# CHAPTER 2

# LITERATURE REVIEW

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 INTRODUCTION

There are numerous reasons to plan for the inevitable disasters that are faced by businesses, industries, and government at the level of computer system. However, the main objective of a disaster recovery planning is to protect people or organization assets, typically data, software and hardware. Therefore, the level of protection must include physical threats, natural disasters, internal vulnerabilities, personal misuse, unauthorized users and programming errors.

According to current estimates, data stored on networks is growing at the rate of more than 40% per year. [4]Therefore, there will be a need for effective disaster recovery planning in order to protect this large amount of data from natural disaster, from unauthorized intrusion, from computer system and human error. According to a 3M study, 60% of data loss is due to user error, such as accidentally deleting or overwriting or not saving the files. Power problems, application errors, hardware problems and disk crashes account for much of the rest.[5]

Hence, the concept of disaster recovery planning is reviewed in order to design and propose the best data backup strategy for a network environment. The various types of local and remote data backup strategies are also discussed in this chapter. Furthermore, the perspective and model of File Transfer Protocol are being studied in order to supplement the knowledge on remote data backup using the FTP. Eventually, a synthesis of the local and remote data backup system will be developed.

## 2.2 OVERVIEW OF DISASTER RECOVERY PLANNING

An effective disaster recovery planning can be viewed as a "three-legged stool." Each component of the stool serves a specific purpose to ensure a solid support resting on a solid foundation. [7] In other words, a disaster recovery planning consists of three components with notation given as off-site storage, users and backup site, and the absence of any of these components or communications between them may impair the whole planning.

### 2.2.1 OFF-SITE STORAGE

Locations that won't be affected by the disaster or far away from the location of disaster may used as places to store vital data files, critical applications and records. These backed up resources in the off-site storage are protected from hazards and vulnerabilities of an unexpected incident. Consequently, the restoration of data may be accomplished in the event of catastrophic disaster and it requires the latest version of a software or the most recently generated data files.

### 2.2.2 BACKUP SITE

In order to maintain the continuity of operation, backup alternative strategies must be devised based on the analysis of the applied system. Maximum allowable downtime and critical applications based on the policy are the examples of criteria that must be considered.

### 2.2.3 USERS

Disaster recovery planning is created to protect users and assets. Thus, the development of disaster recovery planning needs the participation of users to provide the requirements

of planning and maintenance. For instance, as a system grows, the requirements of the user community enlarge. Hence, user communications are subject to continuous modification and expansion of the planning.

### 2.2.4 THE NEED FOR RISK MANAGEMENT IN DISASTER RECOVERY PLANNING

In the disaster recovery planning, risk management sits an important role to determine the existing levels of risk to the assets and to identify the strategies that could reduce the threat. The need of disaster recovery planning will be discussed in Chapter 3.

## 2.3 STRATEGIES OF DISASTER RECOVERY

Recovery strategy means the interim ability to process data while a full recovery of the primary computer site is underway. [7] Hence, with the aim of setting up a dependable system, the strategies of disaster recovery must be configured and analyzed accurately. Following are the lists of available recovery option that provide fault tolerance:

### 2.3.1 REMOTE MIRRORING

Remote Mirroring is storage based disaster recovery solution that provides the ability to copy data in real time to a remote destination. There are two systems with notation given as primary and secondary system. If the primary system fails, the secondary system can keep the system operating without loss of data or interruption to the user. Thus, this is a fault tolerant approach that has an essential role in providing availability especially in transaction processes. Taking into account the drawback in redundancy, remote mirroring usually applied in critical transaction to handle data that has been identified as critical.

### 2.3.2 DISK CHANNEL DUPLEXING

As an extension to disk mirroring, this scenario provides more redundancy by duplexing a second channel into a single server. Thus, it protects against bus channel failure.

### 2.3.3 STANDBY DATABASE

To eliminate time used to restore data after disaster, there is an option to pre-load a secondary machine with the Operating System, Application and Database. A Standby Database can be arranged such that a backup copy of the database can be loaded and applied to the secondary site at a predetermined interval.

Furthermore, this option can work along with the Database journaling software running on a dedicated machine. Journaling means that the application writes database changes and updates to a separate file. This file can be used to rebuild the database if it somehow becomes corrupted or otherwise damaged.

### 2.3.4 STANDBY APPLICATION

Changes in application are less frequent than changes in database. Therefore, the copy at the secondary site can be refreshed less often, typically weekly, monthly or quarterly based on the estimated frequency of changes in the applications. As in the case of standby database, applications may be pre-loaded into a dedicated machine to be used during disaster.

### 2.3.5 STANDBY OPERATING SYSTEM

Restoration of Operating System is the initial step in disaster recovery and must be accomplished in an efficient and timely manner. Therefore, the preparation of preloaded disk with the Operating System is one of the solutions.

### 2.3.6 STANDBY SHARED MACHINE

Shared machine is on standby ready to be used for recovery during disaster.

### 2.3.7 DISPLACEMENT

This option requires top identification of displacement equipment. Hence, it must match the damaged equipment and must not be time critical. During a disaster, the less critical function will be displaced by the highly time critical function, so that the equipment and facilities of the displaced function could be utilized for recovery.

### 2.3.8 REPLACEMENT

This requires that the predetermined equipment is replaced at the recovery site within a specific time for recovery purpose.

There are several criteria that are normally used in the decision making process of choosing a suitable recovery strategy. These variables are generally the same for all organizations, however, the weighting for each criteria may differ. The criteria are as depicted in Table 2.1.

Table 2.1: Identification of Decision Variables

| | |
|---|---|
| Cost | Cost of implementing the option |
| Availability | Account in the accessibility of the option |
| System Compatibility | Determine the level of compatibility of the option |
| Testing | The convenience level of testing using this option |
| Data Communication | The speed of communication that can be established using this option |
| Support | Determine level of assistance can be expected if this option is used |

## 2.4 OVERVIEW OF LOCAL AND REMOTE DATA BACKUP SYSTEM

Backup system is one of the approaches of disaster recovery planning. They are interdependent with off-site storage and users plans and procedures.

### 2.4.1 DATA STORAGE DEVICES

There are many backup technologies currently in use. It varies from simply copying data to tape cassette or other storage devices to an effectively online system redundancy. The storage environment comes in a hierarchy of technologies. Technologies with the best performance come at the highest cost. Magnetic tape storage, magnetic disk drives and magneto-optical devices are commonly implemented in the backup system due to their large capabilities.

#### 2.4.1.1 MAGNETIC TAPE STORAGE

Magnetic tape stands at the lower end of the storage technology spectrum in terms of performance--it is relatively slow and has a correspondingly low media cost for high capacity. [8]As a serial-access media, tape is not as efficient as magnetic and optical storage that allows random access of information on the media. However, this may be acceptable for backup and restore operations, which read and write data files sequentially.

Using tape backup system is fast and straightforward, and suits well in either the independent mainframe or the openness network environment. In both environments, a tape server can connect directly to the network so that the backup processing is either run by network administrator or by every registered user.

### 2.4.1.2 MAGNETO-OPTICAL STORAGE

This technology offers a lower media cost and maintain an adequate performance in data storage. Using optical storage devices have the full advantages of storing the less frequently used data by transferring it from the expensive magnetic storage to cheaper optical disc. Hence, it can provide a more cost-effective storage and backup environment. Since optical medium is a random access unit, which allows direct access to a specific file, implementation of optical backup system will ease the data retrieval process.

While optical media itself is relatively cheap, its use may involve some fairly expensive hardware.[8] This hardware is either write-once read-many (WORM) or erasable optical devices that permits unlimited access. This may include *CD-Recordable* (CD-R) or *CD-Rewritable* (CD-RW).

### 2.4.1.3 MAGNETIC DISK DRIVE

The fastest and most expensive storage technology is magnetic disk drive. It is the main storage device used in every computer system. Due to a high demand, the capacities of magnetic drives have historically risen faster than their costs. Consequently, making this technology affordable for an increasingly larger portion of the information storage paradigm. Removable magnetic ZIP disk is one of the technologies that provide for a cost- effective, flexible and easily deployable storage solution.

### 2.4.1.4 REDUNDANT ARRAYS OF INEXPENSIVE DISKS (RAID)

The basic idea of RAID is to combine multiple small, inexpensive disk drives into an array of disk drives, which yields performance exceeding that of a Single Large Expensive Drive (SLED).[6] Additionally, this array of drives appears as a single logical storage unit or drive.

As a fault tolerance process, RAID consists of five types of array architectures, RAID-1 through RAID5. Besides these five architectures, RAID-0 array is referred as a non-redundant array of disk drives. These levels are summarized in Table 2.2. [9]
"Striping" is the fundamental of RAID technology to concatenate multiple drives into a logical storage unit. Each drive's storage space is partitioned into stripes that are various of sizes. These stripes are then interleaved in a round-robin manner. As a result, the combined space is composed alternately of stripes from each drive.

Table 2.2 Levels of Redundancy in RAID Technology [9]

| Level | Redundancy Description |
|---|---|
| 0 | Data is "stripped" across all drives in the array. It is the fastest and most efficient array type but offers no fault-tolerance. |
| 1 | Used of disk mirroring and critical data will be written to two disks. |
| 2 | Distribution of data based on bit-by-bit basic. Bits in each byte are spread across the number of disks in the array. |
| 3 | Addressing, parity and error correction information tables is provided in one drive, while the others perform data stripping. Can be used in data intensive or single –user environments, which access long sequential records to speed up data transfer. |
| 4 | As in the case of RAID-3, addressing, parity and error correction information tables is provided in one drive, while the others perform blocks stripping. RAID-4 offers no advantages over RAID-5. |
| 5 | Blocks of data are striped across all drives in the array. Parity and error correction are distributed across all drives. Best choice in multi-user environments, which are not written performance sensitive. |

Storage systems have an important role in preventing data loss, offering adequate capacity that can easily scale as storage needs grow. Nevertheless, it is always utilized in parallel with disaster recovery strategies to optimize the backup planning so that they provide fast access to data without interruptions and are always prepared for equipment failures.

### 2.4.2 DATA BACKUP STRATEGIES

Disaster recovery planning is not completed without taking into account data backup strategies. These strategies include analysis of backup cycles, off-site storage, documentation of backup procedures and testing on the backup system. [9]

#### 2.4.2.1 BACKUP DOCUMENTATION, TESTING AND SCHEDULES

Documentation of backup and restore procedure should be managed to avoid human error. The documentation may include the time, date and backup progression. It should be written in short and precise covering the basic procedures used, so that non-technical person can understand it.

Testing of backup system in a regular and scheduled basis should be performed to maintain and amend system. Data that has been backed up must perform a restoration testing procedure to ensure that it is same as the originals. Thus, every program files that have been restored must executable; all data files can be opened by their application. All of the testing procedures are advised to perform on dedicated, aside area on a file server.

Backup cycles define the approach of backup procedures and can be generally classified as full daily backups, incremental, differential, selective backups. Identification of the backup cycles is based on the time and backup media efficiency. These are illustrated in Table 2.3.

Table 2.3 Backup Cycles[9]

| **Backup Cycles** | **Weekly Backup** | **Daily Backup** | **Restoration Procedures** | **Drawbacks** |
|---|---|---|---|---|
| Full Daily Backups | -NA- | All files | Perform full restoration | Use time and backup media inefficiently |
| Incremental Backups | All files | Only those files that have changed since the last backup | Full restoration requires restoration from the last full backup and follows by each partial backup in order. | 1. More complex process required.<br>2. Files deleted since the last full backup will be restored. |
| Differential Backups | All files | Only those files that have changed since the full backup | Full restoration requires restoration from the last full backup and follows by each partial backup in order. | More complex process required. |
| Selective Backups | -NA- | Only specified files | Only specified files | Requires an tightly controlled environment to identify change of data |

-NA- = Not Applicable

### 2.4.2.2 LOCAL DATA BACKUP SYSTEM

In a local area network, backing up multiple workstation drives is time consuming and requires the cooperation of users. Consequently, it is advisable to keep data and files on file servers, which are always backed up properly. Typically, there are four common ways to connect these backup servers to the network:

#### 2.4.2.2.1 WORKSTATION-BASED BACKUP

The backup storage device is attached to a workstation. The software in that workstation runs backup and restore processes. Thus, backups require a user to be logged in to a workstation, and may present security risk.

#### 2.4.2.2.2 FILE-SERVER-BASED BACKUP

The backup storage device is attached to a file server. Backup and restore processes executed by the software in the file server, act as a server-based application. Thus, scheduled backups do not require a user to be logged in at a workstation.

#### 2.4.2.2.3 DEDICATED BACKUP SERVER

The backup system is independent from the network being backed up. Thus, it is facilitated by its own CPU and acts as an isolated and dedicated unit. This approach provides similar security to the file-server-based systems.

#### 2.4.2.2.4 HOST-BASED BACKUP

Host-based system use mainframe or midrange computer resource to back up data.

### 2.4.2.2.5 COMPARISON OF LOCAL DATA BACKUP OPTIONS

Table 2.4 lists down the advantages and disadvantages of various local data backup options that have discussed.

Table 2.4 Advantages and Disadvantages of Various Local Data Backup Options[9]

| | **Advantages** | **Disadvantages** |
|---|---|---|
| **Workstation-Based Backup** | Simplicity in installation of backup system. | Weak in performance and security.<br>Server backup speed in limited by the speed of network. |
| **File-Server-Based Backup** | Faster in speed of backup and restore processes.<br>Provides higher level of security | Centralized management system.<br>Affect the performance of network during the failure of backup system. |
| **Dedicated Backup Server** | Performance of network is independent from the failure of backup system. | Performance of backup server limited by the speed of network. |
| **Host-Based Backup** | Provides high security level centralized backup and management. | Costs is significantly higher |

### 2.4.2.3 REMOTE DATA BACKUP SYSTEM

Natural disaster as an inevitable risk should be taken into account in disaster recovery planning. Backing up data remotely or storing backup media off-site is an absolute necessity to recover after a major site disaster, such as a fire or an earthquake. One of the approaches to perform remote data backup is by transferring data over a dedicated line to a remote server. Contacting an off-site storage company, which will pick up storage media in schedule and placed in a secure environment, can be also performed. Hence to avoid regional disaster, several options of remote data backup are provided:

#### 2.4.2.3.1 OFF SITE STORAGE VENDOR

The storage media is picked up and stored in a quality facility designed to protect the integrity of the media. They are returned by a professional organization and are easily accessible for disaster recovery purposes.

#### 2.4.2.3.2 COMPANY INTERNAL OFF SITE STORAGE

The premise behind this option is similar to the off site storage vendor method mentioned in the previous section. The difference is that the company manages its own pickup and delivery of backup tapes and manages its own off site storage facility.

#### 2.4.2.3.3 BACKUP OVER THE WAN TO THE MAIN SERVER

The remote server automatically sends the data over a dedicated line to the main server where it is stored in a scheduled backup. When the main server is backed up, the remote site data is included in the backup. In the event of a disaster at the remote site, the data can be quickly retrieved from the main server over the WAN.

#### 2.4.2.3.4 ELECTRONIC VAULTING

Electronic Vaulting is one of the newer methods of network backup. This method allows a company to electronically transmit data to a remote server through remote storage shadowing or mirroring. This process is similar to backing up over the WAN in that the backups are transmitted over a dedicated line. In the event of a disaster, the data can be recovered from the remote site via a dedicated line.

### 2.4.2.3.5 COMPARISON OF REMOTE DATA BACKUP OPTIONS

Table 2.5 shows the comparisons of various remote data backup options that have discussed.

Table 2.5: Benefits and Drawbacks of Remote Data Backup Options[10]

| | Benefits | Drawbacks |
|---|---|---|
| **Off Site Storage Vendor** | Highly effective method of storing critical data backups, protect integrity of media and ease in accessibility. | Costs of transportation are the main issue because it needs a high frequent pickup of small volume of backup tapes. |
| **Company Internal Off Site Storage** | The benefits are similar to the off site storage vendor. | Costs of maintaining this service are extremely high, these include tape transportation, facility upkeep and associated labor costs. |
| **Backup over the WAN to the Main Server** | Centralized data backup with extremely disciplined and controlled method. Requires little human management. | Requires secure lines that can provide substantial bandwidth. Major configuration is needed to set up the WAN to transmit, receive, and schedule the backups. |
| **Electronic Vaulting** | The benefits are similar to the previous option. | Great dependence on the type of connection that the LAN has to the remote site, in the case of bandwidth. |

## 2.5 REVIEW ON FILE TRANSFER PROTOCOL

### 2.5.1 PERSPECTIVE

The File Transfer Protocol (FTP) is defined by RFC 959 published in 1985. It provides facilities for transferring to and from remote computer systems. Usually the user transferring a file needs authority to login and access files on the remote system. The common facility known as anonymous FTP actually works via a special type of public guest account implemented on the remote system.

Objectives of FTP are to: [42]

1. promote sharing of files, computer programs and data,
2. encourage indirect or implicit use of remote computers,
3. shield a user from variations in file storage systems among hosts,
4. transfer data reliably and efficiently

### 2.5.2 FTP MODEL

An FTP session normally involves the interaction of five software elements. There are[43]:

1. User Interface:

   This provides a user interface and drives the client protocol interpreter.
2. Client Protocol Interpreter (PI):

   This is the client protocol interpreter. It issues commands to the remote server protocol interpreter and it also drives the client data transfer process.
3. Server PI:

   This is the server protocol interpreter, which responds to commands issued by the client protocol interpreter and drives the server data transfer process

4. Client Data-Transfer Process (DTP):

   This is the client data transfer process responsible for communicating with the server data transfer process and the local file system

5. Server DTP:

   This is the server data transfer process responsible for communicating with the client data transfer process and the remote file system.

During an FTP session, two separate network connections will be established between the PIs (control connection) and between the DTPs (data connection). These connections use TCP.

The well-known port for FTP is number 21, hence, the FTP server will listen on this port for control connection request. The client then will send a control message, along with the port number on which the client is prepared to accept an incoming data connection request.

The advantages of utilizing separate connections for control and data are to:

1. Maintain different appropriate qualities of service among the two connections,
2. Avoid problems of providing escape and transparency for commands embedded within the data stream.

Clients always initiate on file transfer setup, however either the client or the server may be the sender of data. Besides, data transfer mechanism is also used for transferring directory listings from server to client.

### 2.5.3 FTP SERVICE COMMANDS

The FTP service commands define the file-transfer or the file system function requested by the user. There are transmitted as ASCII strings starting with three or four upper case ASCII characters followed by optional arguments and a CR/LF pair at the end of the

command.[42] Table 2.6 shows a list of all commands. The ones marked with an asterisk are rare and rarely implemented.

Table 2.6 List of FTP Service Commands [43]

| String | Meaning |
|---|---|
| ABOR | Abort transfer. |
| *ACCT | Some systems associate both accounts and users with file system objects. |
| *ALLO | Allocate space for file about to be sent. Parameter specifies number of bytes. |
| *APPE | Append file to existing file. |
| CDUP | Change to parent directory on remote system. |
| CWD | Change working directory on remote system. |
| DELE | Delete file on remote system. |
| HELP | Elicit "helpful" information from the server. E.g. a list of commands supported. |
| LIST | Send a list of file names in the current directory on the remote system on the data connection. |
| MKD | Make directory. |
| MODE | Specifies transfer mode. Parameter is S,B or C. |
| NLST | Send a "full" directory listing of the current directory on the remote system on the data connection. |
| NOOP | Do nothing. |
| PASS | Supplies a user password. Must occur immediately after the USER command. |
| *PASV | Specifies that the server data transfer process is to listen for a connection request from the client data transfer process. |
| PORT | Specify the client port number on which the data transfer process is listening for a connection request. |
| PWD | Show current directory name on remote system. |
| QUIT | Logout or break the connection. |
| *REIN | Reinitialize. Logout without breaking connection. A new USER command for a different user would follow. |

| | |
|---|---|
| *REST | Restart transfer from server marker. |
| RETR | Get file from remote system. |
| RMD | Remove directory. |
| *RNFR | Specifies old path name of file to be renamed. Follow with RNTO command. |
| *RNTO | Specifies new path name of file to be renamed. |
| *SITE | Site specific server services. |
| *SMNT | Structure mount. Supplies the remote system path name of a file system structure. |
| *STAT | Elicit status information. |
| STOR | Store file on remote system over-writing the file if it already exists. |
| *STOU | Store unique. Does not over-write existing files. |
| STRU | Specifies file structure. Parameter is F,R or P. |
| *SYST | Report operating system type on remote system. |
| TYPE | Specifies representation (file) type. Parameter is one of the characters A,E,I,L for file type followed by N,T or C for format control or a number specifying the local byte size. Only TYPE A and TYPE I are common. |

### 2.5.4 FTP REPLIES

Replies to FTP commands are devised to ensure the synchronization of requests and actions in the process of file transfer, and to guarantee that the user process always knows the state of the server. It starts with 3 digit ASCII numbers with an optional message.
A long reply may be sent as several messages with a dash after the three digits on the first message and no dash after the three digits on the last message. Table 2.7 and Table 2.8 show the encoding of FTP reply codes.

Table 2.7 FTP Reply Codes 1 [43]

| Type | Description |
|---|---|
| 1yz | Positive preliminary reply. Expect another reply before sending another command. |
| 2yz | Positive completion reply. The last command completed successfully. |
| 3yz | Positive intermediate reply. A further command must be sent. |
| 4yz | Transient negative completion reply. The requested action did not take place but can be retried. |
| 5yz | Permanent negative completion reply. The requested action did not take place and should not be retried. |

The "y" digit encodes further information

Table 2.8 FTP Reply Codes 2 [43]

| Digit | Meaning |
|---|---|
| 0 | Syntax error |
| 1 | Information |
| 2 | Connection status. |
| 3 | Authentication and accounting. |
| 4 | Unspecified |
| 5 | File system status |

Table 2.9 Example of Some Typical FTR Replies Messages[43]

| Number | Meaning |
|---|---|
| 125 | Data connection already opens; transfer starting. |
| 200 | Command OK |
| 331 | User name OK, Password Required. |
| 425 | Can't open data connection |
| 452 | Error writing file |
| 500 | Syntax error – unrecognized command |

## 2.6 SYNTHESIS OF LOCAL AND REMOTE DATA BACKUP SYSTEM

Synthesis of the local and remote data backup system will be performed based on the literature review. Apparently, the requirements of this system are to:

1. Provide a scheduled basis backup system that will be performed automatically. This may include day, time and time interval for the backup procedure.
2. Provide documentation or log files of the performed backup procedure. This may include the time, date and backup progression.
3. Perform an incremental backup that will only backup the files that have been changed since that last backup.
4. Provide local data backup system that compatible with the workstation-based backup, file-server-based backup and dedicated backup server option.
5. Provide remote data backup system that backup over the WAN to the remote server, using FTP.
6. Restore data over the WAN from the remote server.