the iSCSI:

- The lowest level - identification of an error and data recovery on the SCSI task level, for example, repeated transfer of a lost or damaged PDU.
- Next level - a TCP connection which transfers a SCSI task can have errors. In this case there is an attempt to recover the connection.
- At last, the iSCSI session can be damaged. Termination and recovery of a session are usually not required if recovery is implemented correctly on other levels, but the opposite can happen. Such situation requires that all TCP

connections be closed, all tasks, under fulfilled SCSI commands be completed, and the session be restarted via the repeated login.

4.6.2.4 Security

As the iSCSI can be used in networks where data can be accessed illegally, the specification allows for different security methods [26]. Such encoding means as IPSec which use lower levels do not require additional matching because they are transparent for higher levels and for the iSCSI as well. Various solutions can be used for authentication, for example, Kerberos or Private Keys Exchange, an iSNS server can be used as a repository of keys.

4.7 Internet Fibre Channel Protocol (iFCP)

iFCP is designed for customers who may have a wide range of Fibre Channel devices (i.e. host bus adapters, subsystems, hubs, switches, etc.), and want the flexibility to interconnect these devices with IP network. iFCP can interconnect Fibre Channel SANs with IP, as well as allow customers the freedom to use TCP/IP networks in place of Fibre Channel networks for the SAN itself [27].

iFCP is a TCP/IP based protocol for connection of FC data storage systems using the IP infrastructure together or instead of FC switching and routing elements. In other words, iFCP is a protocol which provides FC traffic delivery over the TCP/IP transport between iFCP gateways. In this protocol an FC transport level is replaced with a transport of the IP network, the traffic between FC devices is routed and switched by the means of TCP/IP. The iFCP protocol allows connecting current FC data storage systems to an IP network with a support of network services which are necessary for these devices.

Figure 4.10 shows how an iFCP protocol stack looks like [25].

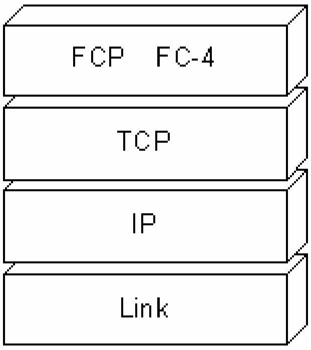


Figure 4.10: Lower levels of the iFCP protocol

According to the specification iFCP:

- Overlays FC frames for their delivery to a predetermined TCP connection
- FC services of message delivery and routing are overlapped in the iFCP gateway device; therefore, network structures and components of the FC do not mix in one FC SAN but are managed by the TCP/IP means
- Dynamically creates IP tunnels for FC frames.

An important feature of the iFCP is that this protocol provides an FC device-to-device connection via an IP network which is a more flexible scheme in comparison to the SAN-to-SAN. For example, if the iFCP has a TCP connection between two FC devices such connection can have its own QoS level which will be different from a QoS level of another pair of FC devices.

4.8 Fibre Channel over IP (FCIP)

Fibre Channel over IP standard takes advantage of the installed base of Fibre Channel SANs, and the need to interconnect these SANs to support mission-critical environments. SANs provide the high performance and reliability required to support business continuance and disaster tolerance environments, including remote backup/archiving, high availability, remote mirroring, and centralized management.

FCIP is a tunnel protocol based on the TCP/IP which is designed for connection of geographically far FC SANs without affecting FC and IP protocols [28]. The most revolutionary protocol among these three is Fibre Channel over IP. It doesn't bring in any changes into the SAN structure and organization of storage area systems. The main idea of this protocol is to make functional integration of geographically remote storage networks.

Figure 4.11 [27] shows the stack of the FCIP protocol.

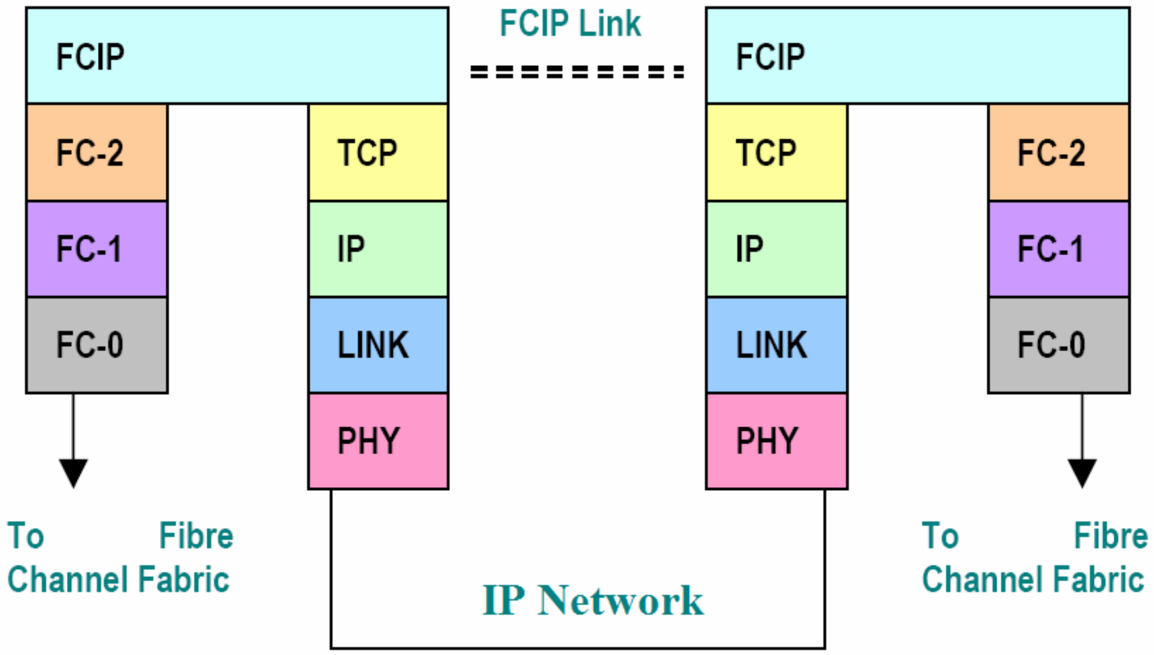


Figure 4.11: FCIP protocol stack

FCIP helps to effectively solve a problem of geographical distribution, and integration of SANs on large distances. This protocol is entirely transparent for existent FC SANs and involves usage of infrastructure of modern MAN/WAN networks. So, if you want to merge geographically remote FC SANs with new functionality enabled you will have to get just one FCIP gateway and connection to MAN/WAN networks. A geographically distributed SAN based on the FCIP is taken by SAN devices as a usual FC network, and it is seen as a usual IP traffic for a MAN/WAN network it is connected to.

The IETF RFC [29] specifies that:

- Rules of encapsulation of FC frames for delivery through TCP/IP
- Rules of using encapsulation for creation of a virtual connection between FC devices and elements of an FC network
- TCP/IP environment for support of creation of a virtual connection and support of FC traffic tunneling through an IP network including safety, integrity of data and a data rate issue.

There are some applied problems which can be successfully solved using the FCIP protocol: remote backup, data recovery and a shared data access. With high-speed MAN/WAN communications one can also use synchronous data doubling and a shared distributed access to data storage systems.

4.9 iSCSI-based SAN vs. FC-based SAN for Disaster Recovery Operations

A SAN Environment typically consists of four major components:

- End user platforms such as desktops/thin clients
- Server systems
- Storage devices and storage sub-systems
- Interconnect entities

The interconnect entities are switches, hubs and bridges. Historically, SANs have been connected using fibre channels. While fibre channels work well in providing high speed SANS, there are some drawbacks associated with this technology [30]. Some of these are:

- High Total Cost of Ownership. Fibre technology is expensive and the Total Cost of Ownership is extremely high.
- Limited Operating Distance. Theoretically fibre channels can operate at distances of 10 kilometers. In practical situations the distance is much less.
- Steep Learning Curve. The SAN is not based on popular technology, rather is based on technology which is new to System Administrators. The cost of training is high and the time required to train people to the level where they can manage the SAN is also high.

The technology that has evolved to solve the problems of the historical FC-based SANs is IP-based SANs. This has taken two directions - Pure IP-based SANs and FC-IP mixed SANs. In the FC-IP mixed SANs, technologies like FCIP and iFCP have evolved. On the pure IP side iSCSI protocol is a promising solution. Typically, the existing Gigabit Ethernet LAN and WAN infrastructure of organizations that go for SAN solutions is IP based. This makes an IP based solution very lucrative and seemingly simpler.

4.9.1. The major differences between iSCSI and FC

Some of the major differences between iSCSI and fibre channels from a technology perspective are:

- FC requires a separate fibre-optic network for the SAN, while iSCSI uses the existing Gigabit Ethernet LAN. FC requires a completely separate set of fibres and switches. However, that does not mean the iSCSI network is for free. iSCSI requires additional network adapters dedicated to the SAN, in addition to the server's ordinary Gigabit Ethernet network interface card. This means using a PCI slot. The iSCSI host-bus adapters will consume a Gigabit Ethernet switch port and add to the traffic on the LAN.

- A server's FC host-bus adapters must be connected directly to the SAN switch. An iSCSI host-bus adapter can connect to a storage router (the iSCSI equivalent of the FC SAN switch) anywhere on the Gigabit Ethernet SAN. There is more flexibility when it comes to building a complex iSCSI-based SAN.

- iSCSI is going to be on the Ethernet fast track. While FC is still in the painful process of migrating from 1Gbps to 2Gbps, with 4Gbps as the next step. Ethernet is rapidly progressing to 10Gbps and 40Gbps. Plus, Ethernet has a strong tradition of genuine multi-vendor interoperability, and FC does not.

Some of the major differences from a Management perspective are:

- Reduced Total Cost of Ownership (TCO) - The major components of the cost of ownership are the initial cost per port, maintenance costs and training costs. All three are low compared to fibre channels in a iSCSI based solution

- Interoperability - ISCSI is now an IETF standard. Organizations such as Storage Networking Industry Association (SNIA) and the University of New Hampshire (UNH) Interoperability Lab have hosted a number of multi-vendor tests and demonstrations to ensure interoperability. Interoperable products exist in the market. This means that iSCSI is a promising hassle-free technology for the end user.

- Simple - An iSCSI SAN utilizes Gigabit Ethernet network components and enables network administrators to continue working in their all familiar IP environment. This simplicity is a great advantage. With more plug and play products coming up in the market, the simplicity factor is becoming more and more important.

4.9.2. Combination of FC and iSCSI for remote data replication

The combination of Fibre Channel and iSCSI is becoming popular to provide remote data replication for disaster recovery. In this environment, users have built a Fibre Channel SAN infrastructure to access primary storage repositories. Having done so, they require the broader capabilities found in many Fibre Channel disk subsystems but are restricted by Fibre Channel's theoretical 10km distance limit. Fibre Channel-to-iSCSI lets them span distances required for disaster recovery.

5. BUILDING HIGH-PERFORMANCE iSCSI-BASED SAN

5.1 Understanding iSCSI for Better iSCSI Performance

The iSCSI protocol has emerged as a transport for carrying SCSI block-level access protocol over the ubiquitous TCP protocol. It enables a client's block-level access to remote storage data over an existing IP infrastructure. This can potentially reduce the cost of storage system greatly, and facilitate the remote backup, mirroring applications, etc. Due to the ubiquity and maturity of TCP/IP networks, iSCSI has gained a lot of momentum since its inception.

On the other hand, the iSCSI-based storage is quite different from a traditional one, e.g. FC-based SAN. A traditional storage system is often physically restricted to a limited environment, e.g. in a data center. It also adopts a transport protocol specially tailored to this environment, e.g. parallel SCSI bus, Fibre Channel, etc. These characteristics make the storage system tend to be more robust, and achieve more predictable performance. It is much easier to estimate the performance and potential bottleneck by observing the workload. While in an iSCSI storage, the transport is no longer restricted to a small area. The initiator and the target can be far apart. The networking technology in between can be diverse and heterogeneous, e.g. ATM, Optical DWDM, Ethernet, Wireless, satellite, etc. The network condition can be congested and dynamically changing. Packets may experience long delay or even loss and retransmission, etc. Thus, the performance characteristics of the iSCSI storage are quite different from the traditional one.

To take advantage of iSCSI protocol to build iSCSI storage systems, it is needed to better understand the iSCSI characteristics, e.g. the performance characteristics for different iSCSI and TCP parameters.

5.2 The iSCSI Data Transfer Model

Figure 5.1 shows the iSCSI architecture model. iSCSI builds on top of TCP transport layer. For an iSCSI initiator to communicate with a target, they need to establish a session between them. Within a session, one or multiple TCP connections are established. The data and commands exchange occurs within the context of the session [20].

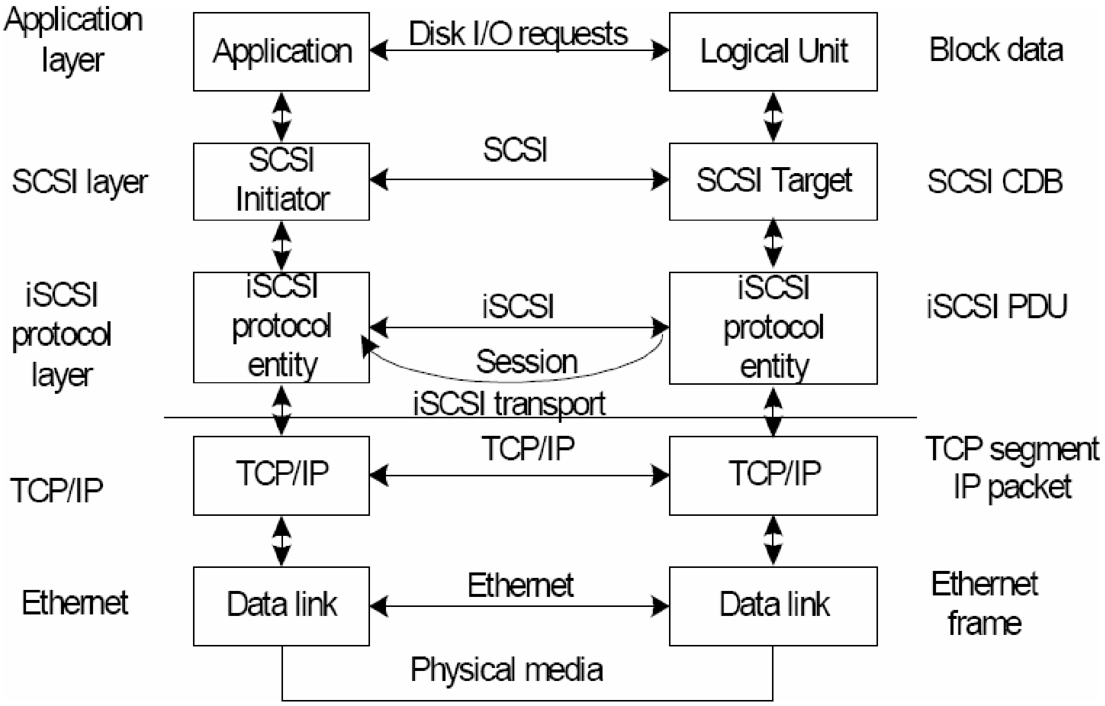


Figure 5.1: The iSCSI model

Figure 5.2 shows iSCSI command execution by illustrating a typical Write command. The execution consists of three phases: Command, Data and Status response [20]. In the Command phase, The SCSI command which in the form of Command Descriptor Block (CDB) is incorporated in an iSCSI command PDU. The CDB describes the operation and associated parameters, e.g. the logical block address (LBA) and the length of the requested data. The length of the data is bounded by a negotiable parameter “MaxBurstLength”. During the Data phase, the data PDUs are transmitted from an initiator to a target. Normally, the initiator needs to wait for “Ready to Receive (R2T)” message before it can send out data (solicited data). However, both initiator and target can negotiate a parameter “FirstBurstLength” to speed up the data transmission without waiting. FirstBurstLength is used to govern how much data (unsolicited data) can be sent to the target without receiving “Ready to Receive (R2T)”. A R2T PDU specifies the offset and length of the expected data. To further speed up the data transfer, one data PDU can be embedded in the command PDU if “ImmediateData” parameter is enabled during the parameter negotiation. This should be very beneficial for small write operation. The Status PDU is returned once the command is finished. Finally, these messages are encapsulated into TCP/IP packets, where the packet size is bounded by MTU parameter in TCP. MTU is determined by the smallest frame size along the path to the destination. Within an Ethernet LAN, the maximum frame size is 1500 bytes (Gigabit Ethernet supports Jumbo frame). The MTU is 1460 bytes (40 bytes for IP and TCP header). When the iSCSI PDU size is greater than the segment size, the PDU is further fragmented into smaller packets. The size of iSCSI parameters like: MaxBurstLength, FirstBurstLength and PDU size all have certain impact to the iSCSI performance. However, the iSCSI performance is also significantly affected by the underlying TCP parameters, such as TCP windows size, MTU size.

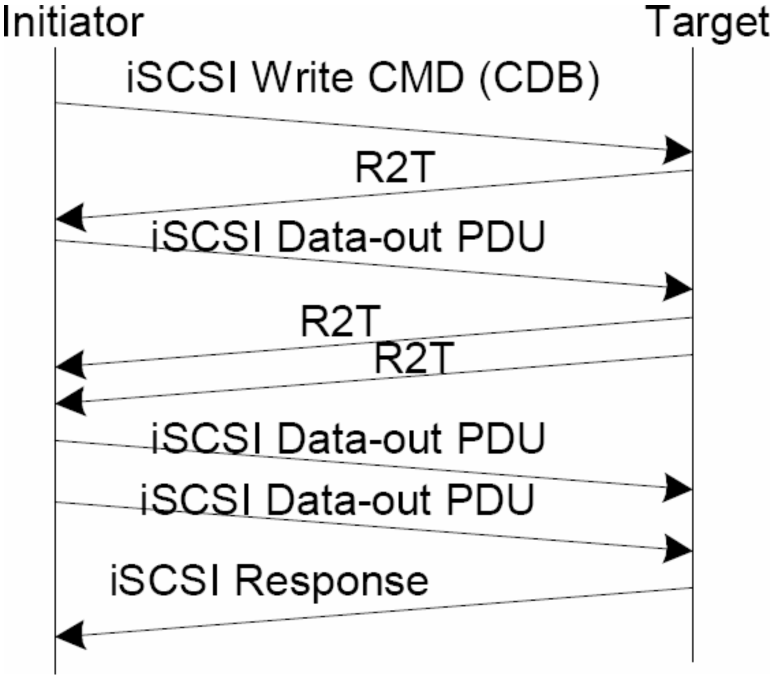


Figure 5.2: The command execution sequence

5.3 Test configuration

A HP Integrity rx2620 Unix server, and three HP EVA storage boxes are used in this test configuration. The server is realized using a system with HP-UX 11v2 Enterprise Edition operating system, two Intel Itanium2 1.6GHz processors, 4GB of memory, two Gigabit Ethernet cards, and two FC host bus adaptors. The one of the storage boxes is connected to server via FC ports, the others are connected to server via EVA iSCSI Connectivity Option. In addition to server and storage boxes, there are also two 2Gbps FC fabric switches and two Gigabit Ethernet switch in the configuration.

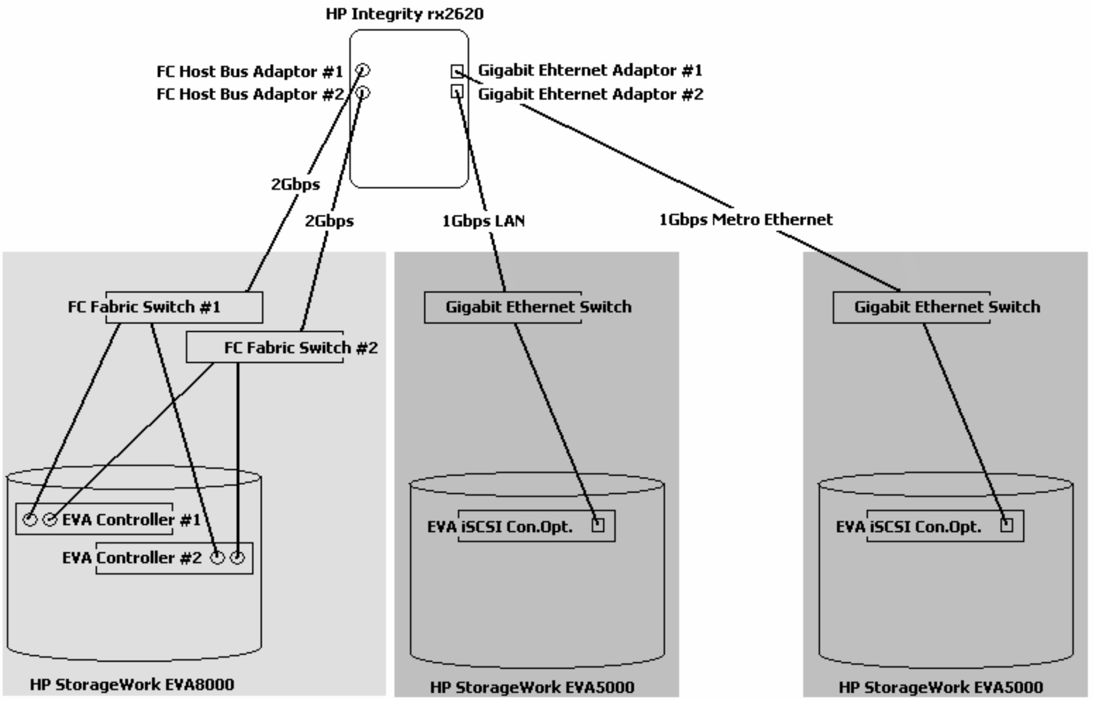


Figure 5.3: Test configuration

While FC-based components such as Host Bus Adaptors, Fabric switches, and EVA controllers are in a redundant configuration, each component of Gigabit-based configuration is single.

5.4 Performance Analysis

In this section, the results of the effect of iSCSI parameters in iSCSI layer and TCP parameters in TCP layer to iSCSI data access performance is presented. It is also examined both how the CPU usage affects the iSCSI performance and how PDU burst size affects the CPU utilization.

The real performance measurement is chosen instead of using a network simulator, e.g. NS2 [31]. A simulation approach can be offer more flexibility, e.g. paper [32]. However, real measurement is much more accurate. Although papers [33, 34] represent the similar approach, these papers doesn't examine iSCSI performance in view of CPU usage.

5.4.1. The Effect of the iSCSI Parameters

The effect of different iSCSI PDU sizes is examined. TCP window size and MTU size are set to 20 and 296, respectively. While figure 5.4 shows the write throughput with varying PDU sizes, figure 5.5 shows the read throughput with varying PDU sizes. It is found out that at larger burst data size, the PDU size makes difference. For a large burst size, e.g. 1024K, with the PDU size of 8K, there will be 128 PDUs, while with PDU size of 1K, and then there will be 1024 PDUs. More PDUs cause more R2T messages, and potentially more waiting for the R2T signals. From these figures, the better performance for larger PDU size is observed as data size increases.

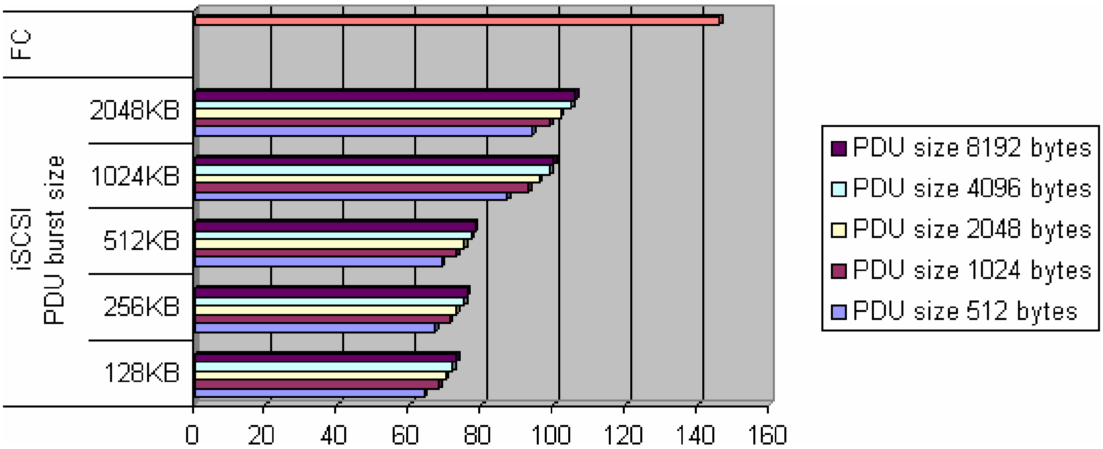


Figure 5.4: Network link utilization for write operations

While PDU size is increased from 512 bytes to 8192 bytes for the same PDU burst size, the average improvement on throughput is about 13%. When PDU burst size and PDU size are increased from 128KB and 512 bytes to 2048KB and 8192 bytes, the maximum improvement on throughput is about 65% for write operations.

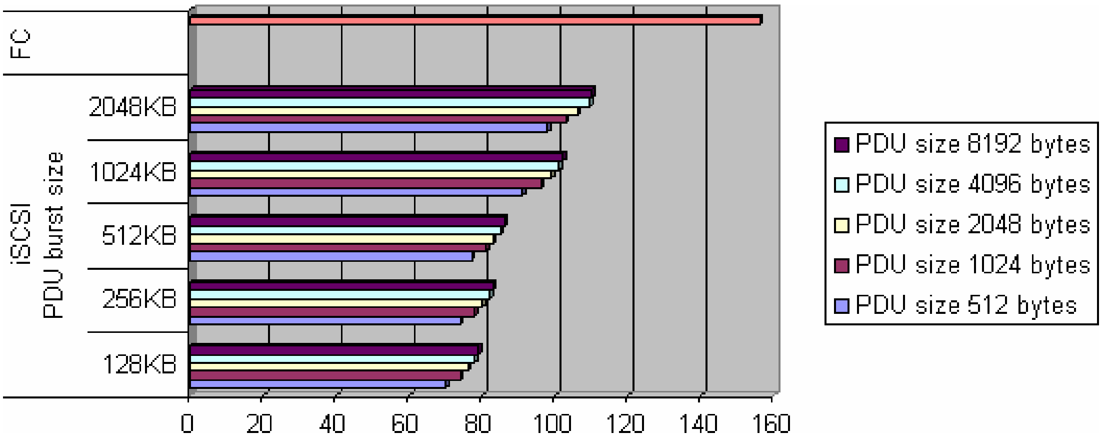


Figure 5.5: Network link utilization for read operations

While PDU size is increased from 512 bytes to 8192 bytes for the same PDU burst size, the average improvement on throughput is about 12%. When PDU burst size and PDU size are increased from 128KB and 512 bytes to 2048KB and 8192 bytes, the maximum improvement on throughput is about 57% for read operations

5.4.2. The Effect of the Network Parameters

In this section, how the network parameters like TCP window size, MTU (Maximum Transmission Unit) and link delays affect the iSCSI performance is investigated. The effect of MTU size is first examined. In the test setting, the TCP window assumes the default value of 20. The links are a Gigabit Ethernet with delay 20us, and a 1Gbps Metro Ethernet with delay 0.5ms. The MTU sizes are 296B, 576B and 1500 bytes respectively.

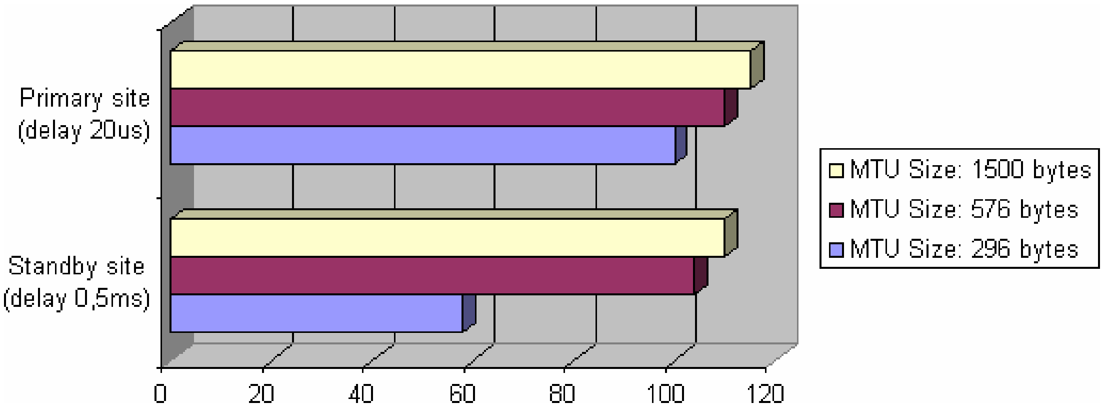


Figure 5.6: MTU size and Link Delay

Figure 5.6 shows the achieved throughput with varying MTU sizes. Normally, for a given delay and MTU size, the maximum throughput that can be achieved is approximately one window per round trip time, i.e. (MTU * window)/2*delay, which implies that throughput is inversely proportional to the link delay for the given MTU. This figure shows that the throughput decreases quickly for smaller MTU sizes, whereas higher MTUs show a gradual decrease even for higher link delays. For link delays equal to 20us, the MTU size does not have much effect on the throughput this is because at short network link delay, bandwidth-delay product is small. The acknowledgement comes back very fast. As the link latency increase, the throughput drops gradually, thus the link utilization is also getting lower. Papers [33, 35] also represent the similar results for the relationship between the link latency and the throughput.

When the MTU size is increased from 296 bytes to 1500 bytes, the improvement on throughput is about 15% in the primary site which link delay is smaller. As the MTU size is increased from 296 bytes to 1500 bytes, the improvement on throughput is about 89% in the secondary site which link delay is higher. This results show that MTU size is much more important in the environment that has higher link latency.

How the TCP window variation affects the throughput is examined. Three different TCP window sizes (20, 40 and 80) are used. The MTU size is fixed at 296 bytes. Figure 5.7 shows that the throughput increases with the increase of TCP window for the standby site which has higher link latency. At the short network link latency, e.g. Gigabit Ethernet LAN environment, the throughput is the same for all TCP windows sizes. However, with the increase of network latency, e.g. Metro Ethernet WAN environment, when the window size is too small, the throughput reduces significantly.

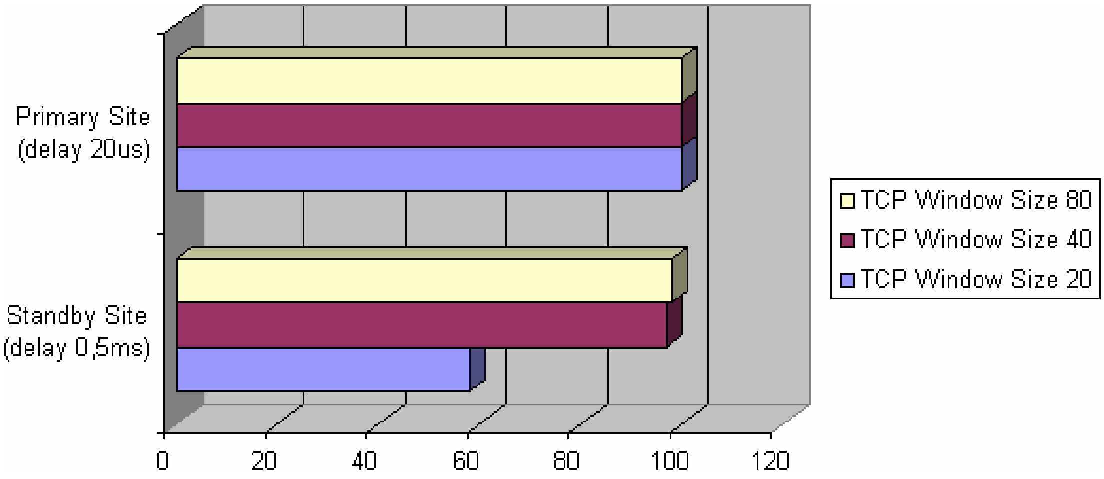


Figure 5.7: TCP window size and Delay

At the higher link delay, which is 0.5ms in the standby site, as TCP windows size is increased from 20 to 80, the improvement on throughput is about 68%.

5.4.3. The Effect of the CPU Utilization

In addition to effects of iSCSI and TCP parameters, how the CPU usage affects the iSCSI performance is investigated in this section. The following single line PERL script was used in order to create a CPU load that stops after n-seconds:

```
perl -e '$SIG{ALRM}=sub {exit}; alarm n; 1 while {}'
```

After running above perl script, Process Resource Manager (PRM) was used in order to cap CPU usage. PRM is a resource management tool used to control the amount of resources that processes use during peak system load.

Figure 5.8 and figure 5-9 show when the processor had any other tasks to perform, the throughput of the iSCSI decreased. After creating CPU loads, throughputs of iSCSI with different PDU burst sizes were gauged. The results were expected because the higher CPU usage was resulted in the lower throughput, because of that the performance of iSCSI depends on the power of processor.

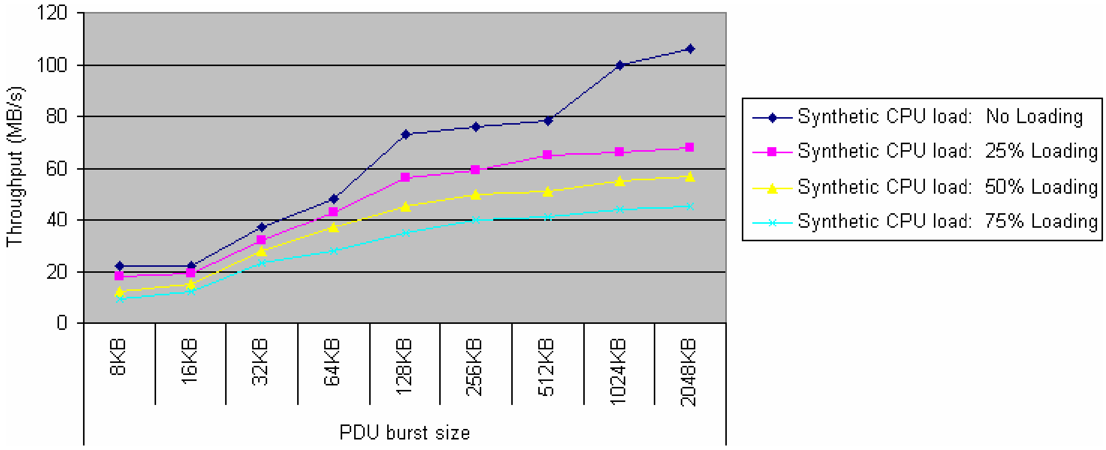


Figure 5.8: Effects of CPU load on write operations

At the 25% CPU load, %50 CPU load, and %75 CPU load, the maximum throughput decreases %35, %46, and %57, respectively for write operations.

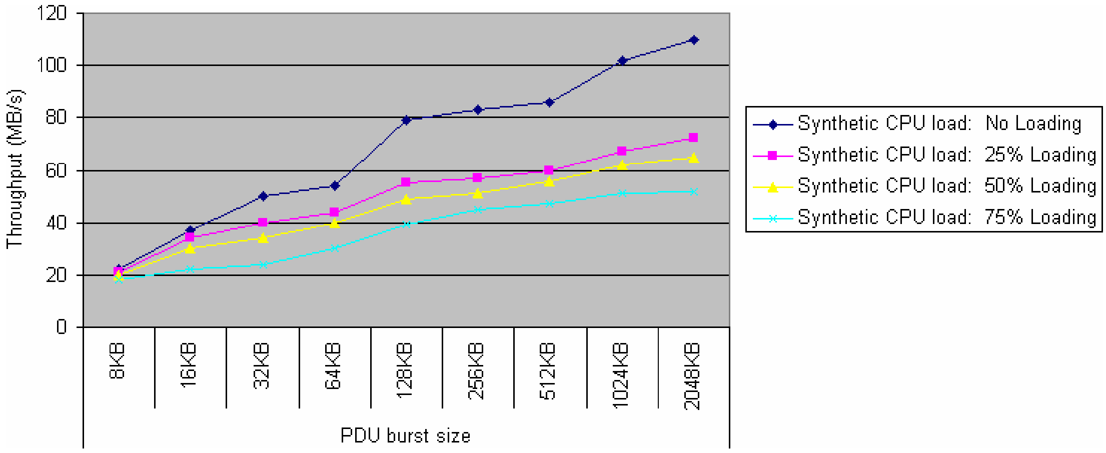


Figure 5.9: Effects of CPU load on read operations

At the 25% CPU load, %50 CPU load, and %75 CPU load, the maximum throughput decreases %34, %40, and %52, respectively for read operations.

5.4.4. The Effect of the PDU burst size on CPU Utilization

In this section, how PDU burst size affects the CPU utilization is examined. iSCSI in itself is fairly low CPU intensive software and even during heavy loads uses very little of the CPU power, but TCP/IP can consume noticeable CPU resources. Figure 5.10 shows when PDU burst size increases, the CPU utilization increases too. It may be assumed that the PDU burst size increases should decrease the amount of CPU cycles being used since there is less overhead with larger bursts. However, the overall size of the data has an important effect on CPU utilization. The larger data sets requires more TCP/IP packets to be sent and as such, more processing on the host CPU. While this thesis examines the effect of throughput on CPU utilization in view of PDU burst size, the one of the papers [36] presented at MSST'05 investigates the same effect in view of I/O request size. Both of them show pretty similar results to each other. The increased network throughput results in high network processing times including memory copies of data from the NIC.

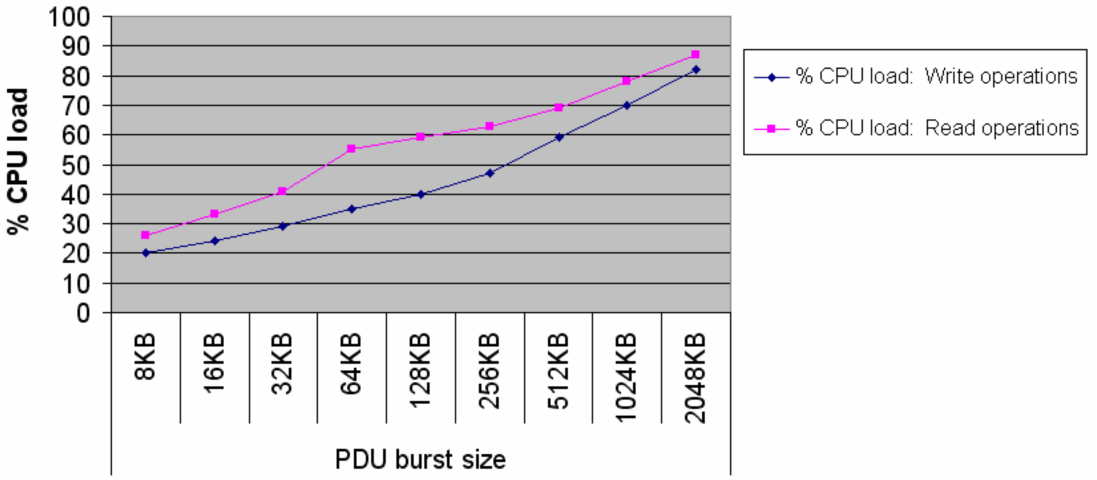


Figure 5.10: Effect of the PDU burst size on CPU Utilization

Although the nature of the measurements also has an effect, this is not the case. Since the utilization measurements were taken as a time average, the smaller bursts were written to memory almost immediately while the next datagram was traveling over the network. The result is that the CPU spends a relatively large amount of time idle, waiting for the next piece of data. With the larger bursts, the data is still being written to memory when the next piece of data arrives. Therefore, the CPU does not spend a lot of time waiting for a new chunk of data but is busy for most of the duration of the transaction.

Increasing PDU burst size from 8KB to 2048KB increases CPU usage 3.1 times and 2.2 times for the write and read operations, respectively.

6. CASE STUDIES

6.1 Best practices in real world

In order to illustrate both how these protocols can be used as complementary solutions and the different usage of them in disaster recovery mechanisms, three case studies are provided. While first of them is a cost effective remote backup solution, two of them are disaster recovery solutions.

6.2 Virtual tape library based remote data backup

There is 500km distance between the primary site and the secondary site. Each site has its own FC-SAN. A robotic tape library that has Fibre Channel interface is used for local backup operation in the primary site. Although the secondary site has a small set of all component of the primary site, it has not a tape library for data backup operations. Company wants to backup data in the primary site to the secondary site. In addition to lack of tape library in the secondary site, it is also impossible to establish any kind of Fibre cable between these sites due to insufficient service provided by Turk Telekom. There is a 34Mbps dedicated line between these sites. In both sites HP StorageWorks Enterprise Virtual Array (EVA) are used as storage box. By default EVA is not an iSCSI enabled disk array, but the EVA which is in the secondary site has “EVA iSCSI Connectivity Option”. EVA iSCSI Connectivity Option extends the capability of EVA to incorporate iSCSI servers without requiring additional storage arrays or management costs. The EVA iSCSI Connectivity Option also offers simultaneous iSCSI and Fibre Channel support.

In the storage box of the secondary site, sufficient amount of space is allocated to the database servers of primary site for two weeks of daily backup. This storage space which is in the storage box of the secondary site is presented to backup server in the primary site. HP Openview Dataprotector software is used as backup solution. Dataprotector can use any file system as a backup media like a real backup tape. This mechanism is called “virtual tape library”. This storage space is configured as virtual tape library in Dataprotector. Backup server in the primary site is configured as iSCSI enabled and connected to the storage box in the secondary site with existent 34Mbps line.

This remote backup solution is faster than traditional tape solutions because of being disk array based backup and cheaper than any Fibre Channel solution because of using existent network line for iSCSI connection without any additional investment.

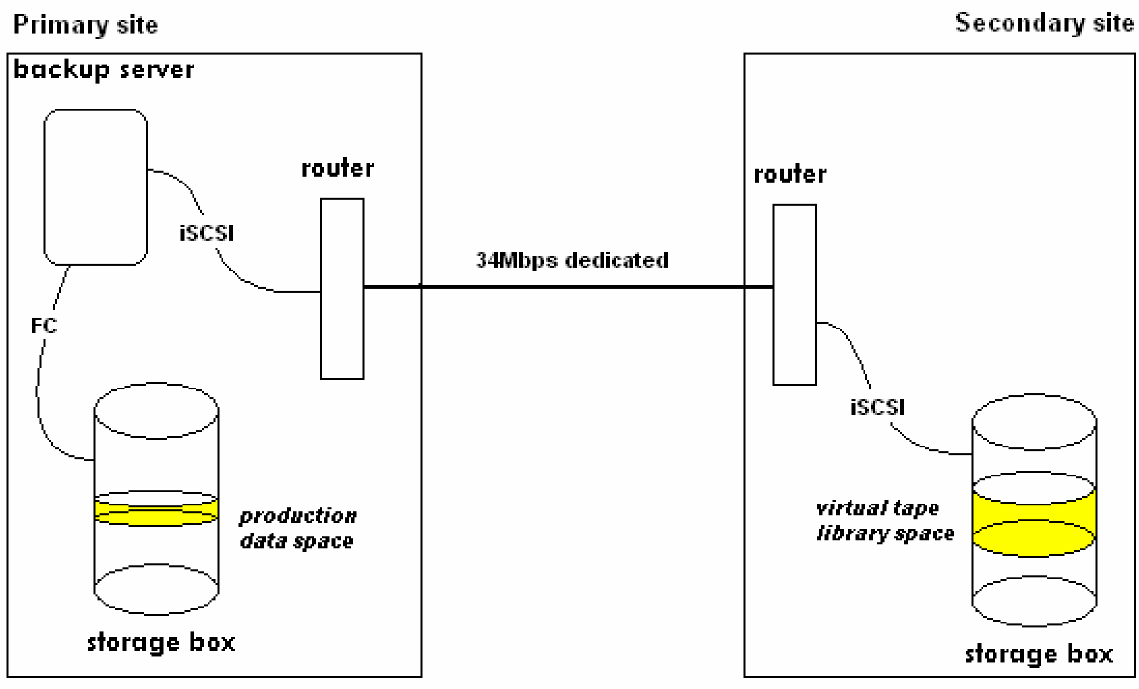


Figure 6.1: iSCSI-based remote data backup

6.3 FC-based SAN for real-time remote data replication

There is 200m distance between the primary site and the standby site. Each site has its own FC-SAN. The standby site consists of a database server and a storage box. In both sites HP StorageWorks Enterprise Virtual Array (EVA) with HP StorageWorks Continuous Access EVA Software (CA) license are used as storage box. Real-time synchronous data replication between the primary and the standby sites is wanted. CA is an array based application to create, manage and configure remote replication on entire EVA product family. CA copies configured disk space from primary storage box to standby storage box over FC connection without needing any server. When a writing request is coming from database server to primary storage box, the following steps are carried out:

- write data to primary storage
- copy data to standby storage
- write data to standby storage
- send commit to database server

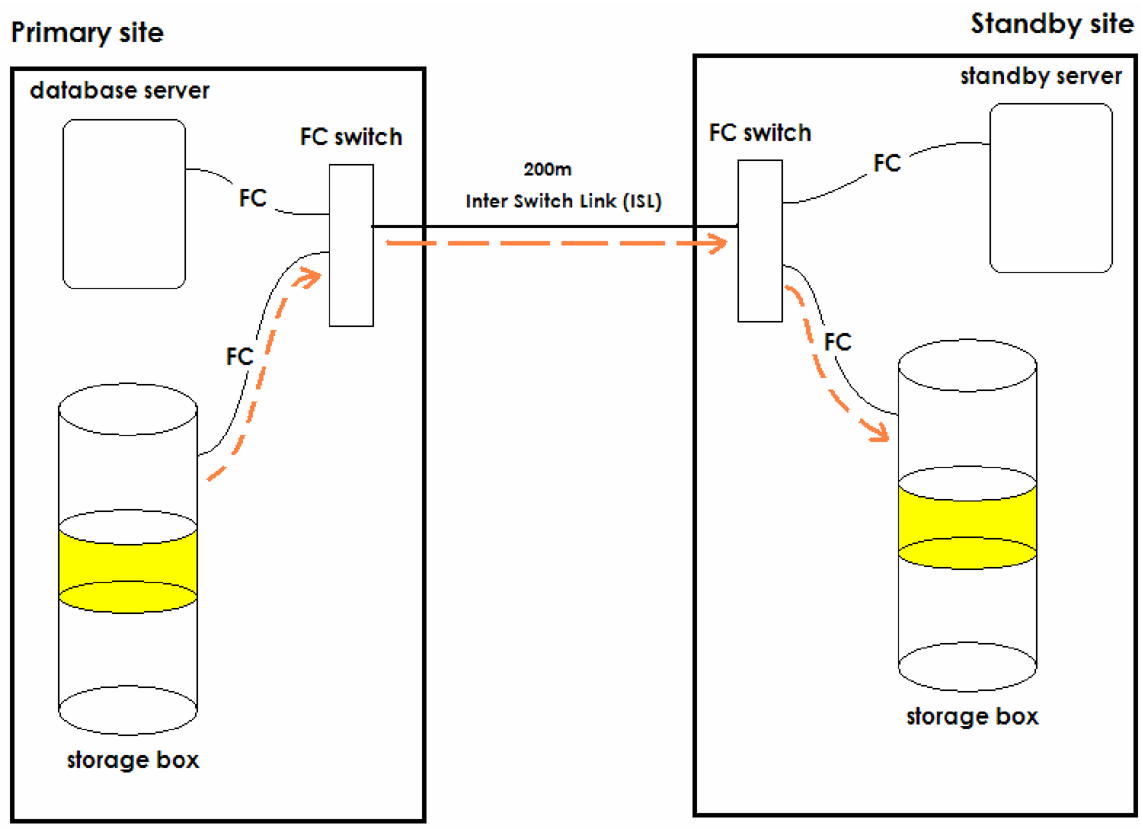


Figure 6.2: FC-SAN based remote data replication

The main advantages of CA are:

- The fastest data replication solution
- There is no server intervention

The main disadvantages of CA are:

- CA software license is expensive, and at least two licenses are required: one of them for primary site, one of them for standby site
- CA uses block level transfer. Even only one byte changes in a block, it is required to copy entire block. Common block sizes are 16MB and 32MB.
- FC-SAN is required for CA configuration.
- A downtime is needed for switching from one storage box to another in a disaster situation.

6.4 iSCSI-based SAN for real-time remote data replication

There is 10km distance between the primary site and the standby site. There is a 1Gbps Metro Ethernet line between these sites. Each site has its own FC-SAN.

Although CA is very fast and easy-to-use, it has its own drawbacks which are described above. It is an expensive solution because of requiring FC-SAN and double software license cost. It does not support automatic failover; it requires manual operations and downtime.

Mission critical businesses are required load balance and failover enabled IT infrastructure, and data storage system are one of the most important parts of this IT infrastructure. Although both remote backup and CA based remote data replication solutions described above are implemented and used for two years in my company, we need an additional data replication mechanism which ensures both failover and load balance between two separate data storage systems.

Our each Unix server has two internal disks which are configured as mirrored disks. The same data is written in both disk simultaneously in this configuration, therefore these two disks are seen as a single disk by operating system. When one of these disks fails, all disk operations are executed by the other disk without any interruption. All disk operations are spanned to both disks.

Can two data storage system be configured in the same manner? The answer is yes. In order to mirror two data storage system, basic logical volume mirroring can be used. Before describing this mechanism, I want to talk about basic Unix disk and file system terminology. Unix disk and file system hierarchy can be shown as a layered structure:

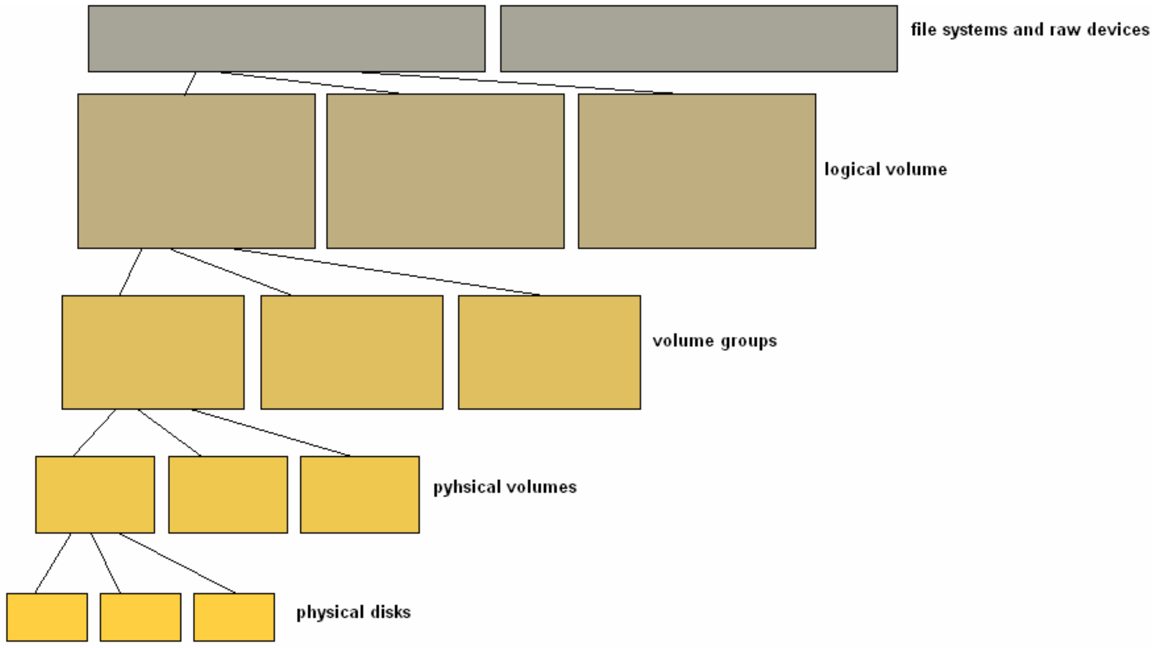


Figure 6.3: Unix disk and file system hierarchy

A physical volume consists of physical disks. At least one physical disk is required to form a physical volume. A volume group consists of physical volumes and at least one physical volume is required. A logical volume consists of volume groups and at least one volume group is required. A file system a raw device consists of logical volumes and at least one logical volume is required.

Logical volume can also be configured as mirrored logical volumes. Data in the new copies is synchronized. The synchronization process can be a little bit time consuming, depending on the amount of data and the performance of the server.

Briefly, I want to mention about HP Auto-Port Aggregation (APA). Two or more Ethernet ports can be configured as a single Ethernet port by using APA. A single IP is assigned to APA configured Ethernet port. The total throughput is the sum of all throughputs. APA configured Ethernet port supports failover and load balancing.

We assume that:

- One of the two logical volumes is in the primary data storage and the other is in the stand by data storage
- These logical volumes are configured as a mirrored logical volume
- Server is connected to the primary storage via FC, and is connected to the standby data storage via iSCSI
- FC ports is 2Gbps ports

- iSCSI ports is 2Gbps APA configured Ethernet ports.

In this configuration, data is written both data storage simultaneously, and therefore data is synchronized in both data storage. If the primary data storage fails for any reason, all disk operations is going to continue via the other data storage which is connected with iSCSI without any downtime, through the nature of mirrored logical volume.

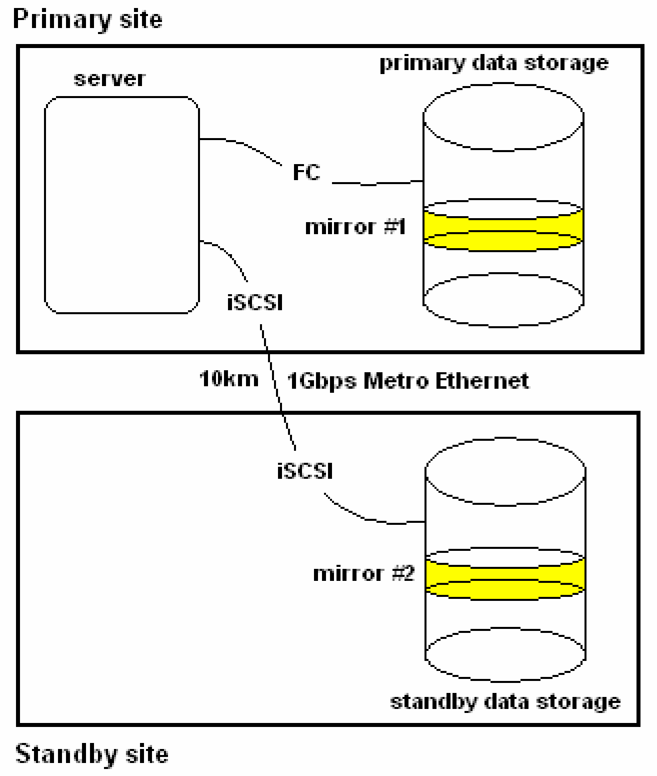


Figure 6.4: iSCSI-SAN based real-time remote data replication

7. CONCLUSION AND FUTURE WORK

In this thesis, it is overviewed that Business Continuity Planning and Disaster Recovery Planning are very important for any organization which has mission critical business. Remote data replication of SAN is the basis of data protection, and data protection is the basis of disaster recovery. If backup and recovery mechanisms were put in place by these businesses, they generally involved periodic data backups (perhaps every week) and even less frequent tape shipments to off-site storage warehouses. While the nature of these businesses may allow for this, it could still take days to recover from unplanned downtime resulting in loss of revenue. However, decreasing the RTO and RPO for these organizations has always been too expensive to consider. At the same time, large organizations and those businesses with considerable budgets (Class 3 and 4 environments) have invested heavily in expensive Fibre Channel hardware, resources, and trained personnel to enable fast data backup and recovery to remote locations. While these solutions perform exceptionally well, they are costly and any upgrades will no doubt be very expensive as well.

This thesis shows how iSCSI can be used as an alternative to expensive FC-based SAN solutions. In the Class 1 and 2 settings, iSCSI can be implemented easily. The only costs to these organizations are associated with the network infrastructure and servers/workstations. In most cases, this infrastructure already exists in some sense and thus the cost is reduced. Using Internet, these organizations can capture every write operation to a local disk and forward an encapsulated iSCSI packet to a remote location.

The case study that is described in chapter 6.4 shows that iSCSI is very promising as a disaster recovery medium for large organizations with high RTOs and RPOs. In order to maintain high recovery time and recovery point objectives, these organizations generally employed powerful but expensive Fibre Channel hardware. In these cases, performance cannot be sacrificed to save money. With Fibre Channel able to operate at 4 Gbps (and 10 Gbps in the future), it was the only solution that high performance environments could count on. iSCSI can function very well at Gigabit speeds and will be able to take advantage of the large throughput available once 10 GigE networks are available. In addition, iSCSI can already match the performance of Fibre Channel operating at 4 Gbps because of multiple connections per session (APA). Since APA allows a collection of separate physical links to be used as a single session, it is possible to provide link aggregation thereby matching the performance of 4 Gbps FC implementations. Furthermore, it would be interesting to see how APA could be used to provide for fault tolerance and load balancing.

For 10 GbE settings, a number of new challenges are presented concerning memory and CPU requirements. A thorough investigation of these hardware requirements for 10 Gbps iSCSI must be undertaken to shed light on future solutions that will allow iSCSI to perform at increased throughputs.

This thesis shows how the iSCSI parameters and the underlying TCP parameters affect the iSCSI performance. Chapter 5 shows that with larger burst data size, the larger PDU size will help the throughput. Increasing TCP window size and MTU also affects the end-to-end performance. But this effect is more pronounced for higher link delays. The further study for the iSCSI storage system in a diverse network such as fat network (long latency, high bandwidth), wireless network may be done.

Although iSCSI has numerous advantages, such as the usage of commodity IP-based infrastructure and thus, reduced cost and management effort, there are numerous questions associated with the impact of iSCSI on system performance. In this thesis, it is also tried to show that using iSCSI without increasing system resources, e.g. number of CPUs, a mount of memory has a significant impact in all applications running in the server.

REFERENCES

[1] **Myers, J.**, 2003. Effective and Efficient Disaster Recovery Planning, *Disaster Recovery Journal*, Fall 2003 Vol.16, Issue.4

[2] **Smith**, **D. J.,** 2001. Business Continuity Management: A Good Practice Guideline, *International Journal of Business Continuity Management*, Vol.2, Issue.1.

[3] **Gartner, Inc.**, 2001.*USA Today,* October 2002.

[4] **Marcella, A. J., Stucki, C.,** February 2004. Business Continuity, Disaster Recovery, & Incident Management Planning: a Resource for Ensuring Ongoing Enterprise Operations, The Institute of Internal Auditors Research Foundation.

[5] **Sarrel, M. D.**, 2006. Building a Disaster Recovery Plan, *PC Magazine*, January 15.

[6] **Greer**, **G.,** May 13, 2002. Business Continuity: How the Events of September 11 are affecting business plans for disaster recovery and business resumption, *Paper*, Management of Information Systems, Baylor University, Texas.

[7] **Fontana, J.**, 2001. Alternate Site Recovery Techniques, *White Paper*, Network World, November 5.

[8] **Juran, I., and Merryman, J.**, November 2002. Disaster Recovery Strategies with Tivoli Storage Management Redbooks Second Edition, IBM, San Jose, CA, USA.

[9] **Department of Information Systems**, 2002. Draft Interagency White Paper on Sound Practices to Strengthen the Resilience of the IS Financial System, *White Paper*, Securities and Exchange Commission Release No. 34-46432; File No. S7-32-02.

[10] **Weems, T.**, 2003. How far is Far Enough, *Disaster Recovery Journal*, Spring 2003, Vol.16, Issue.1.

[11] **Evan, M., and Stern, H.**, 2003. Blueprints for High Availability (second edition), Wiley Publishing, Indianapolis, IN, USA.

[12] **Mayer, P.**, October 2003, Data Recovery: Choosing the Right Technologies, *White Paper*, Datalink.

[13] **Brocade, Inc.**, March 2001. Storage Area Networks: An Innovative Approach to Backup and Recovery, *White Paper*, Brocade Communications Systems, Inc., San Jose.

[14] **Tate, J., Lucchese, F., and Moore, R.**, 2003. Introduction to Storage Area Networks, *Red Books*, IBM.

[15] **Schiattarella, E.**, 2005. Performance Analysis of Storage Area Network Switches, *IEEE Workshop on High Performance Switching and Routing*, Hong Kong, 12-14 May 2005.

[16] **Meggyesi, Z.**, August 1994. Fibre Channel Overview, *High Speed Interconnect Pages*, CERN, Budapest, Hungary.

[17] **Kraft, C., 1994**. Design and Implementation of iFCP, *Master Thesis*, University of Colorado, USA.

[18] **Wallace, D., and Ham, B.**, 2001. ANSI NCITS Project 1235-D, Fibre Channel Physical Interface (FC-PI), *Technical Report*, rev.13, December 9.

[19] **Satran, J.**, January 2003. iSCSI Specification, *IETF Draft*, draft-ietf-ips-iscsi-20.txt.

[20] **Meth, K. Z., and Satran, J.**, Apr. 2003. Design of the iSCSI Protocol, *Twentieth IEEE/Eleventh NASA Goddard Conference on Mass Storage Systems & Technologies*, Paradise Point Resort, San Diego, California, USA, April 7-10, 2003.

[21] **Hufferd, J. L.**, 2003. iSCSI: The Universal Storage Connection, Addison-Wesley, New Jersey.

[22] **Kembel, R. W.**, 2004. Transitioning to IP Storage Networks and iSCSI: Using iFCP, *SNIA PowerPoint slide presentation*, SNIA.

[23] **Mearian, L.**, 2003. SAN after Fibre Channel, *Technical Report*, SNIA.

[24] **Clark, D. D., and Tennenhouse, D. L.**, September 1990. Architectural considerations for a new generation of protocols, *ACM Symposium on Communications Architectures and Protocols*, September 1990.

[25] **Savyak, V.**, 2004. iSCSI Review, available at http://www.digit-life.com/articles2/iscsi/

[26] **Tang, S., Lu, Y., and Du, D.**, 2002. Performance Study of Software-Based iSCSI Security, *IEEE Security in Storage Workshop 2002*, Greenbelt, Maryland, USA, December 11, 2002, 70-79.

[27] **Dayal, R.**, March 2004. iSCSI- Internet Small Computer System Interface, *White Paper Technology Review*, Uttar Pradesh, India.

[28] **Clark, T.,** December 15, 2001. IP SANS: An Introduction to iSCSI, iFCP, and FCIP Protocols for Storage Area Networks, Addison-Wesley Professional.

[29] **Rajagopal, M., Rodriguez, E., and Weber, R.**, July 2004. Fibre Channel Over TCP/IP (FCIP), *IETF RFC*, available at http://www.ietf.org/rfc/rfc3821.txt

[30] **Voruganti, K., and Sarkar, P.**, April 2001. An Analysis of Three Gigabit Networking Protocols for Storage Area Networks, *20th IEEE International Performance, Computing, and Communications Conference*, Embassy Suites Phoenix North, Phoenix, Arizona, USA, April 4-6, 2001.

[31] **McCanne, S., and Floyd, S.**, 2002. The Network Simulator (NS-2), USC Information Sciences Institute, available at http://www.isi.edu/nsnam/ns

[32] **Molero, X., Silla, F., Santonja, V., and Duato, J.**, September 2000. Modeling and Simulation of Storage Area Networks, *The 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, San Francisco, California, USA, 29 August - 1 September 2000, pp. 307.

[33] **Aiken, S., Grunwald, D., and Pleszkun, A.**, Apr. 2003. Performance Analysis of iSCSI protocol, *Twentieth IEEE/Eleventh NASA Goddard Conference on Mass Storage Systems & Technologies*, Paradise Point Resort, San Diego, California, USA, April 7-10, 2003.

[34] **Lu, Y., and Du, D.**, August 2003. Performance Evaluation of iSCSI-based Storage Subsystem, *IEEE Communication Magazine*, August 2003.

[35] **Teck, W., Hillyer, B., Shriver, E., Gabber, E., and Ozden, B.**, June 2002. Obtaining High Performance for Storage Outsourcing, *Joint International Conference on Measurement and Modeling of Computer Systems*, Marina Del Rey, California, USA, June 15-19, 2002.

[36] **Xinidis, D., Bilas, A., and Flouris, M. D.**, April 2005. Performance Evaluation of Commodity iSCSI-based Storage Systems, *22nd IEEE/13th NASA Goddard Conference on Mass Storage Systems & Technologies*, Monterey, CA, USA, April 2005.

CURRICULUM VITAE

Hakan Arıbaş was born in Erzurum, on the 19th of March, 1973. He received his B.Sc. degree from Istanbul Technical University, Electronics and Communication Engineering Department in 1995. Upon graduation, he was employed as a Software Design Engineer at NETAŞ for two years. After two years of working at Pamukbank as Systems Administrator, he had been doing military service at Turkish Navy Headquarters as Military IT Officer. He had been working at Oyak Teknoloji as Unix Systems Administrator for two years. He has been working at Merkezi Kayıt Kuruluşu A.Ş. as Project Manager since August 2003. He has been married since October 2002, and has a child. His areas of interest and expertise include but not limited to: Unix systems, storage area networks, project management.