Chapter 5
Storage Infrastructure for Big Data and Cloud

Anupama C. Raman
IBM India Pvt Ltd, India

ABSTRACT

Unstructured data is growing exponentially. Present day storage infrastructures like Storage Area Networks and Network Attached Storage are not very suitable for storing huge volumes of unstructured data. This has led to the development of new types of storage technologies like object-based storage. Huge amounts of both structured and unstructured data that needs to be made available in real time for analytical insights is referred to as Big Data. On account of the distinct nature of big data, the storage infrastructures for storing big data should possess some specific features. In this chapter, the authors examine the various storage technology options that are available nowadays and their suitability for storing big data. This chapter also provides a bird's eye view of cloud storage technology, which is used widely for big data storage.

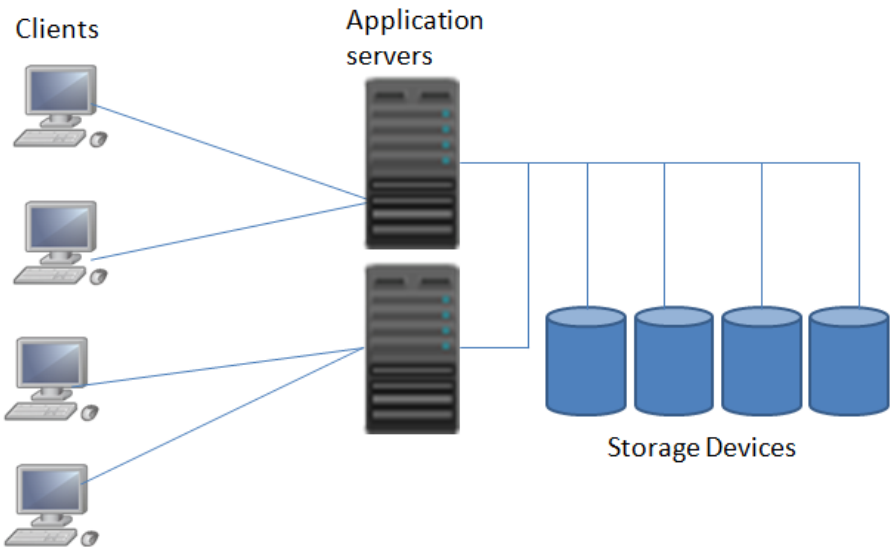
INTRODUCTION

In the initial stages of its evolution, Storage Area Network (SAN) was perceived as a client server system with the server attached to a collection of storage devices by means of a bus. In many scenarios, the client systems were directly connected to the storage devices as well. These storage architectures were referred to as Direct Attached Storage (DAS). The high level architecture diagram of a DAS system is given in Figure 1.

There are three main tiers in the architecture given above, they are:

DOI: 10.4018/978-1-4666-5864-6.ch005

Figure 1. Architecture of DAS



1. Tier one is comprised of the client devices which are connected to the application server using some kind of a switch.
2. Tier two comprises of the application servers where the applications are hosted. The application servers have (Input /Output) I/O controllers to control input/output operations to the attached storage devices. The I/O controllers are designed to work according to the specific interfaces which are used for connecting to the storage devices. If the attached storage devices support different types of interfaces, there will be an I/O controller for each type of interface.

The following are the some of the key types of connectivity interfaces supported by the storage devices in a DAS system:

Small Computer System Interface (SCSI)

It is a set of American National Standards Institute (ANSI) standard electronic interfaces. Parallel SCSI (also called as SCSI) is one of the most popular forms of storage interface. It is mainly used to connect disk drives and tapes to the servers or client devices. It can be also be used to connect other devices such as printers and scanners. Communication between the source (servers/ client devices) and the attached storage devices are done using the SCSI command set. The latest version of SCSI which is SCSI ultra 320 provides data transfer speeds of 320 MB/s. There is also a serial version of SCSI called Serial Attached SCSI (SAS). It offers better performance and scalability when compared to SCSI. SAS currently supports data transfer rates of 6 Gb/s.

Integrated Device Electronics/ Advanced Technology Attachment (IDE/ATA)

The term IDE/ATA denotes the dual-naming conventions for various generations and variants of this interface. The IDE component in IDE/ ATA provides the specification for the controllers connected to the computer's motherboard or communicating with the device attached. The ATA component is the interface for connecting