Storage is a key element in an enterprise IT infrastructure. Large organizations typically have huge storage demands, which they try to address by various means. In fact, an organization's worth could be directly related to the bits & bytes it has on its disks! Therefore, carefully addressing the storage requirements remains a very important task for any enterprise. Cloud storage works through data center virtualization, providing end users and applications with a virtual storage architecture that is scalable according to application requirements. In general, cloud storage operates through a web-based API that is remotely implemented through its interaction with the client application's in-house cloud storage infrastructure for input/output (I/O) and read/write (R/W) operations.

The Authors Dr. Sandip Roy, Assistant Professor, Dr. Rajesh Bose, Deputy Manager(Industry) and Ms. Srabanti Chakraborty, Assistant Professor-all have individually or as co-authors published papers in leading journals on cloud computing and IoT.

978-613-9-44607-0

Storage Technology

Sandip Roy
Rajesh Bose
Srabanti Chakraborty

Evolution and Exploration of Storage for Cloud Computing

Storage Technology

LAP LAMBERT Academic Publishing

Roy, Bose, Chakraborty

Sandip Roy
Rajesh Bose
Srabanti Chakraborty

Evolution and Exploration of Storage for Cloud Computing

Sandip Roy
Rajesh Bose
Srabanti Chakraborty

Evolution and Exploration of Storage for Cloud Computing

Storage Technology

LAP LAMBERT Academic Publishing

Imprint

Cover image: www.ingimage.com

Publisher:
LAP LAMBERT Academic Publishing
is a trademark of
International Book Market Service Ltd., member of OmniScriptum Publishing Group
17 Meldrum Street, Beau Bassin 71504, Mauritius

Printed at: see last page
ISBN: 978-613-9-44607-0

Evolution and Exploration of Storage for Cloud Computing

Dr. Sandip Roy,
Dr. Rajesh Bose and Srabanti Chakraborty

About the Authors

Dr. Sandip Roy, an Assistant Professor and Head of Department of Computer Science & Engineering of Brainware Group of Institutions-SDET, Kolkata, West Bengal, India. He has awarded his Ph.D. in Computer Science & Engineering from University of Kalyani, India in 2018. Dr. Roy received M.Tech. degree in Computer Science & Engineering in 2011, and B.Tech. in Information Technology in 2008 from Maulana Abul Kalam Azad University of Technology, West Bengal (Formerly known as West Bengal University of Technology). He has authored over 30 papers in peer-reviewed journals, conferences, and is a recipient of the Best Paper Award from ICACEA in 2015. He has also authored of 3 books. His main areas of research interests are Data Science, Internet of Things, Cloud Computing, and Smart Technologies.

Dr. Rajesh Bose is an IT professional employed as Deputy Manager with Simplex Infrastructures Limited, Data Center, Kolkata. He has completed PhD from University of Kalyani, Kalyani, West Bengal. He graduated with a B.E. in Computer Science and Engineering from Biju Patnaik University of Technology (BPUT), Rourkela, Orissa, India in 2004. He went on to complete his degree in M.Tech. In mobile communication and networking from Maulana Abul Kalam Azad University of Technology, West Bengal (Formerly known as West Bengal University of Technology) - WBUT, India in 2007. He has also several global certifications under his belt. These are CCNA, CCNP-BCRAN, and CCA (Citrix Certified Administrator for Citrix Access Gateway 9 Enterprise Edition), CCA (Citrix Certified Administrator for Citrix Xen App 5 for Windows Server 2008). His research interests include cloud computing, IoT, wireless communication and networking. He has published more than 45 referred papers and 4 books in different journals and conferences. Mr. Bose also has experience in teaching at college level for a number of years prior to moving on to becoming a full time IT professional at a prominent construction company in India where he has been administering virtual platforms and systems at the company's data center.

Ms. Srabanti Chakraborty, an Assistant Professor and Head of the Department of Computer Science & Technology of Elitte Institute of Enginering & Management, Kolkata, West Bengal, She received her M. Tech. degree in Computer Science & Engineering in 2012 from Maulana Abul Kalam Azad University of Technology (Formerly known as West Bengal University of Technology), West Bengal and MCA degree in 2009 from Indira Gandhi National Open University (IGNOU). She has authored over 4 peer-reviewed papers in National and International journals and conferences. Her main areas of research interests are Cloud Computing, Internet of Things (IoT) and Big Data.

ACKNOWLEDGEMENT

This kind of book is shaped for distributing knowledge. I cannot but trust it that devoid of the benevolence of the benign God, it was perfectly preposterous to shape the bottom line of the book in the first position. I am grateful to the God as He blessed me with the power to write. Moreover, I am thankful to Him as He has gifted diligence to me to execute what this takes to provide this book with the subject-matter, structure in which this has been offered for the readers for whom the book has been connoted.

We are blessed as we could communicate with some of the most conversant experts in the IT sector for the period of epitomizing this book. Feasibly, most of those experts have supplied appealing information and profound intuitiveness of the procedure by which cloud computing have been progressed to the juncture where it is at present.

My co-writers, Dr. Rajesh Bose and Srabanti Chakraborty, have been involved to provide form and essence to the book. Devoid of their priceless role, much of the research which has helped shape this book would have been preposterous. In spite of the existence of the connoisseurs in their relevant domains, they have relaxed much reliance and buoyancy in my competence to take part in the most important role in compiling this book.

Devoid of my parents, my targets and aspiration would have been apprehended. My mother, Mrs. Kalyani Roy, my father, Mr. Sankar Prasad Roy whose authorities have been at the nucleus and spirit of my endeavors in conveying my study to culmination. With their approvals and supportive words that I value, I accumulated the power to go forward through succeeding waves of tryouts and ordeals. They were the ones to understand the worth of technology years ago. I appreciate their chastity and vision that has empowered me to attain this humble target.

My own inputs, nevertheless, would not have become visible without my family's immense support in my mission and passion to write my first book. I am really grateful to my wife, Tanaya, who has consistently supported me in my efforts to compile the primary outlines and run of the chapters. Devoid of her unwavering support and devotion, it would have been preposterous to enlarge my perspectives and go on with my fight for my career. I wholeheartedly dedicate this book to her.

Besides, I am thankful to my friends and colleagues who have offered me their wholehearted support at all junctures of writing.

SUMMARY

Storage is a key element in an enterprise IT infrastructure. Large organizations typically have huge storage demands, which they try to address by various means. In fact, an organization's worth could be directly related to the bits & bytes it has on its disks! Therefore, carefully addressing the storage requirements remains a very important task for any enterprise. Storage has moved from the traditional Direct Attached Storage (DAS) to Network Attached Storage (NAS) and now to Storage Area Networks (SAN). SANs offer several advantages over conventional storage models such as consolidation of storage, server less backup improving the server resource utilization, and improved storage utilization.

As storage usage increases exponentially, improving storage efficiency is critical for many data centers today. There are many viable solutions to achieve this objective. These include data migration, thin provisioning, content and quota management, and data deduplication. However, the solution you implement is based on what you believe is a suitable or effective approach to achieve storage efficiency in your production environment. If the intent is to reduce the cost of storing data at the file system level by achieving space savings without affecting end-user experience, and to propagate these space savings within the storage environment at the file system level, an appropriate technology to consider is data deduplication.

Cloud storage works through data center virtualization, providing end users and applications with a virtual storage architecture that is scalable according to application requirements. In general, cloud storage operates through a web-based API that is remotely implemented through its interaction with the client application's in-house cloud storage infrastructure for input/output (I/O) and read/write (R/W) operations.

A cloud database is a database that typically runs on a cloud computing platform, access to it is provided as a service. Database services take care of scalability and high availability of the database. Database services make the underlying software - stack transparent to the user. Cloud platforms allow users to purchase virtual-machine instances for a limited time, and one can run a database on such virtual machines. Users can either upload their own machine image with a database installed on it, or use ready-made machine images that already include an optimized installation of a database. With a database as a service model, application owners do not have to install and maintain the database themselves.

Explaining the Working Principle of Cloud Storage offers a synopsis of the cloud to reader and also explore the infrastructure requirements for setting up a SAN, important design considerations and challenges involved in architecting a SAN, some proprietary & open source file systems that form the basis of low-level storage in a SAN, and factors to consider in calculating the return on investment (ROI) which one may obtain on adopting a SAN. This book assists readers to comprehend what the cloud storage is and what is the method to work with it, though it isn't a portion of their regular accountability. Writers illustrate the working principle of cloud storage in realistic expressions, assisting readers to comprehend the procedure of controlling cloud services and offer worth to their trades via converting information to the cloud. This book provides readers a theoretical knowledge and architecture for going ahead with cloud storage that looks for being comprehensive directions to the cloud.

FOR WHOM THE BOOK IS COMPILED?

All over the world, corporate strategy designers and they who have authority for the work of conveying the borderlines of their specific companies' information technology architecture have initiated to embrace in a prodigious means. Once when it was an emerging technology, service providers and end-consumers continued the range of coveted and acquired presentation stratums. Most was an issue of trial and error.

Due to growing rivalry, most of the cloud based business keeps equilibrium cautiously. This has turned out to be significant. Consequently, managers and Data Center executives can make a cloud elucidation which will be economical for them to operate supply. For the businessmen and students, the context of service rank management, service rank purposes, and ranks of confirmations can be bewildering, immature lump of concepts. This book has been shaped to augment and explain several of these subject matters associated with cloud computing services rank deals.

Enterprises and beginners on a financial plan would discover the segment of on cloud service invoice of engrossment. We expect that for safety executives and data centre executives, the division of executing safety stratums and catastrophe redemption schemes would shape the nucleolus of the book.

We wait to get synopsis made by our readers. Getting charged up by the students, IT experts and those endowed with the assignment of constituting the track to drifting to cloud, the writers created this book. Under no circumstances is this book finished. Nevertheless, it is expected that it will assist them keen in cloud computing to notice enthusiasms irrespective of probable intricacy and murk of the clouds of computing.

Table of Contents

CHAPTER 1

INTRODUCTION OF STORAGE TECHNOLOGY

1.1 Introduction to Storage Technology

With the varying data storage requirements, storage has moved from the traditional Direct Attached Storage (DAS) to Network Attached Storage (NAS) and further to high performance networks dedicated for storage referred by Storage Area Network (SAN). Although NAS are intelligent storage devices that connect to networks and provide file access/ storage facilities to clients, SANs best address the extensive storage requirements in a distributed environment.

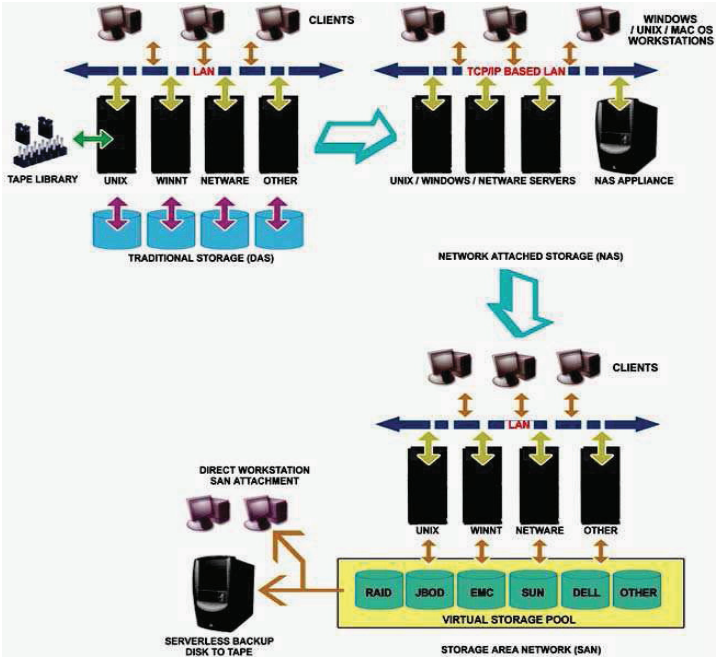


Figure 1.1: Move from DAS to NAS & SAN

1.2 NAS vs SAN: A Technical Perspective

There is still some confusion amongst new users of storage technologies, regarding the use of a NAS and a SAN. The table below lists and clarifies the important differences between these. Although NAS and SAN address a different set of issues, they are complementary solutions rather than competing technologies. Today IT experts are talking about a fusion of NAS and SAN wherein enterprises can make use of a NAS as part of their IT infrastructure while making use of a SAN for backing up data from multiple servers as well as a NAS, at a high speed. This is also popularly referred to as NAS-SAN Convergence.

NAS	SAN
An Independent device	A network of storage devices
Attached on the primary LAN	Acts as a secondary network to LAN
Connected by Ethernet	Connected using Fibre Channel (FC) or SCSI over IP (iSCSI)
Uses standard protocols such as TCP/IP, CIFS, NFS, HTTP	Uses FC or SCSI protocol for data transfer. The recent IP-based SAN also make use of TCP/IP
NAS appears as a single node on network	SAN appears as extra storage for each server
NAS follows a client/server model	SAN provides direct access to the disks
Best suited for file sharing and applications involving data transfer of short duration and volume	Best suited for data-intensive applications, and mission-critical database applications
No reduction of load on main network as the device itself is connected on main network	Reduces the load on the main network, thus reducing backup and recovery time

Table 1.1: Differences between a NAS and a SAN

1.3 Why and When to Adopt a SAN

There could be many benefits in adopting a SAN. Some of the significant ones are enumerated below.

- **Consolidation of Storage** – With the traditional storage model, administrators have to manage multiple storage devices. Backup for each of these storage devices is also a cumbersome process. Consolidation of these individual storage entities could solve many of such problems.
- **Server less Backup/Improved Server Resources** – Server less backup allows disk storage device to copy data across the high-speed links of the SAN directly to a backup device without any intervention of the server. Data is confined to the SAN boundaries and the clients get uninterrupted access to the server resources.
- **Better Utilization of Storage Facilities** – With SAN, one may improvise storage as per ones requirements, thus considerably increasing the utilization of storage. In the traditional storage model, even though we may have plenty of vacant space on a storage drive attached to a server, it may not be possible to make use of the same for another system, which has run out of storage space.
- **Integration with a Disaster Recovery Solution** – Integration with a disaster recovery solution or a replication solution also becomes easier, as with a SAN, the challenge is confined to looking for solutions only for an integrated and consolidated storage space rather than a scattered and distributed storage space.
- **Improved Scalability** - While individual resources and servers have a restriction on the number of storage and interconnected units they can attach, a SAN is not affected by such constraints, leading to a higher scalability.

The decision of adopting a SAN may be based on a number of factors. Some of these are as follows:

- ✓ The primary factor is the extensive growth in the storage requirement. Many experts are of the view that storage size of approximately 3 TB or more should be a reasonable point at which one can start looking towards SAN. Many others feel that the number of servers exceeding say 20 could be considered as the starting point for the adoption of SAN. However there is a lot of subjective element involved in the decision of adopting a SAN and varies from organization to organization

- ✓ When a network grows with a heterogeneous mix of servers and their corresponding storage, managing these becomes a challenge

- ✓ Need for a disaster recovery solution

- ✓ Budget availability

- ✓ The initial investments for setting up a SAN could be very high. However, the Return on Investments (ROI) over a period of time may justify the high initial investment. The section on SAN ROI discusses factors that need to be taken into account in order to calculate the ROI from a SAN

1.4 SAN Infrastructure

SANs involve a variety of technologies and related equipments and devices. This section takes a brief look at the popular SAN technologies,

- ➢ **FC** – FC is a technology designed for very high performance low-latency data transfer among various types of devices. The FC protocol is based on the SCSI protocol and makes use of the common SCSI command set over the FC protocol layer. The FC protocol may be implemented both over optical fibre as well as copper cable.
- ➢ **FC Switch** – An FC switch provides multiple simultaneous interconnections between pairs of ports with the resultant increase in total bandwidth. FC switches are used to implement FC fabric interconnection.

Figure 1.2: Fibre Channel Switch

- **FC Hub –** An FC hub is used to implement the FC Arbitrated Loop (FC-AL) protocol. Hubs pass signals arriving from one port to the next port in the loop. It is up to the devices to intercept and process signals addressed to them.
- **FC Arbitrated Loop (FC-AL)** – FC-AL is an FC topology that provides a solution for attaching multiple communication ports in a loop. In an FC-AL, communication is not broadcast as it is in architectures like Ethernet. Instead it is transmitted from one device to the next with each device repeating the transmission around the "loop" until the data reaches its destination. The devices arbitrate for access to the loop before sending data.

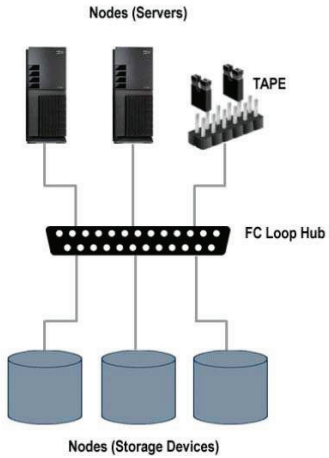


Figure 1.3: Fibre Channel Arbitrated Loop Architecture

- **Fabric –** A 'fabric' is a network of FC switches providing interconnectivity and scalability. It is used to describe a generic switching environment. With a fabric the bandwidth is not shared.

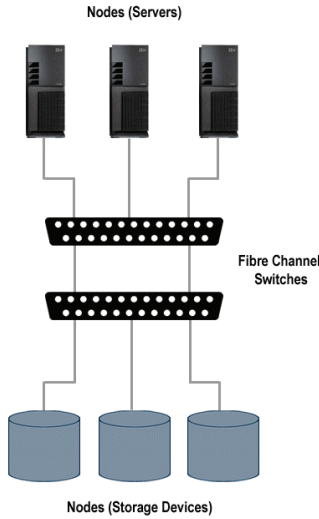


Figure 1.4: Fibre Channel Fabric Using Switches

- **Host Bust Adapters (HBA) –** An HBA provides the interface between a server and the SANs network. Every HBA has a corresponding device-driver, which handles the I/O and control requests.

Figure 1.5: Host Bus Adapter (HBA)

The HBA connecting a server to a SAN may be an FC HBA (for an FC based SAN) or an iSCSI HBA (for an IP based SAN). Some of the latest HBAs may have the support for both FC and iSCSI.

- **Storage Subsystem**

  - ✓ **Small Computer Systems Interface (SCSI)** – This is a parallel interface standard used for attaching peripherals (including disk drives) to a computer. SCSI enables faster data transmission rates compared to other popular interfaces like IDE, serial and

parallel ports. In addition to this, many devices can be attached to a single SCSI port. Server grade systems and SAN storage boxes invariably use SCSI disk drives.

- ✓ **Redundant Array Of Inexpensive Disks (RAID) –** This is a mechanism for providing disk fault tolerance. Five types of RAID architectures have been defined. They are RAID-1 through RAID-5. Each provides disk fault-tolerance and offers different trade-offs in features and performance. In addition to these five redundant array architectures, it has become popular to refer to a non-redundant array of disk. drives as a RAID-0 array. Possible approaches to RAID include hardware RAID and software RAID. Internal hardware RAID solutions involve presence of a RAID controller inside the server. In external hardware RAID solutions, the hardware RAID controller and the disk drives are housed separately from the server in a high-availability external RAID enclosure. The external hardware RAID controller-based storage system may be attached directly to the SAN.

- ✓ **Just a Bunch of Disks (JBOD) –** This refers to a set of disks that has not been configured into a RAID array but can be used as if they were a single volume. This can be used for applications, which require more storage space than that offered by the disks individually.

- ✓ **Gigabit Interface Converters (GBIC) –**This is a removable transceiver. It interconverts electrical and optical signals for high-speed networking. GBICs are used in all types of FC devices including switches and HBAs. Initially targeted to support FC data networks, the GBIC standard was quickly adopted for use with Gigabit Ethernet installations as well. By providing hot-swap inter changeability, GBIC modules give net administrators the ability to tailor transceiver costs, link distances, and configure overall network topologies to meet their requirements.

Figure 1.6: Gigabit Interface Converter (GBIC)

- ✓ **Internet SCSI Protocol (iSCSI) –** iSCSI protocol enables deployment of a SAN over conventional Ethernet based network. The iSCSI protocol uses TCP/IP as its network transport protocol and is designed to leverage TCP/IP for block storage needs. However, there are a few challenges in the acceptance of iSCSI as the SAN interconnect. TCP/IP has traditionally been tuned to favour short and bur sty user transmissions as against large and continuous data transfer requirements of storage. However, several vendors have announced their support for usage of iSCSI to reduce

the overhead processor. Once this overhead becomes comparable to that of FC, iSCSI would present itself as a serious competitor of FC (which has already begun).

- ✓ **FC-iSCSI Gateway (SAN Gateway) –** In a heterogeneous SAN containing both FC and iSCSI-based devices, an FC-iSCSI gateway provides the internetworking of iSCSI devices with FC devices. The gateway maps selected iSCSI devices into the FC SAN and selected FC devices into the IP SAN.

- ✓ **SAN Management Software –** A SAN management software, as the name suggests, assists in the management of the SAN environment. The tasks of any typical SAN management software include discovery and mapping of storage devices, switches, and servers; monitoring and alerting on discovering devices and logical partitioning and zoning of the SAN. With increase in the number of vendors providing SAN products, the complexity in SAN environments has increased tremendously. This has made the management of a SAN extremely challenging, as a good SAN management tool should be able to perform well in a multi-vendor heterogeneous SAN environment.

- ✓ **Important players in the SAN segment -** Some of the important names in the SAN hardware and server segment are EMC, Network Appliance, Brocade, Qlogic, McData, IBM, HP, Cisco, Compaq, Broadcom, Emulex, and Fujitsu. However, most of the SAN hardware vendors provide management and application software as well. There are other names such as Veritas, BMC Software, and Sun, which specialize in SAN management and application software space.

1.5 Architecting a SAN-Important Design Considerations

There are typically two kinds of SAN architectures, which are currently popular and gaining ground. These are:

FC-based SAN
IP-based SAN

There are many common steps and phases involved in building both FC and IP-based SANs. This section takes a detailed look at these phases.

- ➢ **Important Phases in SAN Architecture**

The important phases in building a SAN include planning, design and implementation, setting up security, applications and management. Each of the phases is detailed in the subsequent sections.

Planning-Data Collection and Analysis - Data Collection lays the foundation for designing the SAN. The info collected should be concrete and accurate. An elaborate plan containing the needs and views of all the stakeholders (system administrator, storage administrator,

network administrators) should be formulated. Questions regarding the estimated storage space requirements in the next few years, identification of the important applications and services that need to be available continuously, performance requirements of the same, backup strategy and so on should be adequately answered. As a starting point for building the SAN, one may prepare an inventory of the current storage infrastructure and carry out an analysis for identifying components to be used with the future SAN setup. The analysis would also incorporate the identification of components needed in the SAN apart from the port requirements. Use of a template such as a SAN Inventory Worksheet provided by Brocade may be a good way to begin this process.

Since SAN may involve components from a variety of companies, it is very important to verify that the components selected are compatible with each other. SAN vendors provide a compatibility list consisting of detailed information on compatibility of their products with products from other vendors. The details should be verified very carefully in order to ensure that compatibility/interoperability issues do not arise while implementing the SAN. A partial view of the compatibility list for Legato Net Worker with HBAs from different vendors is given below.

OS	Vendor	Model	Bus Type	Firmware
AIX 4.3.3	Bull*	LP7000e	PCI	3.22A1
AIX 4.3.3	Cambex	PC1000	PCI	
AIX 4.3.3	IBM**	6227	PCI	3.22A1
AIX 4.3.3	IBM**	6228	PCI	3.82A1
AIX 4.3.3	Emulex**	LP9002DC/LP9002DC	PCI	3.81A3
AIX 5.1L	Cambex	PC1000	PCI	

Figure 1.7: Partial View of SAN Component Compatibility List

Design & Implementation of the SAN - As mentioned earlier, the data collection phase also involves identification of services and applications viz-a-viz the requirements such as performance requirements, availability requirements and scalability requirements. These requirements may vary from organization to organization and translate directly in the design of the SAN. This means that a particular SAN design may be used to address only a particular business problem. There may be associated trade-offs with each design. For instance, a high availability requirement may be associated with a redundant component, and therefore higher costs. Keeping the cost of the SAN low may lead to reduced availability and performance. This section takes a look at some sample SAN implementations, built using the components identified in the data collections and analysis phase, catering to different business problems and requirements.

High Availability (HA) SANs – Although an HA-SAN may also address other issues as discussed earlier, the primary goal of an HA-SAN is to provide continuous availability of resources and services. There could be several ways of implementing an HA-SAN. For

instance, incorporating redundant paths within the fabric from the server to the storage may satisfy HA requirements. The figure 8 below shows one such HA-SAN architecture. Here, each server has multiple adapters and is connected to multiple fabrics. In the event of the failure of one fabric, the servers can communicate using the remaining fabric. Such systems may use specialized multipathing software to ensure that the hosts get a single view of the devices across the two HBAs. In some configurations, it may be possible to link the switches into a single HA fabric.

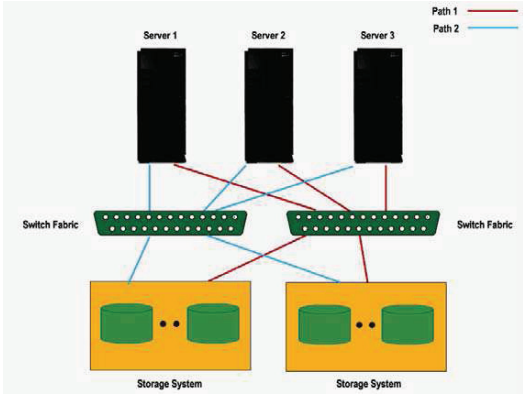


Figure 1.8: A High Availability SAN

An HA-SAN may also involve the use of clustering, which provides continuity of data access in spite of server or application failure. The connectivity provided in a SAN makes it more feasible to apply clusters. However, clustering has its own associated challenges and design issues. Therefore it is advisable to take up clustering in the later stages of SAN implementation.

Server less Backup using SAN – As mentioned earlier, one of the major advantages of a SAN is enabling server less backup. In this the disk storage device can copy data directly to a backup device across the high-speed links of the SAN without any intervention of the server. With backup being handled by SAN devices, enterprise servers can concentrate on application processing rather than getting involved in tasks such as backup. Server less backup is enabled by a protocol-aware and intelligent SAN appliance, which can recognize protocols from many heterogeneous systems and transmit data at high speeds to the tape or tape libraries. The figure 9 below displays a SAN setup for facilitating server less backup.

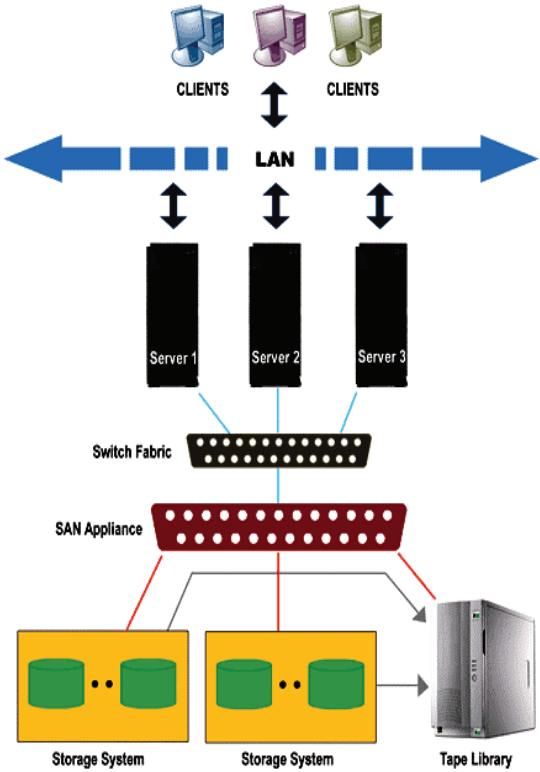


Figure 1.9: Server less Backup over SAN

Setting up Partitioning and Security-Zoning/LUN Masking - Generally, most of the times SANs are intended to host mission critical data. Security of the data therefore has to be giving great importance. This section discusses zoning and LUN masking which are important mechanisms of ensuring SAN security.

Zoning – This is a logical separation of traffic between host and resources. Usually, on a SAN, storage resources are shared across servers and users. Therefore by separating the traffic between hosts and the resources, one may restrict the accessibility of the servers to a particular set of storage resources. Zoning allows segregation of a node on an FC switch on the basis of a physical port, name or address. There are two types of zoning: Hard Zoning and Soft Zoning.

Hard Zoning – This is implemented in the hardware (switch) by linking physical ports to the FC fabric. On using this, if two ports are not authorized to communicate with each other, then the communication between these two ports is blocked. This is supposed to be safer and easier to implement but is less flexible than soft zoning.

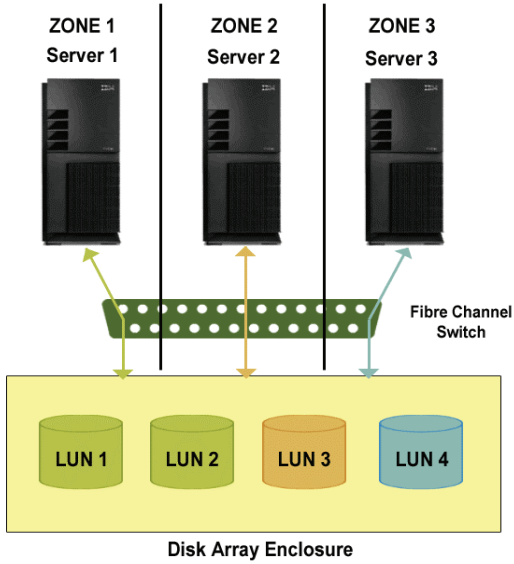


Figure 1.10: Hard Zoning

Soft Zoning – This is implemented in software and relies on the World Wide Name (WWN), which makes use of a name server that runs inside the fabric switch and allows or blocks access to a particular WWN in the fabric.

What Kind of Zoning Should One Use?
As mentioned earlier, soft zoning being based on the WWN, is much more flexible than hard zoning. However, it is much less secure. It is possible to spoof a WWN and write to devices that are not visible. By limiting access to specific ports, hard zoning eliminates this possibility. Wherever possible, it is advisable to make use of hard zoning.

Logical Unit Number (LUN) Masking – LUN refers to the individual piece in the storage system that is being accessed. LUN masking restricts access of servers to specific logical storage units (devices) assigned to them. It is a RAID-based feature that binds the WWN on a particular HBA on a server to a specific LUN. This feature cannot be implemented by using zoning. The hardware connections to other LUNs may exist but LUN masking makes those LUNs invisible to the servers. LUN masking can be implemented using hardware or software and is often built into SAN components such as storage controllers and routers. A combination of hard zoning and LUN masking provides a robust security layer.

Setting up SAN Applications - Once the physical components in a SAN have been put in place, the next important step is setting up applications that leverage the SAN infrastructure for fulfilling the business requirements. Some important SAN applications include Virtualization and server clustering applications. Storage Virtualization normally involves a

specialized server running storage virtualization software, which acts as a gateway between the storage and the servers. Two of the many virtualization packages are FalconStor's Ipstor and Data Core's SAN symphony. As another example, in an HA-SAN, hosts and devices must have multiple adapters. In the case of a host, multiple adapters are typically treated as different storage buses. Additional multi-pathing software such as Compaq Secure Path or EMC Power Path is required to ensure that the host gets a single view of the devices across the two HBAs.

SAN Management - The primary tasks of SAN Management include discovery of storage devices such as HBAs, servers and switches, raising alerts for newly discovered devices, logical partitioning and zoning, configuring and provisioning storage and generating notifications and alarms for any failures. As mentioned earlier, the increase in the number of SAN vendors have led to an increased complexity in SAN environments, thereby making it a challenge to manage a SAN. Because of this reason, storage administrators have to make use of a number of independent management applications that are tied to hardware from their respective vendors. A good SAN management tool should be able to perform well in a multivendor heterogeneous SAN environment.

Additionally, with the release of the Storage Networking Industry Association's (SNIA) Storage Management Initiative Specification (SMIS), which aims to facilitate interoperability between storage products from different vendors, storage management has taken a major step forward. Today, compliance to SMIS is becoming one of the key features to look for while purchasing a storage product. Some of the popular SAN Management solutions include EMC Control Center, HP Open View Storage Area Manager, Tivoli Storage Network Manager, BMC Storage Network Manager, and Veritas SAN Point Foundation Suite.

1.6 Integration of FC & IP-Based SAN

The emergence of new standards such as Internet SCSI Protocol (iSCSI), FC over IP (FCIP) and Internet Fibre Channel Protocol (iFCP) has made the integration of FC and IP-based SANs possible. This section takes a brief look at these standards and discusses deployment considerations for the standards.

iSCSI – This protocol enables deployment of a SAN over the conventional Ethernet based network. The iSCSI protocol uses TCP/IP as its network transport protocol and is designed to leverage TCP/IP for block storage needs. The iSCSI protocol involves encapsulation of SCSI commands and data into TCP/IP packets and subsequent transmission of the same over the network. At the receiving end, these encapsulated packets are de-capsulated and SCSI frames are retrieved. The protocol is used on servers, storage devices and protocol transfer gateway devices. For example, the FciSCSI gateway provides the inter-networking of iSCSI devices with FC devices. Initial iSCSI deployment targets those enterprises that have not made investments in FC-based SAN. With the advancements in Ethernet speeds, iSCSI-based SANs are truly becoming feasible.

Figure 1.11: FC-iSCSI Gateway

FCIP – FC over IP protocol provides a means to tunnel FC over IP based networks while keeping the FC packet and addressing intact. FCIP enables interconnection of FC based SANs making use of TCP/IP as the underlying transport and extends the interconnectivity of SANs across much greater distances. FCIP is designed to leverage the installed base of an FC SAN.

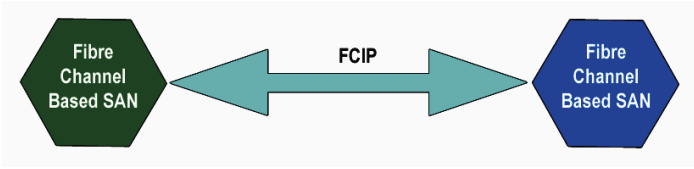


Figure 1.12: FCIP Tunnel

iFCP – This protocol is used between SAN islands to create large SANs also referred to as Storage Wide Area Network (SWAN). iFCP is used for communication between a pair of FC devices in contrast to FCIP, which facilitates an 'extended fabric'. iFCP is a gateway-to-gateway protocol where TCP/IP switching and routing components complement or replace the FC fabric. iFCP deployment comes into picture in scenarios where there has been considerable investment in both FC as well as IP based network infrastructure and where there may be a requirement to extend the IP based services to FC devices and SANs.

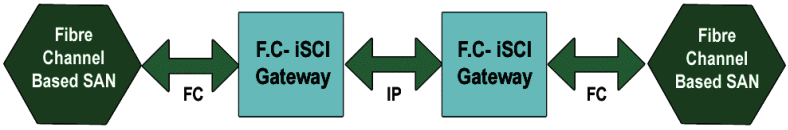


Figure 1.13: FC-iFCP Gateway

1.7 File system for SAN

For a proprietary SAN solution, one may not bother about the internals, for instance, the kind of file system to be used for low-level storage. However, while building SANs from custom components, the choice of an appropriate file system becomes an important consideration. This section takes a cursory look at some of the popular proprietary and open source file systems available for SAN.

Clustered XFS

Clustered XFS, or CXFS is a shared file system from SGI, designed as an extension to and based on 64-bit XFS. It enables data sharing by allowing SAN attached systems to directly access shared file systems. While file data flows directly between systems and disks, CXFS metadata is managed using a client server approach and passes through a metadata server for each CXFS file system. This metadata server acts as a central clearing house for metadata logging, file locking, buffer coherency and other necessary coordination functions. The CXFS metadata requests are routed over a TCP/IP network to the metadata server. The data requests are routed over FC to storage media.

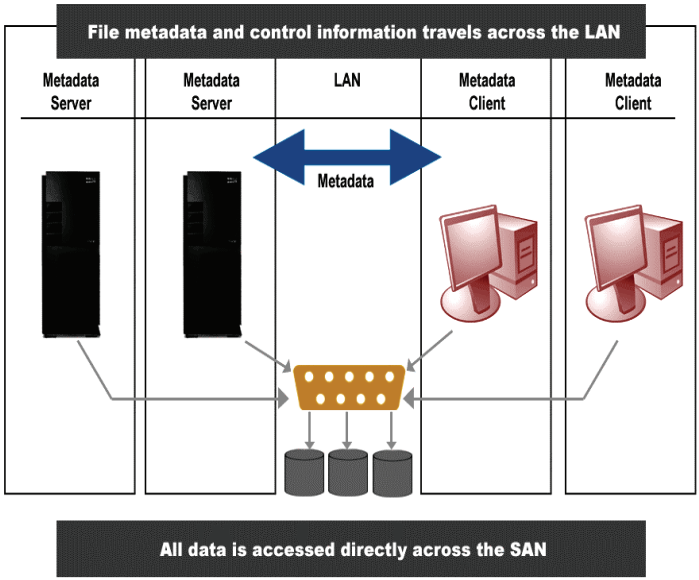
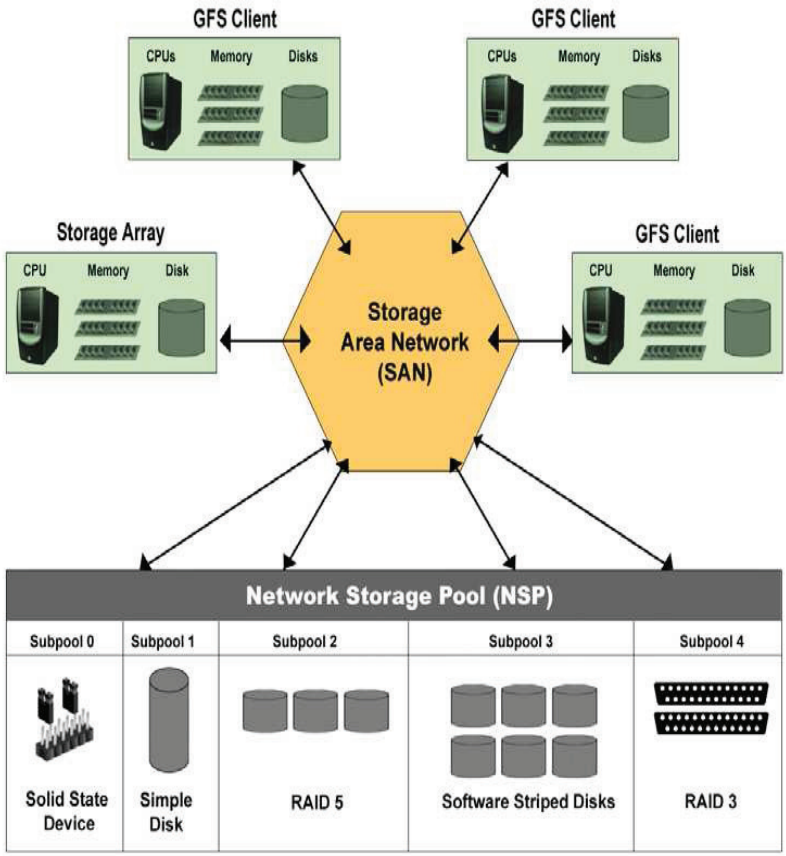


Figure 1.14: CXFS Architecture

Global File system

Red Hat's Global File system (GFS) (previously owned by Sistina) is a native cluster file system on Linux that allows multiple servers on a SAN to have concurrent Read and Write access to a shared data pool. Originally GFS was developed and released under the GNU GPL by the Department of Electrical and Computer Engineering, Parallel Computer Systems

Lab, Binary Operations Research Group (B.O.R.G), University of Minnesota, USA. It is now owned and developed by Red Hat Inc. The file system appears to be local on each node and GFS synchronizes file access across the cluster. GFS is fully symmetric, that is, all nodes are equal and there is no server, which could be a bottleneck or the single point of failure. GFS uses Read and Write caching while maintaining full UNIX file system semantics. GFS supports journaling and recovery from client failures.



Figur1.15: GFS Architecture

Polyserve

Polyserve Matrix Server (MxS) from Polyserve Inc. provides a clustered file system wherein the tasks that the cluster performs are distributed across its members. This enables symmetry across the nodes. For instance, both the metadata as well as the lock management are spread across all nodes in the cluster. It allows servers attached to a SAN to read and write data

concurrently with full data coherency. All access to storage from the Matrix Server Filesystem are handled by a Virtual Device Layer which provides persistent, cluster-wide device names and allows access to shared devices to be controlled by the Matrix Sever clustering infrastructure. Polyserve Matrix Server also integrates with Oracle9i Real Application Clusters to enable the use of Oracle features that rely on a cluster file system to operate in a multi-node environment.

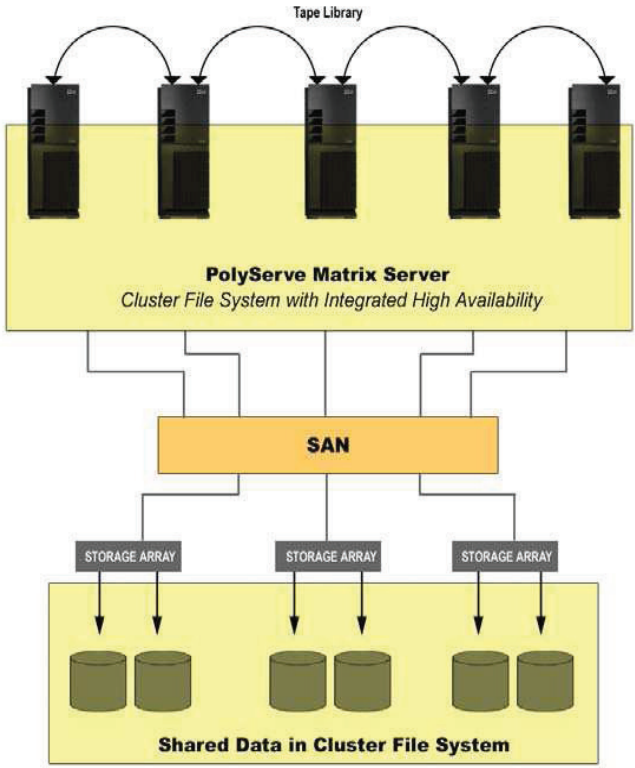


Figure 1.16: Polyserve Symmetric Architecture

Open Global Filesystem (OPENGFS)

This has its roots in the GFS project originally sponsored by the University of Minnesota from 1995 (when the project GFS 4.x was an open source) to 2000 (when it was made proprietary). Open GFS was started shortly thereafter, based on the 4.x source. It provides simultaneous sharing of a common storage device by multiple computer nodes. It further coordinates storage access so that the different nodes do not write on each other's data space while providing simultaneous read access for sharing data among the nodes. The current

versions can use cluster-aware volume managers such as the Enterprise Volume Management System (EVMS) and/or device mapper such as the DM device mapper. Recently, the Open DLM lock module was also added which is more efficient than Open GFS's memexp protocol. The lock module attaches to Open GFS via a plug in interface. Open GFS is an important component in many clustering projects including the Open Single System Image (SSI) clustering project.

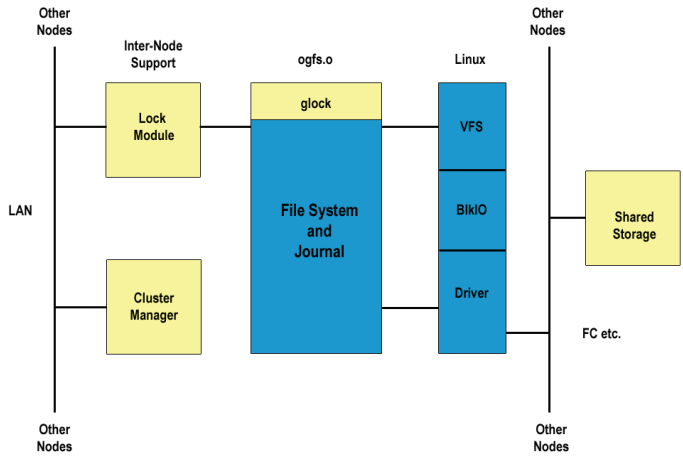


Figure 1.17: High Level Architecture of OpenGFS

1.8 When did Storage become so critical?

Looking at a modern virtual infrastructure, the hypervisor does an excellent job of making server hardware a commodity component, regardless of the technology used - VMware, Xen or Hyper-V. By making servers 'virtual' and IT department can move them between hardware swiftly and without downtime, leaving the physical machine to be viewed as little more than a CPU with a bit of memory. Conversely, in a VM environment, the storage system grows in importance as it becomes the underpinning of the entire infrastructure. In a traditional environment, most (if not all) physical servers have their own internal system disks and only rely on the SAN for application storage. In a virtualized environment the traditional system disks are provisioned from the central storage which not only adds load but

also randomizes the data pattern as many virtual servers all contend for the same disk resource. Consider this example; customer “A” looks to consolidate and virtualized their infrastructure. They have 25 Windows server, 5 Linux servers with a small SAN of 18 SAS drives to support MS Exchange, an ERP system, 2 small SQL databases and user home directories. The customer in this scenario will often invest in several new servers with a drastically increased number of CPU cores and large amount of memory but will neglect the storage. The customer likely considers using the existing small SAN as the VM storage. The physical environment contains at least 78 SAS drives. However, the general view is that this can be consolidated into just 18 disks within the existing SAN. This is where problems begin because not only are the number of spindles drastically reduced, but the workload is simultaneously randomized and put into contention for access to storage resources. The result is poor performance due to short sighted, inefficient storage design.

1.9 Performance vs Capacity

In the classic disk drive market, the rules of “Moore’s Law” can be observed with significant capacity increases every 18 months. However, what we do NOT see is significant increase in spinning disk performance. Disk vendors have continually upgraded the disk interface in an attempt to mask the shortcoming. But this does not affect sustained performance as it is limited to the physical mechanics which have remained unchanged for many years. This is clearly visible when benchmarking a SATA1 vs. SATA2 drive or a 3Gbit SAS vs. 6Gbit SAS drive. The sustained performance results remain the same. In the past this may not have presented a challenge as disk capacity remained so low that most SAN solutions have included upwards of 50+ disks to provide any useful capacity. This many disks provided plenty of IOPs per GB of capacity. In the current technology climate a prevalence of cost-effective SATA drives provide the same capacity with a fraction of the number of disks required. This significantly lowers the number of IOPs per GB and, if used in the wrong environment like a highly transactional database or large virtual server infrastructure, these disks and their IOP capability will bottleneck long before the capacity limit is reached. It is also worth looking at the approximate break-even cost points of different disk technologies. At the 100-3000 IOP range, SATA drives provide a very cost effective platform with pricing usually provided in price per GB. At the 3000-10000 IOP range SAS drives are usually the default technology as reaching this level of performance with SATA requires a vast amount of spindles and thus, wasted capacity. High performance disks are typically priced per GB, but sometimes per IOP. At the 10000+ IOP range, SSD begins to make financial sense. Within this range it is typical to find only small capacity requirements as only a fraction of a customer’s overall storage requires such levels of performance (keeping in mind that an average customer’s overall storage is over 70% static and never used). Once the 8000+ IOPs marker is reached, resellers frequently price per IOP as a price per GB becomes misleading and unattractive to the non-storage-savvy purchasing department. With this in mind, it is clear that balancing performance, capacity and cost is key to deigning an efficient storage system in a virtualized environment. Consider the graphs and tables to further explain these concepts.

To understand the importance of these concepts, consider that in a standard (non-virtualized) environment each server utilizes its own independent local disk (sometimes referred to as 'boot'). Moving this server to a virtualized environment dictates that it must 'share' its boot disk resource with many other virtual machines causing a state of contention with all VM's competing for performance from the same resource pool. It is very common to find a customer implementing far fewer disks in their virtual infrastructure than they had deployed in the physical model. As such, each virtual machine gets only the performance of a fraction of one disk. This causes a random data pattern which decreases performance.

1.10 I/O Data Patterns- Random vs Sequential

The pattern in which an application or 'host' server reads or writes its data can significantly affect the level of performance the storage system is able to provide. This can be further compounded by the choice of RAID level used as each RAID protection technique has a differing effect on write performance. This is generally referred to as 'parity penalty'. Data patterns are usually referred to as either Random or Sequential. A random data pattern infers that the data is written or read from random areas of the disk platter. This has two main effects on the performance of a RAID system. First, it drastically reduces the effectiveness of the controller cache as cache relies on patterns to 'guess' which blocks of data will be read or written next. In a random data pattern this is not possible as a random sequence of events can never be 'guessed' and, as such, cached. The second crucial effect of random patterns on storage systems is an increased number of 'seeks'. A seek refers to the point at which a disk head has to move to the next requested block of data. If this block of data is randomly placed it means the disks actuator arm and head must move a significant distance to 'seek' the block for each read or write. This adds significant overhead and lowers performance. SATA drives suffer under very random workloads as they only spin at 7200rpm. They utilize larger disk platters causing longer seek and access times (average 8.1ms access time). SAS drives are better suited as they spin at 15,000rpm and have smaller platters. Thus, each seeks takes about half the time when compared to SATA drives (average 3.3ms access time). In extreme high-performance applications, SSD can be used as it has no moving parts and therefore seek times are near non-existent. At this point it is worth noting a new disk type that has entered the storage market. NL-SAS (Near Line SAS) has caused some confusion in recent years as it shares the name with SAS but offers SATA-like capacities. Do not be fooled by this. An NL-SAS drive is nothing more than a SATA disk with a SAS connector and therefore offers the same performance characteristics as a SATA drive. The key elements in understanding the performance of a spinning disk in a random workload are the spin speed (RPM) and access time. The faster a disk spins, the more IOPs it will provide. In stark contrast, a sequential data pattern is one of structure and predictability. Some common examples of applications characterized by a sequential data pattern are data backup and video streaming. In these applications, the files are typically large and are written to the disk in continuous blocks and sectors. With this in mind, the RAID controller and disks can more easily 'guess' and/or cache the impending blocks of data to increase performance. In addition the disk actuator arm and head does not need to move a great distance to seek the next requested block. Such sequential applications are usually designed around MB/s (throughput). This design is rarely limited by disk speed and more commonly limited by the controller and interconnect. So, in a

storage design for sequential applications, SATA, SAS and SSD disks provide very similar levels of performance. The quick rule of thumb is that sequential patterns are those with large or streaming files (backup, archive, video etc.) and are best suited to SATA drives. Random workloads are typically those with very small files or storage requests which have no consistent structure (Virtual servers, virtual desktops, transactional databases, etc.) and are best suited to SAS or possibly SSD. The following illustrations demonstrate Random and Sequential disk patterns.

1.11 The Impact of Raid

Understanding data patterns and disk types is crucial when discussing storage design for specific applications. However, there are additional considerations. The RAID level/type must also be considered. The storage concept of "parity penalty" refers to the performance cost or performance impact of protecting data via RAID. This performance penalty only exists on writes. So, it is important to understand if the environment is write intensive or read intensive. Fortunately, most environments are the latter. These are the RAID protection parity penalties:

- ✓ RAID 0: ~0% overhead vs reads
- ✓ RAID 1+0: ~50% overhead vs read
- ✓ RAID 5: ~75% overhead vs read
- ✓ RAID 6: ~85% overhead vs read

Parity penalty depends on the way a block of data is written by the RAID protection level. Keep in mind that generating parity bits for each stripe of data incurs overhead. These figures are only truly visible in random write scenarios. In a sequential write environment, the RAID controller cache helps mitigate the performance impact. With these write overhead costs in mind, consider some best practices. SSD drives are designed for random workloads, so should typically be configured in a RAID 1+0 to maximize performance (unless an environment is 100% read). SAS drives are also aimed at performance. Therefore, RAID 1+0 or RAID 5 should be used. SATA drives are aimed at capacity with throughput, and due to their huge capacities should be configured in RAID 6. RAID 6 also provides additional security and peace of mind during rebuilds for backup applications where SATA drives are preferred. Note: RAID 5 may be considered when using 2TB drives or smaller. RAID 1+0 may also be considered in very high-scale virtual infrastructures of 2,000+ virtual machines. Best practices when sizing individual RAID arrays within storage systems is to keep RAID arrays between 5 and 15 disks per array. Performance will suffer in RAID arrays larger than 15 disks as stripes grow too large. Arrays with fewer than 5 disks will not have enough spindles to provide good performance. Finally, the layout of volumes/LUNs on multiple RAID sets should be considered. Deploying a simple structure on one LUN per RAID set will reduce the risk of disk-based contention. However, this is not always possible with a RAID set between 5 and 15 disks. When a one to one ratio is not possible, isolating LUN's that host similar applications and data patterns to a RAID set will buy back some performance. Consider the following scenario: A LUN used for backup and a LUN used for SQL hosted on the same RAID set. The two applications with differing data patterns contend for disk resources, requiring the disks to perform additional seeks for the scattered, random data. Again, the end result is reduced performance.

CHAPTER 2

MEASURING STORAGE PERFORMANCE

2.1 What is IOPS?

This part guides IT personnel through the process of measuring the performance of both the newer software-based storage and traditional hardware-based storage. By understanding how to measure storage performance, IT personnel will be able to better predict storage needs as they apply to the needs of the business and develop benchmarks for RFPs and product evaluations.

This part focuses on measuring the impact of the following factors on storage performance and the application of best practices in modern software-defined storage systems:

- ✓ Storage performance metrics (IOPS, throughput)
- ✓ Factors affecting storage performance (RAID penalty, READ/WRITE ratios)
- ✓ Provisioning IOPS in legacy (hardware-defined) storage solutions
- ✓ Measuring Storage Performance in ZFS-based (software-defined) storage solutions
- ✓ Best Practices for RAID and cache sizing in ZFS-based storage

The number of input/output operations a storage device can complete within one second is called Input/output operations per second (IOPS) . The performance characteristics are measured by randomly or sequentially. Depending upon the file size the random and sequential operations are done. When we are concerning large file then sequential operations are done to access of stored operation in contiguous manner otherwise random operations are done to access locations in the storage device in a non-contiguous way.

There are different characteristics for IOPS measures:

- ✓ Sequential Write IOPS: The average number of sequential write I/O operations that occur per second
- ✓ Sequential Read IOPS: The average number of sequential read I/O operations that occur per second
- ✓ Random Write IOPS: The average number of random write I/O operations that occur per second
- ✓ Random Read IOPS: The average number of random read I/O operations that occur per second
- ✓ Total IOPS: The total IOPS when performing mixed read and write operations

Frontend IOPS

Fronted IOPS is the total number of read and write operations per second generated by an application or applications.

Backend IOPS

Backend IOPS is the total number of read and write operations per second which a storage controller sends to the physical disks. This phenomenon is also known as storage IOPS.

The backend IOPS or storage IOPS is calculated by the formula below:

Storage IOPS = Number of RAID Groups × (((Read Ratio × Disk Operations/Sec) + ((Write Ratio × Disk Operations / Sec) / Write Penalty)) × Quantity of Disk in RAID Group) (1)

70% vs 30% read/write ratio for 15K SAS in a single RAID 10, the backend IOPS or storage IOPS is:

2 × (((70% x 180) + (30% × 180) / 2)) × 8 = 2, 448 IOPS

RAID Penalty

Write operation can't be completed until both the data and parity info have been written to the disk. If the any of the write operations are failed, waiting for extra time to write the parity info on to disk. This phenomenon is called RAID penalty. Because WRITEs to a disk are complete only when the data and the parity information have been fully written to the disk, extra time is required for writing the parity information. This extra time is called the "RAID penalty". It applies only to WRITE I/OS, not to READ I/Os. Measurement begins at RAID Penalty 1, which means that there is no RAID penalty. Other common examples are given in the table below:

RAID TYPE	SCENARIO	PENALTY
RAID0	Striping	There is no parity to be calculated, so there is no associated WRITE penalty. The READ penalty is 1 and the WRITE penalty is 1.
RAID1	Mirroring	The WRITE must be to the mirrored pair, so while the READ penalty is still 1, the WRITE penalty increases to 2.
RAID5	Distributed parity	This entails reading old data block, reading old parity block, writing new data block, and writing new parity block for each change to the disk, so while the READ penalty is still 1, the WRITE penalty increases to 4.
RAID6	Dual distributed parity	Now the operations involve reading data, reading parity1, reading parity2, writing data, writing parity1, and writing parity2 for each change to the disk. The READ penalty is still 1 but the WRITE penalty is now 6

Table 2.1: RAID Type

So this is the Different RAID level penalties

RAID level	Read	Write
RAID 0	1	1
RAID 1 (and 10)	1	2
RAID 5	1	4
RAID 6	1	6

IOPS Calculation

IOPS is measured by the number of I/O operations, i.e. READs and WRITEs, per second, and can be classified as follows:

- ✓ Per Disk IOPS is the rated IOPS of a single SATA/SAS/FC disk of varying RPMs.
- ✓ Frontend IOPS is the IOPS of the application, installed on storage LUN, which consumes storage. This is the IOPS classification used when talking about a requirement for 100, 200, 1,000, or 1 million IOPS.
- ✓ Backend IOPS is the IOPS required by the storage subsystem to deliver the required frontend IOPS and is dependent on RAID penalties.

The number of input/output operations is done per second. The formula of IOPS calculation is given below:

IOPS per disk = 1 / (((average read seek time + average write seek time) / 2) / 1000) + (average rotational latency / 1000)) (2)

Range of IOPS

Disk type	RPM	IOPS range
SATA	5,400	50-75
SATA	7,200	75-100
SAS/FC	10,000	100-125
SSD	N/A	5,000-10,000
SAS/FC	10,000	100-125

Following table containing the example of how calculating the per disk IOPS.

METRIC	HOW IT IS CALCULATED
Average READ seek time	Rated and published by disk vendors in data sheets and other product specifications.
Average WRITE seek time	Rated and published by disk vendors in data sheets and other product specifications
Average rotational latency	Half the time required for a rotation in milliseconds (ms). For example, 7200 RPM (120 rotations per second) translates to one rotation every 8.33 ms. Half the rotation takes 4.16 ms. Thus, the average rotational latency for a 7200 RPM drive is 4.16 ms.
IOPS per disk	1/(((average read seek time + average write seek time) / 2) / 1000) + (average rotational latency / 1000)). Below are three example calculations: For a 7200 RPM disk, per disk IOPS = 1/((((8.5+9.5)/2)/1000) + (4.16/1000)) = 1/((9/1000) + (4.16/1000)) = 1000/13.16 = 75.98. For a 10K RPM SAS/FC disk, per disk IOPS = 1/((((3.8+4.4)/2)/1000) + (2.98/1000)) = 1/((4.10/1000) + (2.98/1000)) = 1000/7.08 = 141.24 For a 15K RPM SAS/FC disk, per disk IOPS = 1/((((3.48+3.9)/2)/1000) + (2.00/1000)) = 1/((3.65/1000) + (2/1000)) = 1000/5.65 = 176.99 These examples illustrate the reason for minor variations in the rated disk IOPS from different models/vendors for the same RPM disks.

The total IOPS of the application is calculated by the following formula:

$$\text{Total IOPS} = \text{Read IOPS} + (\text{RAID level based write penalty} \times \text{Write IOPS}) \quad (3)$$

To calculate the number of disks needed to meet a front end IOPS requirement on a legacy (hardware-based) storage system, use the following equation:

$$\text{Total number of Disks required} = ((\text{Total Read IOPS} + (\text{Total Write IOPS} \times \text{RAID Penalty})) / \text{Disk Speed IOPS}) \quad (4)$$

2.2 Solid - State Disk (SSD)

Solid-state disk (SSD) devices typically offer access speeds 200 or so times faster than hard disks, so far the high price of solid-state memory has kept these devices confined to niches in the data center.

However, storage professionals are beginning to see more situations where the advantages of these high-speed storage devices, based on battery-backed DRAM or flash memory, outweigh their cost. As the technology improves and the price drops, SSD is becoming more than just a high-priced Band-Aid to be slapped on storage hot spots. SSD is now being used in applications such as transaction processing and improving storage area network (SAN) performance. According to flash drive and SSD maker Samsung, demand will grow from 2.2 million solid state disks last year to 173 million this year and 9 billion by 2010. Although the vast majority of these SSD devices will be used in laptops and consumer electronics products, demand for SSDs is growing for servers and other enterprise storage applications. As prices of flash memory and RAM continue to drop, larger solid state disks are becoming economically feasible for more storage uses. As prices drop, other advantages of SSD, such

as low power consumption and reliability, are playing a larger role in purchase decisions. Not that SSD is even close to the price per gigabyte of hard disks. For example, Violin Memory Inc. sells its 1010 appliance starter kit with 128 GB of memory for around $50,000. However, SSD manufacturers are paying more attention to usability, interoperability and management issues. Features like remote management are becoming common on solid state disks and other products are now able to work with them. For example, Microsoft's Vista is SSD-aware. Familiarity with the technology has also helped adoption. Similar products, such as USB modules, are becoming more common in business and everyday life. It seems like every IT person has at least one thumb drive tucked into their desk these days. This exposure to solid-state storage helps eliminate some of the uncertainty about SSD by extension. Finally, SSD consume less power. SSD typically uses much less power and produces less heat than hard disks. Violin Memory claims that its 1010 appliance uses less than one watt per gigabyte, about one tenth the power consumption of a typical hard disk. New SSD products are emerging to enhance enterprise storage. In July, Solid Data Systems announced a 1 TB SSD array, called Storage Spire, that uses Fibre Channel to connect the SSDs. The Storage Spire array supports up to eight 4 GB Fibre Channel connections and multipath failover for higher reliability.

Storage Spire is designed as a direct replacement for a RAID array for applications like transaction processing. Traditionally, the main use of SSD in IT has been as very high-speed caches to handle hot spots, such as indexes in databases. Solid Data claims that using the product greatly reduces the size of queues and thereby increases system stability. The prevailing wisdom is that SSD will never replace hard disk. While that's probably true, it's worth remembering that hard disk never completely replaced tape either. The real question is what the mix of SSD and hard disk will be in the storage architecture. That will most likely be determined by issues such as reliability and power consumption, as well as price.

2.3 Flash storage: What it is and how it works

Flash storage is the technology of the moment, providing high-performance random I/O capabilities far in excess of what can be achieved with mechanical hard drives. But what is going on inside a flash drive? Why are writes much more troublesome than reads in flash? Why are flash drives' lifetimes limited? And what are flash storage makers doing to overcome these issues? In this part, we look at exactly what flash storage is, how it is managed at controller level and some of the clever work that storage makers do to get the best out of solid state.

Flash deconstructed

When we talk about flash storage, we usually mean Nand flash, which is solid state memory made of millions of Nand memory gates on a silicon die. Flash technology recently reached its 30th birthday and manufacturers continue to push the boundaries of density on a single chip, which now extend into three dimensions with technology such as V-Nand from Samsung. Flash is similar to system memory in that there are no moving parts, but it has the additional property that its contents are not lost when power is turned off. Data is stored in cells, which gives us the terminology used to describe the main forms of flash, namely SLC,

MLC and TLC. SLC stands for single-level cell in which each memory cell records only a single value (of two states) – on or off, 0 or 1, based on the voltage of the cell. MLC, or multi-level cell, is capable of storing up to four states representing two bits of data – 00, 01, 10 or 11. TLC – triple-level cell – stores three bits in a single cell, using the eight states from 000 to 111. Flash devices such as solid state disks (SSDs) are Nand chips packaged with additional circuitry and firmware known as a controller, which is responsible for managing the reading and writing process, as well as other ancillary tasks.

Flash reads and writes

Cells on flash storage are arranged into pages (typically 4KB or 8KB in size) and further grouped into blocks of around 128KB to 256KB, with some checksum data. The exact size depends on the flash manufacturer and product line. The properties of Nand flash are such that a single value in a cell can be changed from "1" to "0" but not the other way around without reformatting the entire block, a process known as a program-erase (P/E) cycle. As a result, writing data to flash in place requires the reading of an entire block from flash and into the memory of the controller, updating it with new data, erasing the existing block and writing the data back to the flash device. This inefficient multi-stage process is known as write amplification, where each writes operation to flash requires more than one physical write I/O.

Write amplification is a problem for flash devices because Nand chips are degraded slightly with every write operation and so devices have a finite number of P/E cycles. SLC Nand has a P/E cycle count of around 100,000 per block, but MLC can be as low as 5,000 per block of data. The finite lifetime of flash means that writing data back in place repeatedly (for example, a file or database column rewritten multiple times) can very quickly result in a device failure. For this reason, flash drive manufacturers have employed techniques in the controller to mitigate the shortcomings of flash lifetime.

Wear levelling

Wear levelling is one technique that flash drive manufacturers use to improve device endurance or lifetime. Rather than storing data in the same place, wear levelling distributes write I/O blocks across the entire flash device, always writing to a new empty page. The result is more even wear across all Nand cells and increased device lifetime. In addition to MTBF (mean time between failures), manufacturers also quote a figure known as DWPD (device/drive writes per day), which provides a measure of how many complete drive writes can be sustained over a fixed period (usually three to five years) before the device can be expected to fail. DWPD figures vary greatly, from less than one to as high as 50, depending on whether the device is for the consumer or enterprise market. Naturally, devices with higher endurance attract a higher price.

Value in the controller

Controller circuitry and firmware performs the task of managing I/O back and forth from the Nand chips. Flash drive suppliers have invested significantly in optimising the firmware to work with Nand to deliver improved product lifetimes.

Garbage collection

As we have seen, flash device architectures store data in pages, which are grouped together in blocks for P/E cycles. As techniques such as wear levelling distribute write I/O across an entire device, blocks start to fill up with pages of both in-use (or valid) and invalidated data that has been moved elsewhere in the device. To re-use these invalidated pages, the entire block must be erased. A process called garbage collection manages the movement and consolidation of valid pages between blocks, allowing an entire block to be erased for subsequent re-use. The effectiveness of the garbage collection process can have a direct effect on the performance of flash. When data is initially written to an SSD, the contents are placed on empty or partially filled blocks and very fast write times result. But, at some point, the controller needs to start reclaiming pages for re-use and when this occurs, devices can experience a dip in performance, sometimes called the "write cliff". The quality of algorithms used to perform garbage collection has a direct impact on performance – yet again demonstrating the importance of controller features.

Cutting writes times with Trim

As we have seen, all issues with flash occur when writing to the device. So, if you can cut down on the processes involved in write I/O, it can improve device performance and lifetime. One technique used to avoid write I/O is Trim. This allows the operating system (OS) to flag blocks of data that have been released from the local file system and to begin the erase process before the next write occurs.

Normally, reads and writes occur at page level, but deletes can only occur at the (larger) block level. In the normal write process, deletes occur at block level, but Trim allows the erase part of the P/E cycle to occur earlier. Trim is supported by the major OSs and by the SCSI protocol as the Unmap command, which, in turn, is supported by the major hypervisor suppliers.

Supplier implementations

Flash devices have very different characteristics to hard drives. As a result, array suppliers have had to either develop new architectures designed around flash, or modify existing products to deal with flash drives. Techniques include reading and writing in block sizes to match the drive being used, as seen in EMC XtremIO and HP's 3PAR Store Server systems. Meanwhile, Hitachi Data Systems (HDS) designed its own flash module, which consolidates management functions into a custom controller, rather than using commodity SSDs. In a similar way, Violin Memory implements system-level wear levelling across all custom

modules in its system, rather than on each drive. Some of these flash benefits are implemented in hardware, but typically innovations are achieved through architectural design and software. This should come as no surprise, as increasingly we are seeing storage move towards a software-defined world.

2.4 SAS SSD (Serial-Attached SCSI Solid-State Drive)

A serial-attached SCSI (SAS) solid-state drive (SSD) is a NAND flash memory-based storage or caching device designed to fit in the same slot as a hard disk drive (HDD) and use the SAS interface to connect to the host computer. The most common drive form factors for a SAS SSD are 2.5-inch and 3.5-inch. SAS SSD bandwidth options include 3 Gbps, 6 Gbps and 12 Gbps. A SAS SSD offers faster data transfer rates than a serial ATA (SATA) SSD. In contrast to a SATA SSD, a SAS SSD also supports dual-port operation and builds in features to improve reliability such as advanced error correction and data integrity technology and high signal quality on the cable or backplane. SAS SSDs are generally more expensive than SATA SSDs. They are primarily used in enterprise servers and storage arrays with application workloads requiring high availability, high input/output (I/O) and low latency. Use cases for SAS SSDs include server virtualization, online transaction processing, high-performance computing and data analytics. Drive manufacturers sometimes offer SAS SSDs with different write endurance options. For instance, a high-capacity SAS SSD intended for read-intensive workloads might guarantee only one drive write per day (DWPD), while a lower-capacity SAS SSD intended for write-intensive workloads might support up to 25 DWPD. SAS is a more robust interface than SATA, with dual ports and end-to-end data integrity. SATA is a cheaper interface. SAS SSDs are gaining in popularity because of that robustness and a narrowing of the price differences.

Array-based SSD - An array-based SSD is a solid state drive manufactured in a form factor that can be installed in a typical storage array. SSD arrays are designed to match up with typical hard-disk drive form factors – 3.5 inches, 2.5 inches, or 1.8 inches. 3.5 inch and 2.5 inch SSDs are the most common. These drives are typically NAND flash-based. The most common interfaces for a solid state drive array in traditional hard-disk form factors are: Serial ATA (SATA) and Serial-Attached SCSI (SAS). The random access time of an SSD storage array is about. 1 millisecond, compared to 5 to 10 milliseconds for a hard disk drive. SSD arrays are used in applications that demand increased performance with high input/output (I/O). It is often the top tier in an automated storage tiering approach. Because automated storage tiering decides where to move data based primarily on input/output activity, it does not prioritize those choices based on the individual application. That allows an SSD array to provide improved data access in an easy-to-use package.

2.5 RAID with SSD

RAID seems to be a storage fixture but it was only formally defined in 1987. The major objectives of RAID were always to address the lack of hard-disk-drive (HDD) reliability by improving data availability and to drive up the performance of HDD systems. While RAID is still a de facto storage standard, there is still a question of whether it is an optimum approach – even for HDD systems – because almost all levels of RAID require an overhead to provide the protection. For solid-state storage, there are even more questions about RAID's relevance. This tip will explore the basics of using SSD and RAID together and offer key advice for RAID, SSD and your storage environment. Solid-state storage gives users loads of performance, so RAID's performance enhancements are moot. That puts the focus on data availability and protection. Many flash chips have basic RAID built-in to increase redundancy and longevity; the question is whether more system RAID on top of that helps. Whether your solid-state storage is used as a tier or as a cache is a key consideration: Many vendors' implementations require confirmation from a lower tier of spinning disks before confirming the write. And most cache – aside from read-only – is unlikely to offer immediate data protection.

How RAID works with SSD

If SSDs merely replace some HDDs in a system, then the same RAID can be applied. You'll invariably need the same RAID to be applied within and across RAID groups to allow tiering and movement flexibility. For most high-end systems, RAID 5 or RAID 10 are likely sufficient, but for added security, RAID 6 (double parity) is probably preferable. Of course, the raw reliability of solid-state is also relevant: Many solid-state vendors are now claiming "x" years for a given number of full daily writes that are often considerably better than equivalent figures for HDDs.

Why you might use RAID with SSD

There are many reasons why you might use RAID with SSDs. A PCI solid-state storage card in a server is a popular approach to boost application and storage performance; however, it's effectively a DAS model, which translates to a single point of failure. To help protect against losing data, a simple RAID 1 (using a mirrored flash card) model might be appropriate, albeit expensive. Otherwise, such cards are often implemented as read-only cache so the protection is performed more economically at the HDD level. This all comes back to knowing what you want to achieve regarding the balance of performance and data availability/protection.

Specialized solid-state RAID hardware/software

New and updated RAID controllers are emerging that allow storage systems to use more of the massive performance boost that solid-state storage can provide, as a few SSDs can quickly overrun the capabilities of traditional controllers. The main focus of these new controllers is simply to allow more SSD performance to flow to the server and applications rather than a revamping of RAID. However, it's important to note that, with the new breed of

purpose-built all-flash and hybrid flash/HDD arrays, some vendors are implementing a special version of RAID that's optimized for their particular SSD implementations.

Economics and mixing HDDs and SSDs in RAID arrays

Having established that RAID is – by definition – all about redundancy, the simple truth is that some amount of copied data and/or parity data has to be stored somewhere. The extent and placement of that "extra" data depends ultimately on economics. If money was no object, you'd have multiple copies on multiple solid-state devices. Realistically, that's not economically feasible, so a hybrid approach makes more sense for using RAID with SSD. Prospective users of solid-state storage should give careful considerations to the options offered by various vendors. A hybrid or mixed approach can make sense in the array (although it requires careful pool management) but it doesn't make sense in a RAID group because the performance and reliability characteristics are so different that you'd lose the solid-state advantage in the process. Some of the emerging vendors have sophisticated firmware to fully use the solid state while placing the redundancy data on lower cost media, which can save in the range of 10% to 50% on overall costs (compared to traditional RAID on flash alone) for well-protected, highly available data.

2.6 Solid-State Storage Choices

Performance is the driver for the vast majority of companies considering flash storage arrays. While there are some other benefits, most companies just need faster storage. Some common use cases for flash storage include high-density, virtual server infrastructures, VDI, high-transaction databases and web-facing applications. When shared storage is required (as opposed to putting flash directly into the application server), there are two options available,

a) All-flash arrays (AFA) and
b) Hybrid flash arrays

ALL-FLASH ARRAYS

As the name implies, all-flash arrays are 100% flash storage systems. Some use drive form factor solid-state drives (SSDs) and others use custom flash modules to populate the array chassis. They are available in scale-up and scale-out architectures using both proprietary and commodity hardware nodes. AFAs support either file, block or object storage protocols, and some actually provide a 'unified' solution that supports multiple protocols. Today, most AFAs offer fairly complete services to support data protection, data handling, efficiency, etc. Early on, many all-flash arrays lacked these features. Similarly, storage management feature sets have evolved, with most AFAs providing an administrative experience similar to traditional storage systems. The capacities of AFAs today range from a few tens of terabytes to multiple petabytes, and data reduction technologies like Deduplication can be especially effective in AFAs. Flash as a storage medium costs less to operate than hard disk-based systems. It consumes less power, creates less heat (requiring less cooling) and takes up less physical space in the data center. AFAs, since their storage media is homogeneous, don't require any complex decision-making or data movement. This gives them more consistent

performance and improved scalability, factors that can make them a better fit for large, multi-tenant environments.

HYBRID ARRAYS

Hybrid arrays combine flash, usually in 2.5" form-factor drives, with hard disk drives (HDD) to lower the effective cost and increase effective capacity. They also come in scale-up and scale-out architectures, using purpose built or commodity hardware to provide combined raw capacities (disk and flash) larger than most AFAs.

Current hybrid offerings support block, file and object-based protocols, with unified systems available as well. Storage services and management features are similar to traditional arrays, which makes the switch to hybrid flash easy from an operational perspective. Since most performance requirements are temporary, the storage system can move particular data objects to flash when the compute process needs them and back out to the HDDs when it doesn't. This creates a 'multiplier effect' that enables a smaller amount of flash to accelerate a much larger total data set. Hybrid arrays today use caching or tiering to accomplish this data movement.

2.7 Flash Caching and Tierring

Read caching

Involves keeping a copy of the most frequently accessed data objects in flash so that read requests can be fulfilled without incurring hard disk latencies. Cache capacity is at a premium so, obviously, the better a hybrid system is at keeping the right data in cache, overall performance will improve and more application data will be accelerated. While supporting read transactions is the most common use of flash in hybrid arrays, most also use a cache to accelerate write operations.

Write caching

Involve storing written data in flash first. While it acknowledged written transaction to the host, then it copy to HDDs. Since all data must eventually be copied to hard disk storage, the array must have a large enough write cache (meaning less flash capacity is available for reads) or enough non write time to allow the cache to empty; otherwise, write performance will suffer. Instead of creating a second copy of data in a cache, flash tiering moves 'hot' data objects out of the hard disk area and into flash to support periods of maximum activity. Ideally, all read and write activity is performed in flash. Eventually, data will be copied back into the HDD tier, a process that can be done manually or based on policies, as it is with caching. Speed is critical for the applications that typically drive flash usage and those applications (and users) often get accustomed to flash performance. When unexpected demands cause a cache or tier miss, applications must read data from disk drives. And the drives that are typically used in hybrid arrays are often very slow with high capacities. As a result, latency can cause unacceptable wait times for users, slow online transactions and

bottlenecks in other production applications, among others. For this reason, Workload predictability is the key to the effective use of flash in a hybrid array.

How to choose

Certain use cases don't work well with either AFAs or hybrid arrays so the first step is to identify if there are any conditions in your environment or workloads that make one of these two options a bad fit. For AFAs, the most obvious factor is capacity required and its effective cost. If the application's current or expected data set is too large for available flash or the budget is too small to buy more, then an AFA is not an option. Look at the 'effective capacity' of the flash system, after data reduction, as well as the raw capacity when making your decision. For environments that need 100% consistency and no chance of a cache or tier miss, an AFA is probably the better option. These include the use cases that all-flash storage was first developed for in the financial, internet based and high-performance computing industries. AFA efficiency advantages also make these systems effective for multi-tenant cloud environments that need low overhead and predictable performance as they scale.

AFAs are more attractive when IT can't make the assumptions required to support data movement in a hybrid array. But even when this isn't the case, for many companies the simplicity of all-flash still wins out. If they can afford enough flash capacity to support the applications that need performance they buy an AFA. If not, they buy a hybrid. It is important to realize that if workloads can handle the occasional cache or tier miss, hybrids offer the best economics, by far, allowing more workloads to be accelerated for a given investment. It's also reasonable to assume that performance will improve as users get more familiar with their applications' storage demands and caching parameters are fine-tuned. Since hybrids include high-capacity disk drives, they also provide better and more cost-effective scalability, although at HDD performance levels. In reality, most people under-buy the flash for a hybrid array because they simply don't know how much flash they need or they're more focused on saving money than improving performance. Hybrid vendors are also guilty of under-selling flash because it makes their cost advantages over AFAs more dramatic. By and large, statistics show that a 5% flash to hard disk capacity is typical for a new hybrid configuration, but so is a cache hit rate of two-thirds, meaning one out of three transactions aren't served from cache. In most environments, moving the flash investment to 10% of total capacity can practically eliminate cache misses.

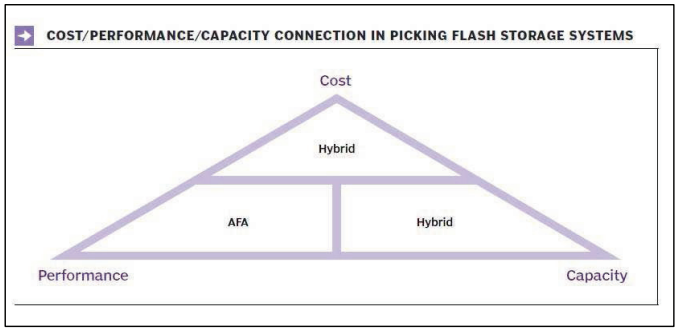


Figure 2.2: Cost/Performance/Capacity Connection in Picking Flash Storage Systems

2.8 Difference between 'usable' and 'effective' flash capacity

Usable flash capacity, which is the capacity before Deduplication and Compression is applied. The effective flash capacity is the array's capacity and compression turned on. Many vendors claim a 4:1 or 5:1data reduction ratio for their systems -- based on typical use. I think that's primarily because of the type of environments they've gone in to. For example, databases have a lot of redundant data, and as such, are very conducive to data reduction. If you are using an array for general purpose storage for a variety of types of data, not just databases, you're likely to see somewhere in the neighbourhood of 3:1 data reduction. So, if I was planning, I would plan for 3:1 even if the vendor says they're able to get 5:1. That 3:1 ratio is a very conservative number, and typically the people who use the systems are very conservative. And so, that's a practical number work with. Many vendors offer an absolute guarantee of at least 2:1. So that's great, take advantage of it.

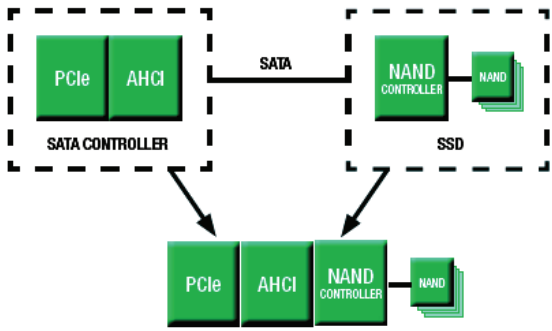
2.9 SSD Interface Comparisons

PCI Express (PCIe) is a general purpose bus interface used both in client and enterprise compute applications. Existing mass storage interfaces (SATA, SAS) connect to the host computer through host adapters that in turn connect to the PCIe interface. The SATA interface was designed as a hard disk drive (HDD) interface, and the SAS interface was designed as both a device interface and a storage subsystem interface/infrastructure. As HDDs and system requirements have evolved, requiring faster interfaces and new features, the SATA and SAS interfaces have gone through several revisions. Solid state drives (SSDs) have quickly added significant new performance requirements to these interfaces, as the data rates of SSDs have gone from tens of MB/sec, to hundreds, and now thousands of MB/sec. In addition to the increase in data rates, the lack of mechanical movement in SSDs have also increased the number of input and output operations per second (IOPS) that these storage

devices can perform. This development has created a need for improved implementations of the existing standards, as well as enhancements to existing interface standards, to manage the new performance requirements while keeping compatibility with existing system architecture. This part discusses the different interfaces and contrasts the various performance and compatibility trade-offs encountered.

SATA Interface

SATA is a low-cost interface designed for point-to-point connection either through a cable or printed circuit board (PCB) trace. The host connection is to an advanced host controller interface (AHCI), which usually resides in the host chipset as the host adapter on the PCIe bus. There are some design issues with this interface that can create a bus overhead of 1µs (or more) for each command. This is not a major issue for HDDs where a 4KB transfer is in the order of 10µs, but SSDs can transfer 4KB of data in 2µs (or less)—thus the overhead becomes significant and the SATA interface less interesting as a high-performance mass storage interface. SATA is still suitable as a low-cost SSD interface where cost, not performance, is the major decision factor. The SATA architecture can also be consolidated into a host adapter that manages the SATA command-set without actually including the physical SATA interface (PHY).



Source: Seagate Technology, 2011

Figure 2.3 Architecture Consolidation

SAS Interface

SAS is also a serial interface, attached to the host through a host adapter, but there are significant differences that make it suitable as an SSD interface:

- Less hardware overhead
- Faster transfer rates
- Wide ports
- Efficient driver-controller interfaces

In addition, SAS includes features not found in SATA that improve reliability and availability of devices connected to the interface:

- Robust serial protocol
- Multiple host support
- End-to-end data integrity
- Dual-port capability
- High degrees of concurrency and aggregation

Less Hardware Overhead - There is not a universal host interface for SAS that would be equivalent to the SATA AHCI controller. Instead, multiple vendors compete in the SAS host adapter market where performance is a major factor—not only to interface individual HDDs but also various RAID systems where the transfer rates of multiple HDD spindles are aggregated for improved transfer speed. Additionally, SAS host adapters are designed to manage higher-performance SSDs and HDDs (such as short-stroked 15K-RPM drives). Since the hardware host adapter and the device driver managing that host adapter are designed as a system, new designs optimized for SSDs are starting to become available and further improve not only transfer rates but also the IOPS.

Faster Transfer Rates - SAS ports currently support up to 6 GB/s data rates. Companies such as LSI and PMC-Sierra are sampling designs currently in development to support 12Gb/s data rates and greater than 2 million IOPS, with the possibility of 24Gb/s in the future.

Wide Ports - Inherent in the SAS architecture is the concept of wide ports—where multiple links can be aggregated to allow multiple, simultaneous paths between one or more hosts and a device. The current SAS drive connector defines two ports for the drive. As a design choice, current HDDs do not support wide port—only dual port, where each port has a different SAS address that prevents configuration as a wide port. Accepted proposals for SAS-3 (12 GB/s) allow an increase in the number of ports on the drive to four, all of which could connect to the same domain, or in pairs to different domains. A very limited number of SSDs can support wide port in addition to dual port on a two-port device.

Robust Serial Protocol - The SAS serial protocol provides for training of the serial transmitters and receivers. This improves the signal quality on the cable or the backplane by compensating for channel length, impedance mismatch and inter-symbol interference. The

SAS serial protocol also manages error detection and retransmission at the hardware level. This allows for faster recovery from intermittent signalling issues.

Multiple Host Support - The SAS interface and switching fabric allow multiple hosts to access the same device. This feature can be used to manage host failures, as well as data path failures for improved data availability.

End-to-End Data Integrity - The SAS interface can verify data integrity through cyclic redundancy checks (CRC) of the data from the time it is created in the host data buffer, through the transfer across the PCIe interface and the SAS interface until it is stored on the device and again read and transferred to the host data buffer. This allows for multiple checkpoints along the path from applications through RAID controllers and at devices. This feature is sometimes called protection information (PI).

Dual-Port Capability - SAS target devices support dual-port operation. This provides the ability to create two fault domains and provides increased availability. Even if a failure occurs in one of the paths to a port preventing access along that path, a device is still accessible using the second port. Historically, Seagate has driven interface adoption in the market. Seagate is working with the SCSI Trade Association (STA) and other industry leaders to leverage the widely deployed, existing SAS infrastructure.

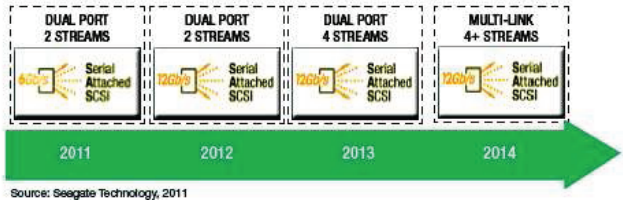


Figure 2.4: SAS Interface Evolution

PCI Express Interface

PCI Express (PCIe) is the fundamental interface that connects peripheral devices to the host processor and through a memory controller to the memory architecture in the system. Both the SATA and SAS interfaces discussed earlier connect through a PCIe interface (or host adapter) to the host processor and memory.

Multiple Links (BW)	X4 (4 × 600MB/s)
Power Available	25W (2.5 inch)
Total Latency	Very low
Multi-Host Protocol	Yes
High Availability	Yes (Dual Port)
Scalability	Excellent
Robust, Proven Protocol Stack	Yes
Hot Swap Serviceable	Yes
Compatible with Existing Management SW	Yes

Source: Seagate Technology, 2011

The PCIe interface is a serial implementation of the original PCI interface that provided a parallel address/data connection between peripherals and host processor/memory. The PCIe interface communicates over one or more lanes that consist of one transmit and one receive serial interface for each lane. Up to 32 lanes can be used to connect a host to a device. The serial data rate on each lane depends on the version of the PCIe standard implemented, the current version is 3.0 and the data rate is approximately 1GB/s. For a 1U server, the PCIe interface is designed to utilize a single connector on a (client) motherboard, or a two-connector, right-angle adapter on a (server) motherboard. A cabling system is also available (though seldom used). A 2U, 4U or 7U server has many more PCIe slots, similar to client implementations. The PCIe specification also uses transmitter (and receiver) training to adapt to the impedance variations of a configuration, but is targeted as shorter length transmission channels than SAS. PCIe switches can accommodate single root I/O virtualization (SR-IOV) and multi root I/O virtualization (MR-IOV)—methods used to improve controller performance in virtual (hypervisor) systems with a single or multiple hosts. SR-IOV is just starting to become generally available in adapters; however, VMware may not yet take advantage of it. MR-IOV is typically not supported on adapters.

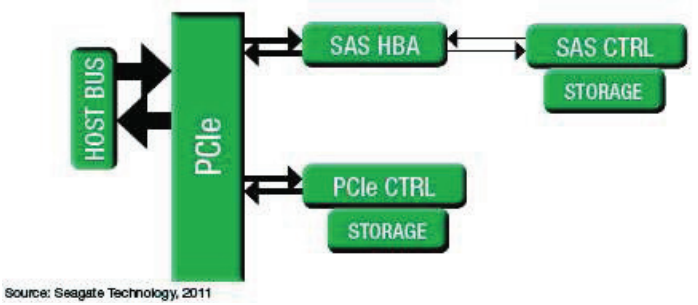


Figure 2.5: SAS Interface Evolution

Storage devices that connect using the PCIe interface do so either through a direct register connection, or through a host adapter that then connects to the device through additional cabling or a backplane-type interface. Currently there are a number of different implementations of both architectures. SATA uses a host bus adapter implementation in the system chipset (south bridge)—either the Intel or AMD AHCI—requiring different AHCI drivers but mapping to compatible IDE legacy implementations. These interfaces also implement various RAID management features. SAS has multiple vendors of HBAs, with additional expanders and RAID controllers available, all using proprietary device drivers and BIOS to satisfy various needs for performance and configurability. The PCIe driver-controller interface is implemented in the NVM Express specification and in the proposed SCSI over PCIe (SOP) specification. The SATA consolidated architecture described above is another example of the PCIe direct register connection.

PCIe SSDs Today - There are two primary types of PCIe SSDs in the market today: the native and the aggregator. The native controller attaches to the host PCIe bus and then directly controls multiple flash memory buses. These typically use a software interface that is proprietary to the manufacturer and used only for the specific device. Some of these implementations place the burden of address translation and other functions on the host CPU and memory. This in turn causes reduced system resources for applications when the devices are used under heavy workloads. Additionally, being relatively new to the marketplace, these unique drives and hardware combinations are sometimes prone to instabilities, as their ecosystems are still evolving. The aggregator model takes a different approach to design. This approach utilizes an existing SAS or SATA RAID controller, to which are attached multiple SAS or SATA SSDs. These are packaged together on a single PCIe card. The RAID controller aggregates the performance of multiple devices to offer high levels of performance. Being based on existing proven enterprise class hardware and software interfaces; these designs are very stable and mature. Additionally, these designs use intelligent controllers that perform address translations and other functions, allowing full use of system CPU cycles and memory by applications, even under heavy I/O workloads.

The Future of PCIe SSDs - Both SOP and NVMe approaches are architecturally similar. However, NVMe is being developed in an industry working group, whereas SOP is being developed in a recognized open standards forum. NVMe is targeted only at use for non-volatile memory devices, while SOP is also being targeted at use for host bus adapters and RAID controllers with features for bridging between various SOP devices. Additionally, SOP heavily leverages existing industry architectures and features, while NVMe uses a new, very limited instruction set and queuing interface.

Interface Benefits and Issues - Each of the storage architectures described has benefits as well as issues. Depending on the overall system design, the benefits of using a specific architecture may outweigh the issues associated with that architecture, and a careful analysis is required to make the appropriate decision. That decision must also include consideration of compatibility with an existing system design. For example, updating a laptop computer system that has an existing 2.5-inch SATA HDD with an SSD would only work with an SSD of the same physical size and with the same (or newer) SATA interface. There will be a limit

on how fast the SSD can be in this case; exceeding the existing host SATA interface speed will not add to the performance of the system. In a similar situation, an enterprise server that is using a short-stroked, 15K-RPM SAS HDD to store a database index can be upgraded using a SAS SSD, which will increase overall system performance, but only to the degree that some other system factor becomes the new bottleneck (CPU, memory, network, adapters, etc.). In a new system architecture, the addition of solid state storage can significantly increase system performance, but only to the extent that the rest of the system architecture can accommodate the increased data rate and data bandwidth. Faster data rates in SSDs also require more power supplied to the device and more heat dissipation required wherever the SSD is mounted.

Another factor is the timing for availability of the operating system device drivers and BIOS support for these new SSD interfaces, as well as the initial reliability of the software.

Interfaces and Flash SSD Latency Facts - There are many misconceptions about what factors add latency and how much they actually affect application performance. When looking at this aspect, it is important to focus on the overall picture, not just one part of it. The overwhelming contributors to latency in SSDs are the flash parts themselves. SLC access times are 25µs+; MLC access times are 50µs+, both assuming no access contention. As queue depths increase, the contention for access to the flash parts can add substantially to latency. Once a flash part starts its access, other requests to the same part must wait. As many as eight flash die share a common bus, which cause die to wait their turn using the bus. Housekeeping activities add additional latency (address translation, garbage collection, wear levelling, etc.). Another aspect is the operating system, which adds latency regardless of the access protocol and interconnects. These include the file system, volume manager, class drivers and context switching overheads. Differences in protocols and interconnects have negligible effects on latency as seen by an application (fractions of a microsecond).

2.10 Quality of service

QoS is a critical enabling technology for enterprise and service providers that want to deliver consistent primary storage performance to business-critical applications in enterprise infrastructure. The type of applications that require primary storage services typically demand greater levels of performance than what is readily available from traditional storage infrastructures today. However, simply providing raw performance is often not the only objective in these use cases. For a broad range of business-critical applications, consistent and predictable performances are the more important metrics. Unfortunately, neither is easily achievable within traditional storage arrays. There is a large imbalance today between the performance and capacity resources within traditional Storage systems. Capacity is plentiful and low cost; conversely, input/output per second (IOPS) are scarce and very expensive. From a provisioning perspective, performance and capacity are rigidly bound together, which only makes matters worse. This bind forces administrators to unnecessarily add storage capacity to increase the amount of IOPS available to a particular application. What results is a wasteful allocation of resources in an effort to overcome the limitations of existing storage architectures. For service providers and enterprise IT, the promise of delivering storage

resources predictably to a broad set of applications without worry has been nothing more than a pipe dream.

The history of QoS

QoS features exist in everything from network devices, to hypervisors, to storage. When multiple workloads share a limited resource, QoS helps provide control over how that resource is shared and prevents the noisiest neighbour (application) from disrupting the performance of all the other applications on the same system. In networking, QoS is an important part of allowing real time protocols such as VoIP to share links with other less latency-sensitive traffic. Hypervisors provide both hard and soft QoS by controlling access to many resources including CPU, memory, and network. QoS in storage is less common. If you seek out QoS within the storage ecosystem you will find that most approaches to storage QoS are "soft" – that is, based on simple prioritization of volumes rather than hard guarantees around Performance. Soft QoS features like rate limiting, prioritization, and tiering, are effective only as long as the scope of the problem remains small. When storage is deployed at scale these soft techniques quickly fail. In fact, these features are all "bolt-on" technologies that attempt to overcome limitations in storage architectures that were never designed to deliver QoS in the first place.

QoS: A critical component of the next generation data center

Quick look across today's storage landscape shows systems with a broad range of capacity and performance resources. On one end of the spectrum, disk-based systems have a high level of capacity and low level of performance. On the other end, flash architectures deliver a very high level of performance while requiring significantly less capacity (and at much higher cost). When viewed from the application perspective, the reality is that most application performance requirements fall somewhere in the middle of these two storage extremes. In order to meet varying application performance requirements, the storage industry has responded by implementing caching or tiering schemes in front of traditional disk-based systems. These schemes apply complex algorithms and predictive methodologies that shuffle data to the right media at the right time to boost performance. Costly, complex, and reactive, this approach does little to bring you closer to the predictable performance required by mission-critical applications. Solving for this disparity requires a more balanced pool of capacity and performance at the system level. From this starting point, a storage system can then deliver performance and capacity scaled independently to serve the unique needs of different applications. This ability to finely allocate capacity and performance resources separately from one to another is a fundamental component of next generation data centres. In these next generation infrastructures raw storage performance is important, but it is the predictable and consistent delivery of that performance which ensures every application has the resources required to run without variance or interruption. In servicing these workloads, IT must consider how well the underlying storage architecture will endure the following conditions:

- ✓ Unpredictable I/O patterns

- ✓ Noisy neighbour applications
- ✓ Constantly changing workload and application performance requirements
- ✓ Deduplication, compression, and thin provisioning processes
- ✓ Scaling of performance and capacity resources on demand

2.11 The Six Requirements for QoS

Adding QoS features to an existing storage platform may solve one performance bottleneck for individual performance conditions, but this approach fails to solve the exponentially larger challenges that occur at cloud scale. A true solution requires a purpose-built storage architecture that solves performance problems comprehensively, not individually. In this part, we'll dive into greater detail around each of the six required components and capabilities of an IT infrastructure that can enable QoS.

Requirement #1: All-SSD architecture

What it enables - Delivery of consistent latency for every I/O. Anyone deploying either a large public or private cloud infrastructure is faced with the same issue: how to deal with inconsistent and unpredictable application performance among apps running simultaneously. The first requirement for achieving this level of performance is moving from spinning media to an all-SSD, or all-flash, architecture. Only all-SSD architecture allows you to deliver consistent latency for every I/O.

At first, this idea might seem like overkill. If you don't actually need the performance of SSD storage, why can't you guarantee performance using spinning disk? Or even a hybrid disk and SSD approach? Fundamentally, it comes down to simple physics. A spinning disk can only serve a single I/O at a time, and any seek between I/Os adds significant latency. In cloud environments where multiple applications or virtual machines share disks, the unpredictable queue of I/O to the single head can easily result in orders of magnitude variance in latency, from 5 ms with no contention to 50 ms or more on a busy disk. All-flash architecture is just the starting point for guaranteed QoS, however. Even a fast flash storage system can have noisy neighbours, degraded performance from failures, or unbalanced performance.

Requirement #2: True Scale-Out architecture

What it enables - Linear, predictable performance gains as system scales Traditional storage architectures follow a scale-up model, where a controller or pair of controllers is attached to a set of disk shelves. More capacity can be added by simply adding shelves, but controller resources can only be upgraded by moving to the next "larger" controller (often with a data migration). Once you've maxed out the biggest controller, the only option is to deploy more storage systems, increasing the management burden and operational costs. This scale-up model poses significant challenges to guaranteeing consistent performance to individual applications. As more disk shelves and applications are added to the system, contention for controller resources increases, causing decreased performance as the system scales. While adding disk spindles is typically seen as increasing system performance, many storage

architectures only put new volumes on the added disks, or require manual migration. Mixing disks with varying Capacities and performance characteristics (such as SATA and SSD) makes it even more difficult to predict how much performance will be gained, particularly when the controller itself can quickly become the bottleneck.

Scaling out is the only way to go - By comparison, a true-scale out architecture adds controller resources and storage capacity together. Each time capacity is increased and more applications are added, a consistent amount of performance is added as well. Scale-out architecture ensures the added performance is available for any volume in the system, not just new data. This solution is critical for both the administrator's planning ability as well as for the storage system itself. If the storage system itself can't predict how much performance it has now or will have in the future, it can't possibly offer any kind of guaranteed QoS.

Requirement #3: RAID-less data protection

What it enables - Predictable performance in any failure condition The invention of RAID 30+ years ago was a major advance in data protection, allowing "inexpensive" disks to store redundant copies of data, rebuilding onto a new disk when a failure occurred. RAID has advanced over the years with multiple approaches and parity schemes to try and maintain relevance as disk capacities have increased dramatically. Some form of RAID is used on virtually all enterprise storage systems today. However, the problems with traditional RAID can no longer be glossed over, particularly when you want a storage architecture that can guarantee performance even when failures occur.

The problem with RAID - When it comes to QoS, RAID causes a significant performance penalty when a disk fails — often 50% or more. This penalty occurs because a failure causes a two- to five-times increase in I/O load to the remaining disks. In a simple RAID10 setup, a mirrored disk now has to serve double the I/O load, plus the additional load of a full disk read to rebuild into a spare. The impact is even greater for parity-based schemes like RAID5 and RAID6, where a read that would have hit a single disk now has to hit every disk in the RAID set to rebuild the original data (in addition to the load from reading every disk to rebuild into a spare).

The performance impact from RAID rebuilds becomes compounded with long rebuild times incurred by multi-terabyte drives. Since traditional RAID rebuilds entirely into a new spare drive, there is a massive bottleneck of the write speed of that single drive combined with the read bottleneck of the few other drives in the RAID set. Rebuild times of 24 hours or more are now common, and the performance impact is felt the entire time. How can you possibly meet a performance SLA when a single disk failure can lead to hours or days of degraded performance? In a cloud environment, telling the customer "the RAID array is rebuilding from a failure" is of little comfort. The only option available is to dramatically under-

provision the performance of the system and hope the impact of RAID rebuilds goes unnoticed.

Requirement #4: Balanced load distribution

What it enables - Eliminates hot spots that create unpredictable I/O latency Most block storage architectures use very basic algorithms to lay out provisioned space. Data is striped across a set of disks in a RAID set, or possibly across multiple RAID sets in a storage pool. For systems that support thin provisioning, the placement may be done via smaller chunks or extents rather than on the entire volume at once. Typically, however, at least several hundred megabytes of data will be striped together. Once data is placed on a disk, it is seldom moved (except possibly in tiering systems to move to a new tier). Even when a drive fails, all its data is simply restored onto a spare. When new drive shelves are added they are typically used for new data only, not to rebalance the load from existing volumes. Wide striping is one attempt to deal with this imbalance, by simply spreading a single volume across many disks. But when combined with spinning disk, wide striping increases the number of applications affected when a hotspot or failure does occur.

Unbalanced loads cause unbalanced performance - The result of this static data placement is uneven load distribution between storage pools, RAID sets, and individual disks. When the storage pools have different capacity or different types of drives (e.g. SATA, SAS, or SSD) the difference can be even more acute. Some drives and RAID sets will get maxed out while others are relatively idle. Managing data placement to effectively balance I/O load as well as capacity distribution is left to the storage administrator, often working with Microsoft Excel spread sheets to try and figure out the best location for any particular volume. Not only does this manual management model not scale to cloud environments, it just isn't viable when storage administrators have little or no visibility to the underlying application, or when application owners cannot see the underlying infrastructure. The unbalanced distribution of load also makes it impossible for the storage system itself to make any guarantees about performance. If the system can't even balance the I/O load it has, how can it guarantee QoS to an individual application as that load changes over time?

Requirement #5: Fine-grain QoS control

What it enables - Complete elimination of noisy neighbours and guaranteed volume performance. Another key requirement for guaranteeing QoS is a fine grain control model that describes performance in all situations. Contrast fine-grain control against today's rudimentary approaches to QoS, such as rate limiting and prioritization. These features merely provide a limited amount of control and don't enable specific performance in all situations.

The trouble with having no control - For example, basic rate limiting, which sets a cap on the IOPS or bandwidth an application consumes, doesn't take into account the fact that most storage workloads are prone to performance bursts. Database checkpoints, table scans, page cache flushes, file copies, and other operations tend to occur suddenly, requiring a sharp

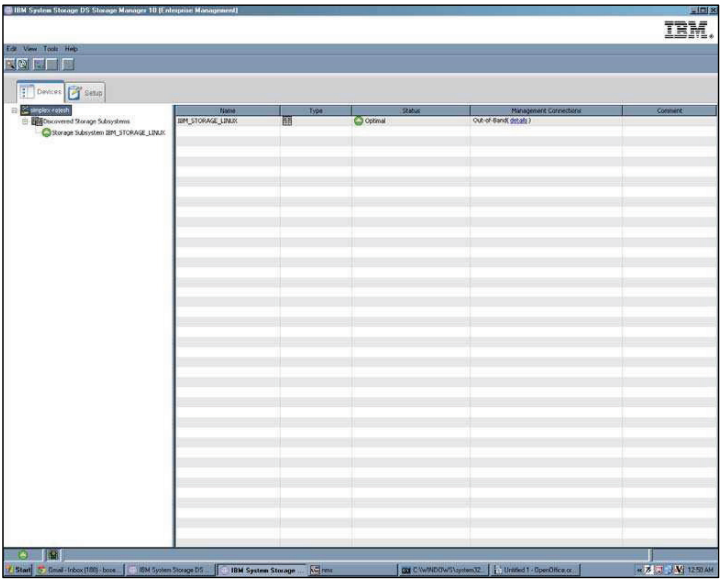
increase in the amount of performance needed from the system. Setting a hard cap simply means that when an application actually does need to do I/O, it is quickly throttled. Latency then spikes, and the storage seems painfully slow, even though the application isn't doing that much I/O overall. Prioritization assigns labels to each workload, yet similarly suffers with burst applications. While high priority workloads may be able to easily burst by stealing resources from lower priority ones, moderate or low priority workloads may not be able to burst at all. Worse, these lower priority workloads are constantly being impacted by the bursting of high priority workloads. Failure and over-provisioned situations also present challenges for coarse-grain QoS. Rate limiting doesn't provide any guarantees if the system can't even deliver at the configured limit when it is overtaxed or suffering from performance-impacting failures. While prioritization can minimize the impact of failures for some applications, it still can't tell you ahead of time how much impact there will be, and the applications in the lower tiers will likely see horrendous performance.

Requirement #6: Performance virtualization

What it enables - The ability to separate provisioning for capacity and provisioning for performance — on demand. All modern storage systems virtualize the underlying raw capacity of their disks, creating an opaque pool of space from which individual volumes are carved. However, the performance of those individual volumes is a second order effect, determined by a number of variables such as the number of disks the volume is spread across, the speed of those disks, the RAID-level used, how many other applications share the same disks, and the controller resources available to service I/O.

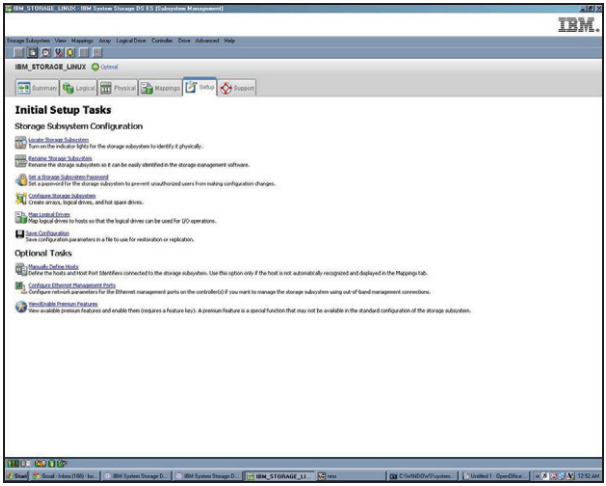
Traditional capacity virtualization does not suffice - Historically this approach has prevented storage systems from delivering any specific level of performance. "More" or "less" performance could be obtained by placing a volume on faster or slower disks or by relocating adjacent applications that may be causing impact. However, this solution is a manual and error-prone process. In a cloud environment, where both the scale and the dynamic nature prevent manual management of individual volumes, this approach just isn't possible. Worst of all, significant raw capacity is often wasted as sets of disks get maxed out from a performance standpoint well before all their capacity is used.

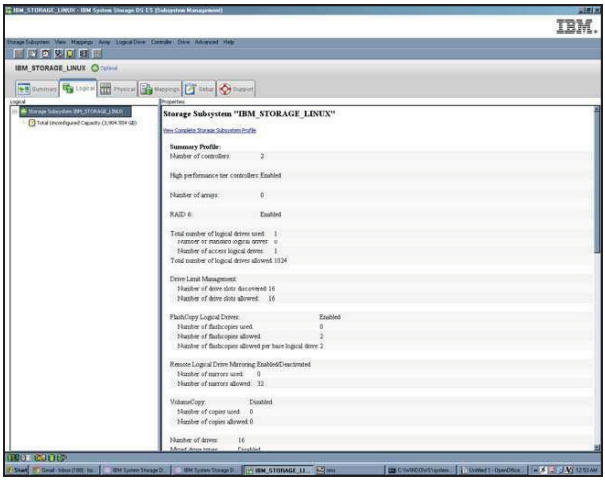
2.12 Storage Configuration



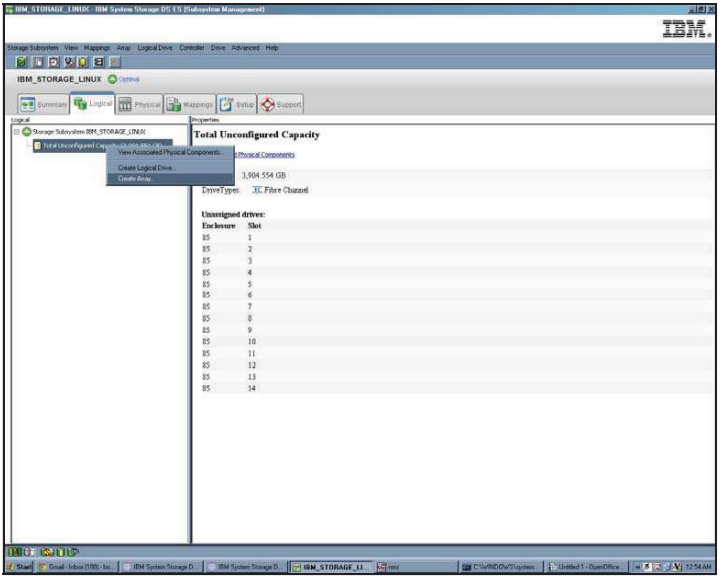
Here, we are using two servers, following IP address and WWN and a SAN two sanswitch and IBM storage.

SERVER: 1.RAC:172.17.1.201-WWN-c6:ec
WWN-c6:e2
RAC2:172.17.1.202-WWN-c6:eb
WWN-c6:c5
SAN S/W:1.172.17.1.183,172.17.1.184
IBM CTRLA: 172.17.1.181, 172.17.1.182

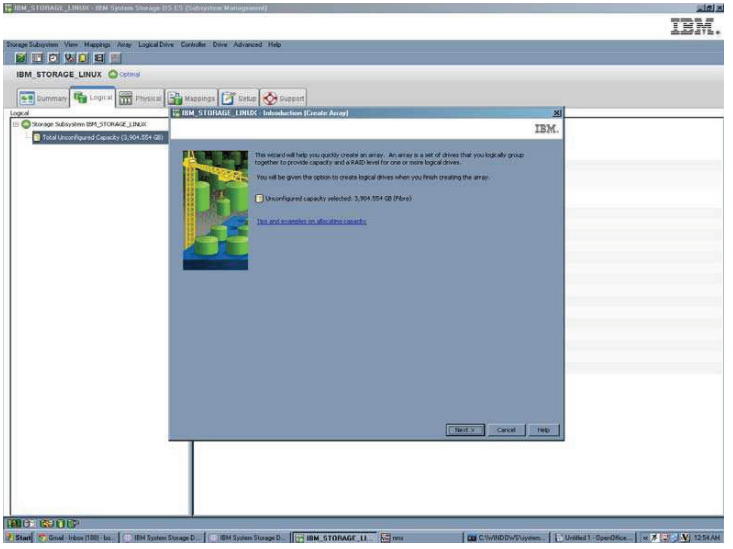


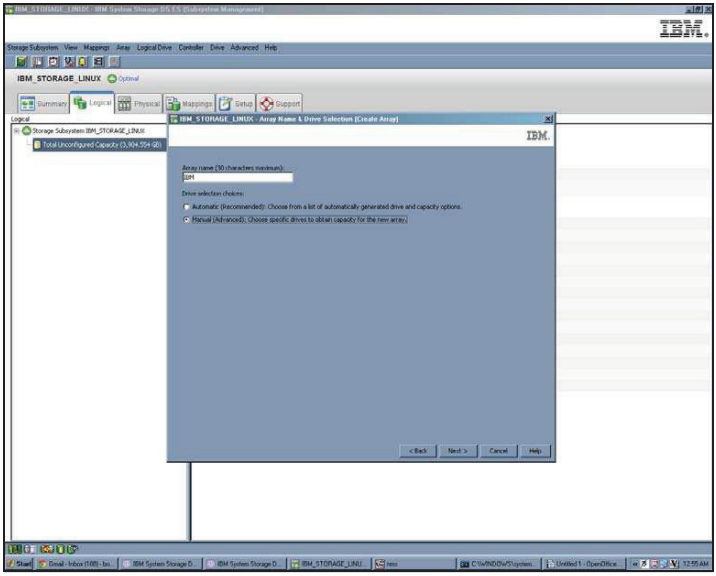


1. First Create Array and give an Array Name

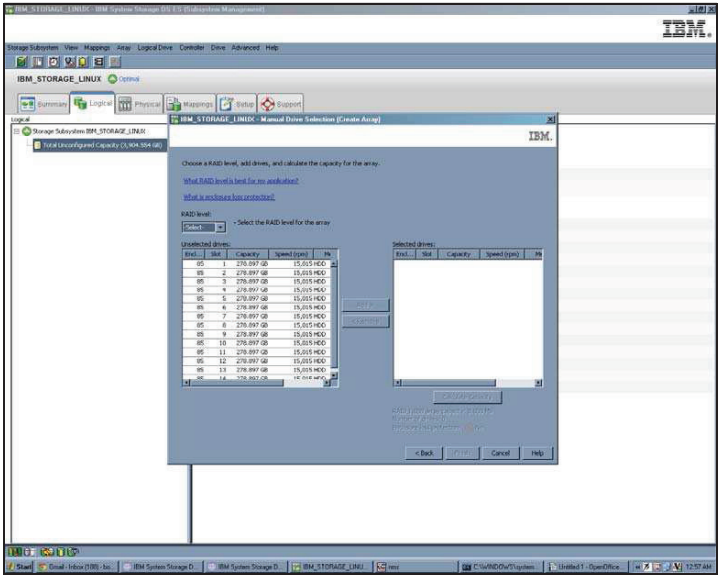


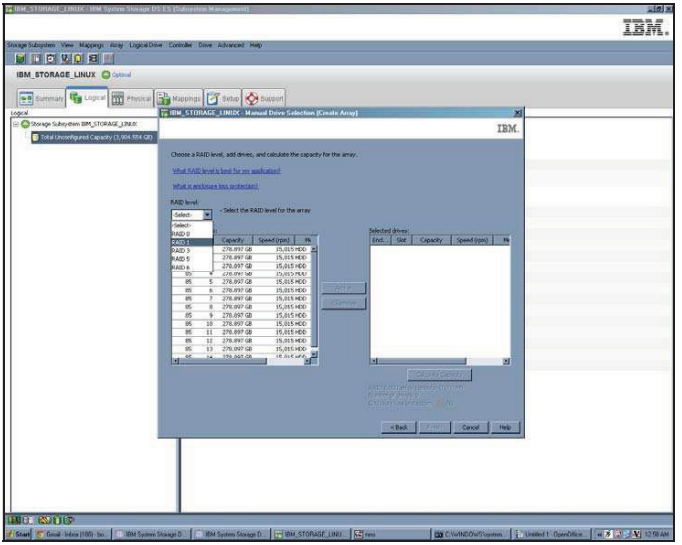
2. Then Create Logical Drive

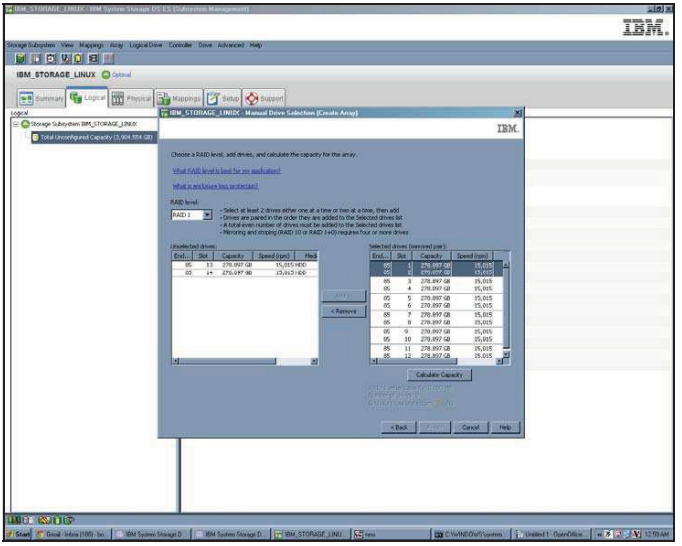




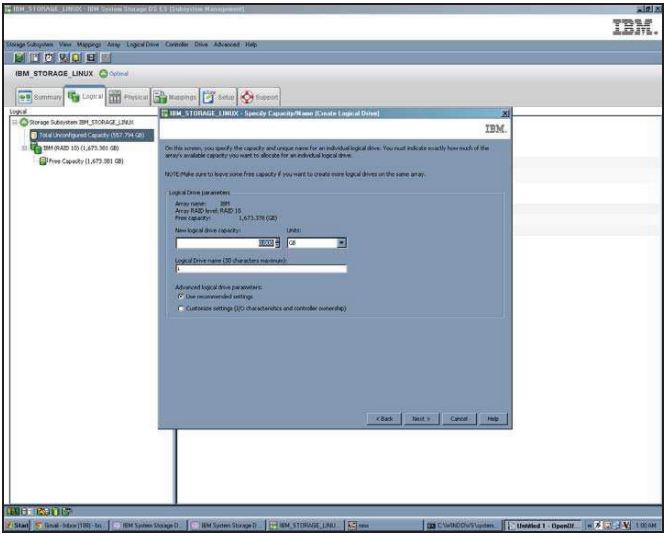
3. Choose RAID1 For our Enviroment:

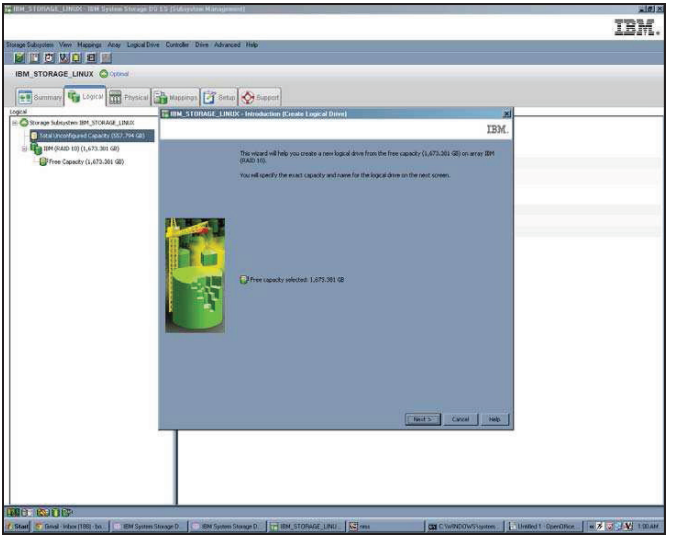




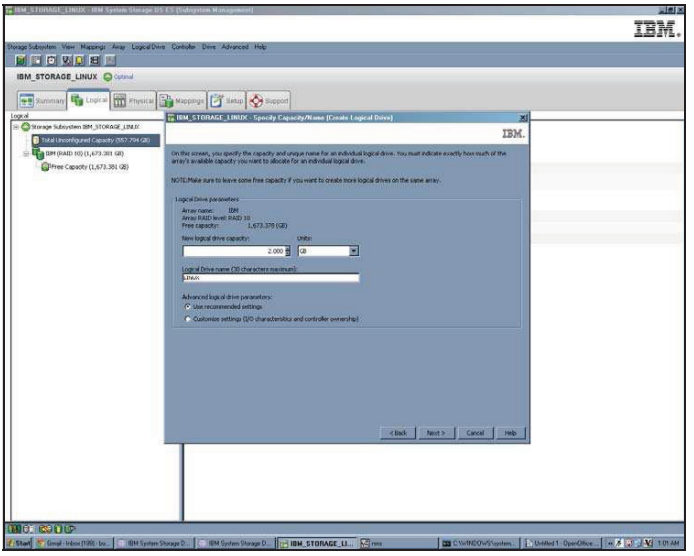


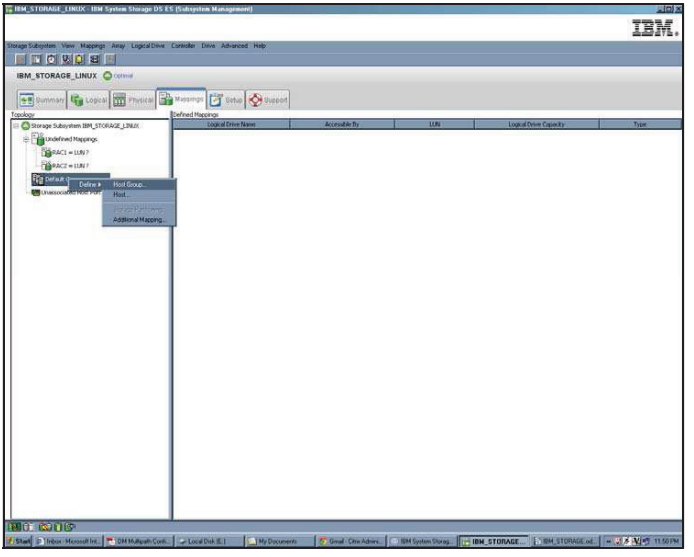
4. Create LUN:

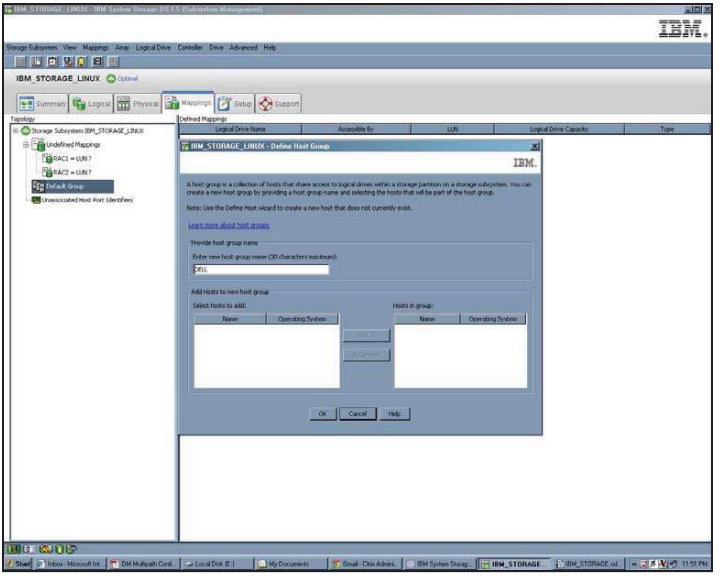




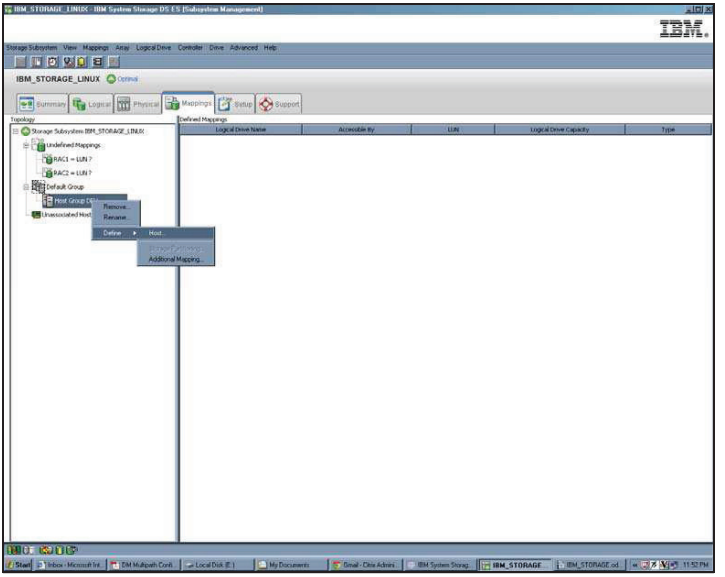
5. Define Host Group

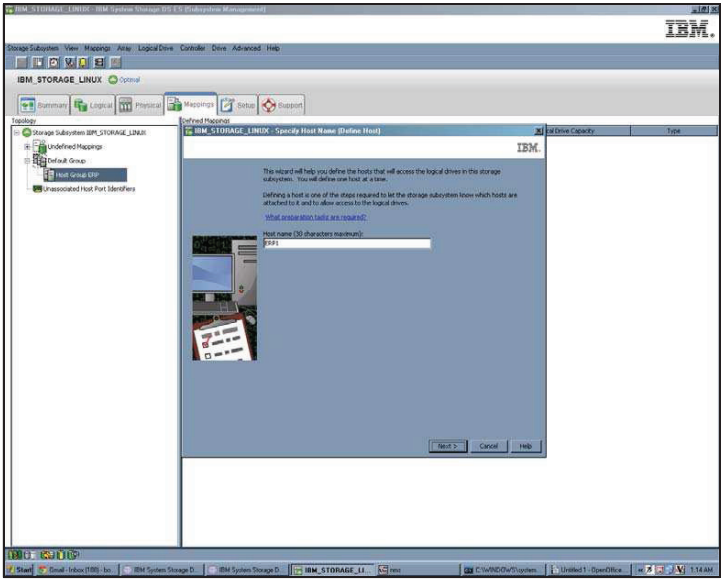


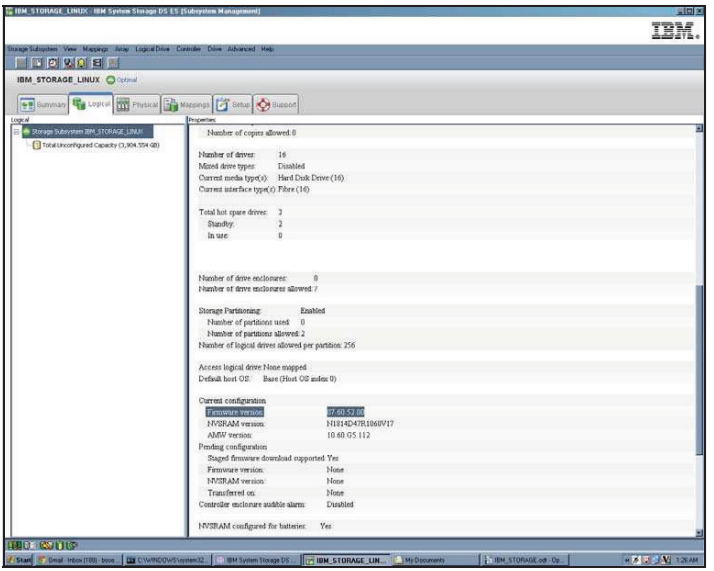


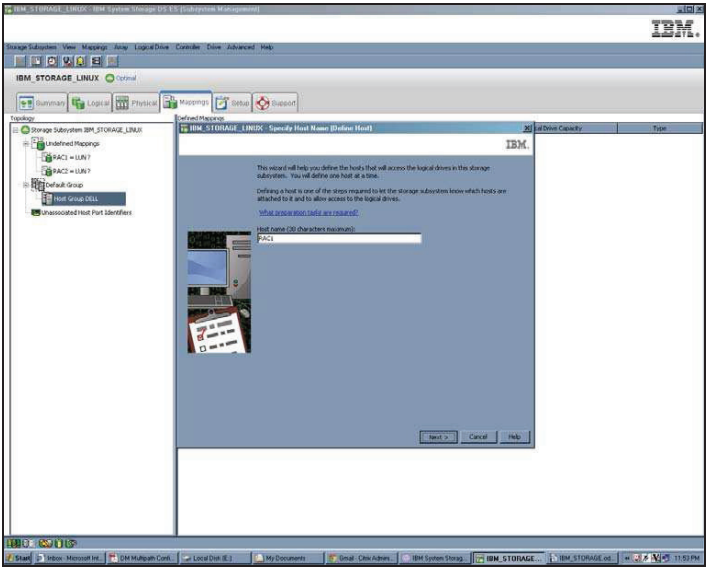


6. Define Host to a Host Group

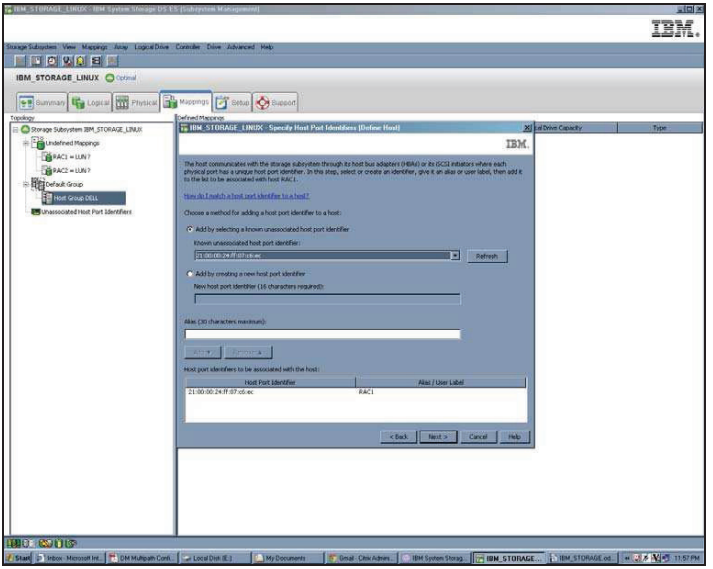


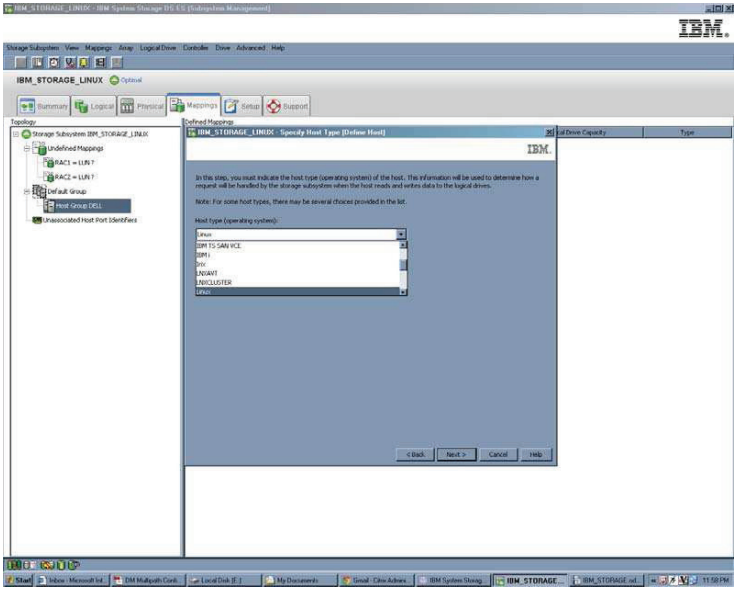

IBM_STORAGE_LINUX
Summary
Logical
Physical
Mappings
Setup
Support
IBM_STORAGE_LINUX - Specify Host Name (Define Host)
This wizard will help you define the hosts that will access the logical drives in this storage subsystem. You will define one host at a time.
Defining a host is one of the steps required to let the storage subsystem know which hosts are attached to it and to allow access to the logical drives.
What preparation tasks are required?
Host name (30 characters maximum):
ERP1
Next >
Cancel
Help

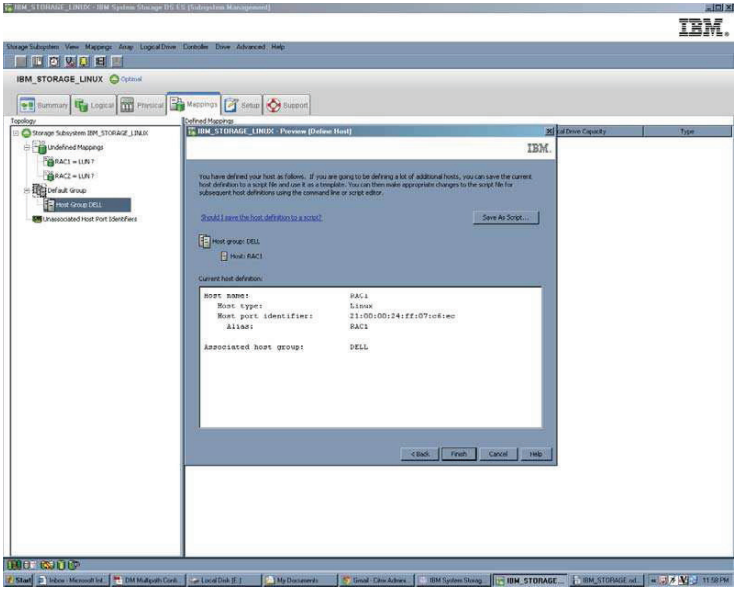

IBM_STORAGE_LINUX
Summary
Logical
Physical
Mappings
Setup
Support
Number of copies allowed: 8
Number of drives: 16
Mixed drive types: Disabled
Current media type(s): Hard Disk Drive (16)
Current interface type(s): Fibre (16)
Total hot spare drives: 2
Standby: 2
In use: 0
Number of drive enclosures: 0
Number of drive enclosures allowed: 7
Storage Partitioning: Enabled
Number of partitions used: 0
Number of partitions allowed: 2
Number of logical drives allowed per partition: 256
Access logical drive: None mapped
Default host OS: Base (Host OS index 0)
Current configuration
Firmware version: 07.60.52.00
NVSRAM version: N1814D47R1060V17
AMW version: 10.60.G5.112
Pending configuration
Staged firmware download supported: Yes
Firmware version: None
NVSRAM version: None
Transferred on: None
Controller enclosure audible alarm: Disabled
NVSRAM configured for batteries: Yes

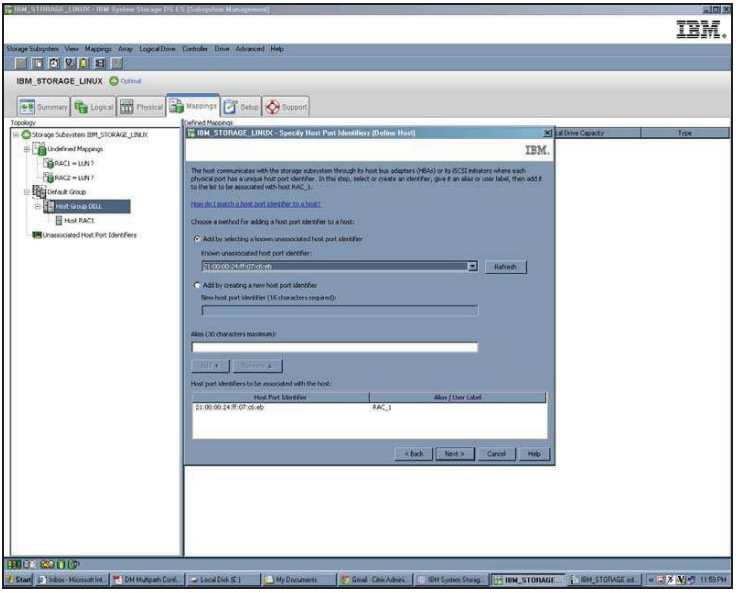


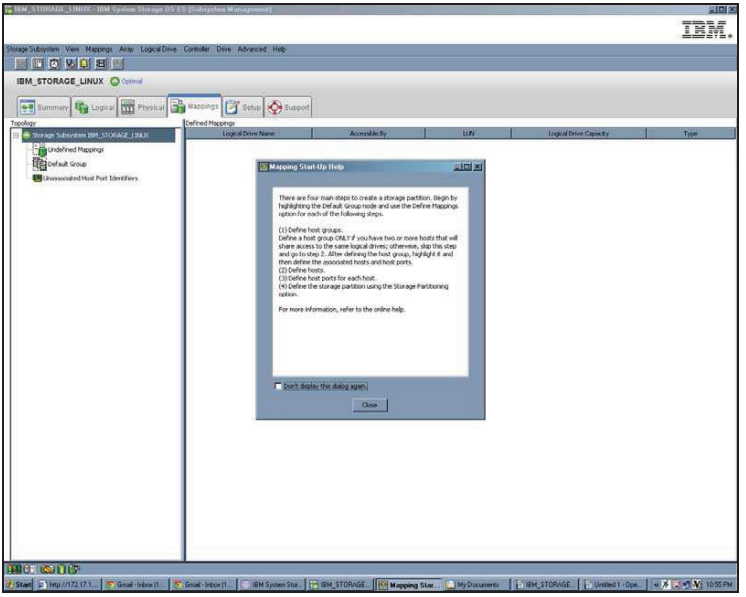
7. Mapping WWN Number:



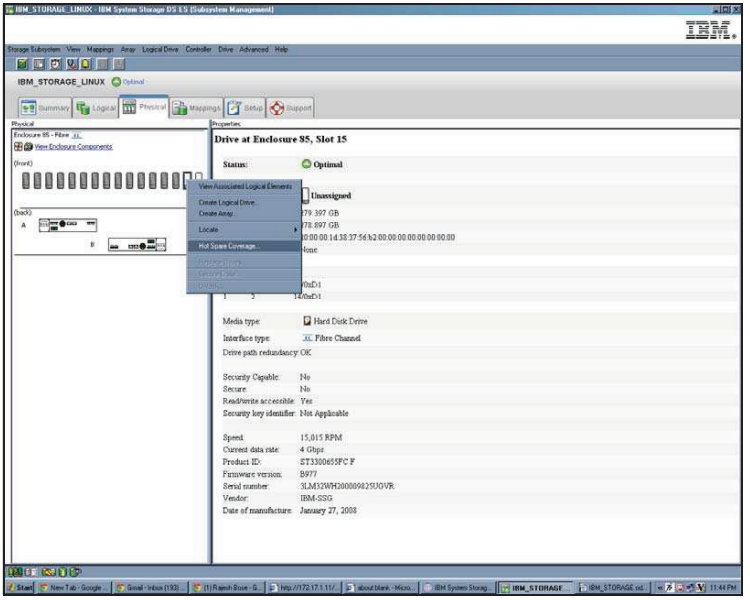

IBM_STORAGE_LINUX
Optimal
Summary
Logical
Physical
Mappings
Setup
Support
Topology
Defined Mappings
Storage Subsystem IBM_STORAGE_LINUX
Undefined Mappings
RAC1 = LUN 7
RAC2 = LUN 7
Default Group
Host Group DELL
Unassociated Host Port Identifiers
IBM_STORAGE_LINUX - Specify Host Type (Define Host)
IBM.
In this step, you must indicate the host type (operating system) of the host. This information will be used to determine how a request will be handled by the storage subsystem when the host reads and writes data to the logical drives.
Note: For some host types, there may be several choices provided in the list.
Host type (operating system):
Linux
IBM TS SAN VCE
IBM-I
HV
LNXAVT
LNXCLUSTER
LNXI
< Back
Next >
Cancel
Help

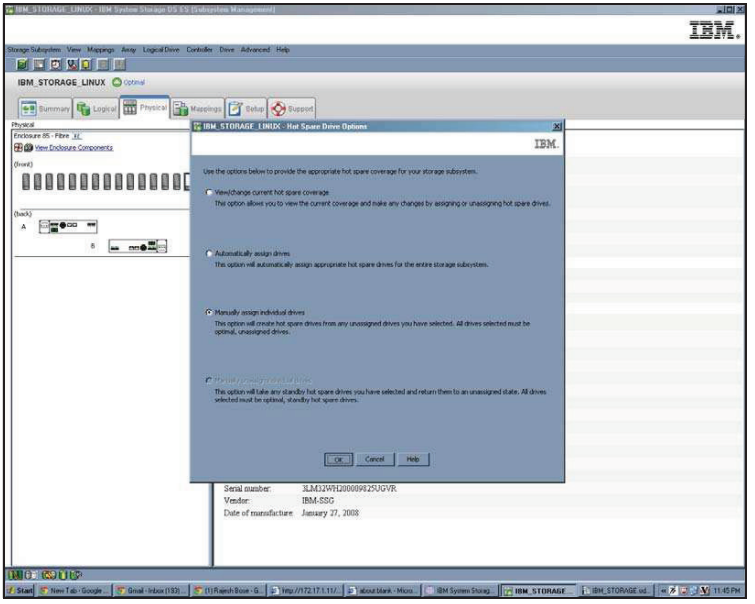

IBM_STORAGE_LINUX
Optimal
Summary
Logical
Physical
Mappings
Setup
Support
Topology
Defined Mappings
Storage Subsystem IBM_STORAGE_LINUX
Undefined Mappings
RAC1 = LUN 7
RAC2 = LUN 7
Default Group
Host Group DELL
Unassociated Host Port Identifiers
IBM_STORAGE_LINUX - Preview (Define Host)
IBM.
You have defined your host as follows. If you are going to be defining a lot of additional hosts, you can save the current host definition to a script file and use it as a template. You can then make appropriate changes to the script file for subsequent host definitions using the command line or script editor.
Should I save the host definition to a script?
Save As Script...
Host group: DELL
Host: RAC1
Current host definition:
Host name: RAC1
Host type: Linux
Host port identifier: 21:00:00:24:ff:07:c6:ec
Alias: RAC1
Associated host group: DELL
< Back
Finish
Cancel
Help


IBM_STORAGE_LINUX
Summary
Logical
Physical
Mappings
Setup
Support
IBM_STORAGE_LINUX - Specify Host Port Identifiers (Define Host)
The host communicates with the storage subsystem through its host bus adapters (HBAs) or its iSCSI initiators where each physical port has a unique host port identifier. In this step, select or create an identifier, give it an alias or user label, then add it to the list to be associated with host RAC_1.
How do I match a host port identifier to a host?
Choose a method for adding a host port identifier to a host:
Add by selecting a known unassociated host port identifier
Known unassociated host port identifier:
Refresh
Add by creating a new host port identifier
New host port identifier (16 characters required):
Alias (30 characters maximum):
Add
Change
Host port identifiers to be associated with the host:
Host Port Identifier
Alias / User Label
21:00:00:24:ff:07:cd:eb
RAC_1
< Back
Next >
Cancel
Help

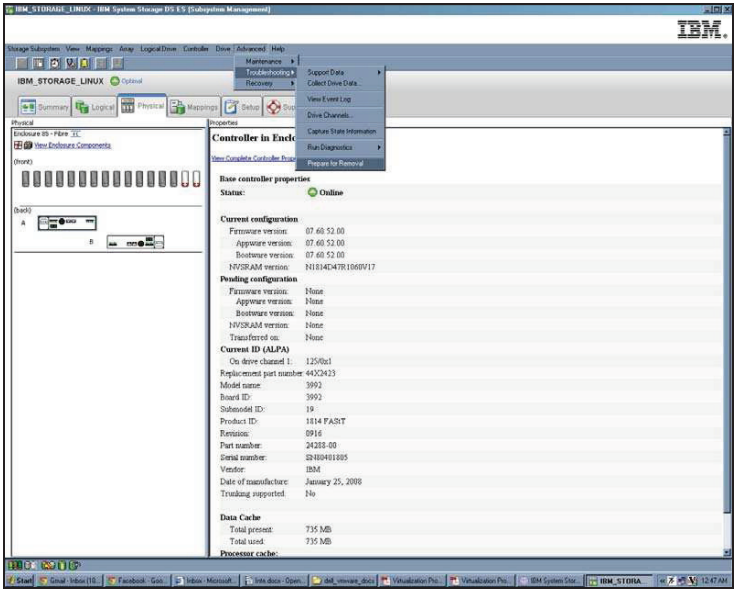

IBM_STORAGE_LINUX
Summary
Logical
Physical
Mappings
Setup
Support
Mapping Start Up Help
There are four main steps to create a storage partition. Begin by highlighting the Default Group node and use the Define Mappings option for each of the following steps.
(1) Define host groups.
Define a host group ONLY if you have two or more hosts that will share access to the same logical drives; otherwise, skip this step and go to step 2. After defining the host group, highlight it and then define the associated hosts and host ports.
(2) Define hosts.
(3) Define host ports for each host.
(4) Define the storage partition using the Storage Partitioning option.
For more information, refer to the online help.
Don't display this dialog again
Close

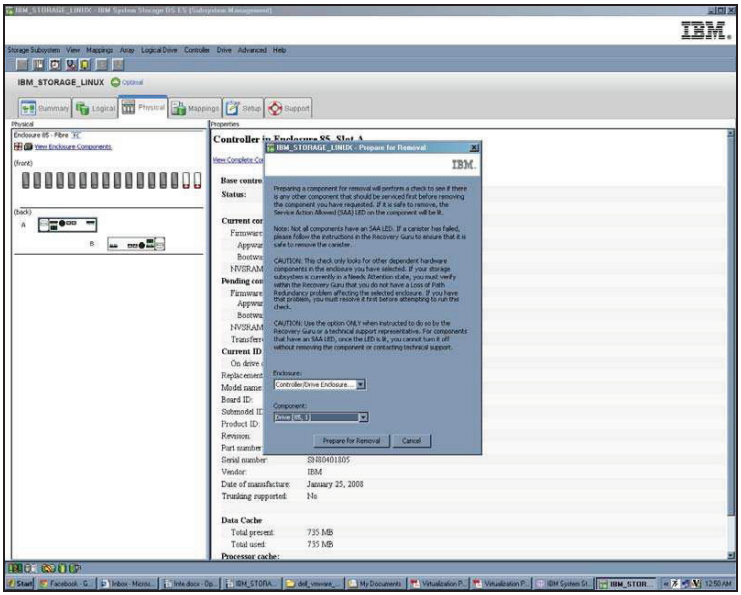
8. Create Hot Spare





9. How to Remove Faulty Disk





CHAPTER 3

FAST AND EFFICIENT BACKUP AND RECOVERY WITH DEDUPLICATION

3.1 Chunking Methods

As storage usage increases exponentially, improving storage efficiency is critical for many data centers today. There are many viable solutions to achieve this objective. These include data migration, thin provisioning, content and quota management, and data deduplication. However, the solution you implement is based on what you believe is a suitable or effective approach to achieve storage efficiency in your production environment. If the intent is to reduce the cost of storing data at the file system level by achieving space savings without affecting end-user experience, and to propagate these space savings within the storage environment at the file system level, an appropriate technology to consider is data deduplication.

For a file server, the main purpose of data deduplication is to increase the file storage efficiency by eliminating redundant data from files stored on a file system hosted by the file server. Though there are many product offerings that feature data deduplication, the objective of any type of deduplication should be to decrease the need for storage space intelligently while being mindful of the user impact.

The purpose of deduplication is to achieve storage efficiency. Storage savings is a means to that purpose. By optimizing the use of the existing storage environment through single instancing and compression, deduplication can save storage costs and lower future storage projections for the data center. Deduplication is enabled at the file system level and is transparent to access protocols.

Before we talk about data deduplication, it may be helpful to remind readers more familiar with storage at an applications level about how files and data sets are represented in conventional disk-based storage systems. The data in a single file or in a single data set is rarely stored in sequential or contiguous blocks even on a single disk system, and in the case of RAID storage, data is almost always written to multiple blocks that are striped across multiple disk systems. In the operating system's file system, the file or the data set is represented by a set of metadata that includes reference pointers to the locations on the disk where the blocks that make up the data set physically reside. In Windows systems the File Allocation Table maps these links; in UNIX/Linux systems the inodes hold the mapping information. Several block-based data storage utilities, including differential snapshots and data deduplication, use a technique in which a single segment or block of data may be referenced simultaneously by multiple pointers in different sets of metadata. The technique for data deduplication also makes use of the idea of using multiple pointers to reference common blocks.

For a great explanation of deduplication we turn to Jeff Bonwick and his experience on the ZFS Filesystem:

Deduplication is the process of eliminating duplicate copies of data. Dedup is generally either file-level, block-level, or byte-level.

Chunks of data -- files, blocks, or byte ranges -- are check summed using some hash function that uniquely identifies data with very high probability. When using a secure hash like SHA256, the probability of a hash collision is about $2^{-256} = 10^{-77}$ or, in more familiar notation, 0.00000000000000000000000000000000000000000000000000000000000000000000000000000001. For reference, this is 50 orders of magnitude less likely than an undetected, uncorrected ECC memory error on the most reliable hardware you can buy.

Chunks of data are remembered in a table of some sort that maps the data's checksum to its storage location and reference count. When you store another copy of existing data, instead of allocating new space on disk, the Dedup code just increments the reference count on the existing data. When data is highly replicated, which is typical of backup servers, virtual machine images, and source code repositories, deduplication can reduce space consumption not just by percentages, but by multiples. Efficient chunking is one of the key elements that decide the overall deduplication performance. There are a number of methodologies to detect duplicate chunk of data using fixed-level chunking and fixed-level chunking using rolling checksums. As described by Won et al., chunking is one of the main challenges in the deduplication system. Policroniades et al. and Kulkarni et al. analyzed the various factors that affect the overall deduplication performance using file-level chunking, fixed-size chunking, and content-aware chunking. Also, various studies including bimodal chunking and Two Threshold Two Denominators (TTTD) emerged to optimize chunking algorithms. Figure 3.1 explains the different chunking module structures in detail. The chunking module contains file-level, fixed-size, variable-size, and content-aware chunking. Input files come to the deduplication system and then the files are moved to the chunking module where the file is divided into small chunks based on the chunking method used. After the chunks division, the unique hash value is assigned to each chunk. In this section, we discuss different chunking modules in detail.

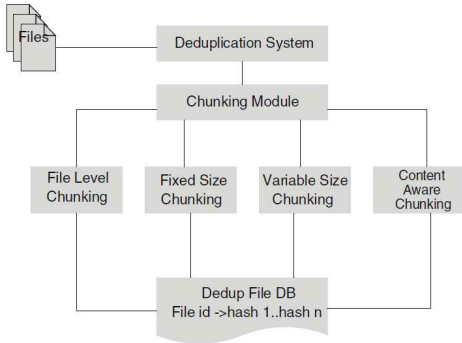


Figure 3.1 Different chunking module structure

3.1.1 File-Level Chunking

File-level chunking or whole file chunking considers an entire file as a chunk, rather than breaking files into multiple chunks. In this method, only one index is created for the complete file and the same is compared with the already stored whole file indexes. As it creates one index for the whole file, this approach stores less number of index values, which in turn saves space and helps store more index values compared to other approaches. It avoids maximum metadata lookup overhead and CPU usage. Also, it reduces the index lookup process as well as the I/O operation for each chunk. However, this method fails when a small portion of the file is changed. Instead of computing the index for the changed parts, it calculates the index for the entire fi le and moves it to the backup location. Hence, it affects the throughput of the deduplication system. Especially for backup systems and large files that change regularly, this approach is not suitable.

3.1.2 Block Level Chunking

Fixed-Size Chunking. Fixed-size chunking method splits files into equally sized chunks. The chunk boundaries are based on offsets like 4, 8, 16 kB, etc. This method effectively solve issues of the file-level chunking method: If a huge file is altered in only a few bytes, only the changed chunks must be reindexed and moved to the backup location. However, this method creates more chunks for larger file which requires extra space to store the metadata and the time for lookup of metadata is more. As it splits the file into fixed size, byte shifting problem occurs for the altered file. If the bytes are inserted or deleted on the file, it changes all subsequent chunk position which results in duplicate index values. Hash collision is likely to happen on chunking method by creating same hash value for different chunks. This can be eliminated by using bit-by-bit comparison which is more accurate, but requires more time to compare the files.

Variable-Size Chunking: The files can be broken into multiple chunks of variable sizes by breaking them up based on the content rather than on the fixed size of the fi les. This method resolves the fixed chunk size issue. When working on a fixed chunking algorithm, fixed boundaries are defined on the data based on chunk size which do not alter even when the data are changed. However, in the case of a variable-size algorithm different boundaries are defined, which are based on multiple parameters that can shift when the content is changed or deleted. Hence, only less-chunk boundaries need to be altered. The parameter having the highest effect on the performance is the fingerprinting algorithm.
One of the widely used algorithms for variable-size chunking is Rabin's algorithm to create the chunk boundaries. Basically, the variable-size chunking algorithm uses more CPU resources. Based on the characteristics of the fi le such as content, size, image, color, etc., we can apply variable-size chunking and fixed-size chunking. An ADMAD scheme method proposed by Liu et al. applies not only the fixed or variable chunking method, but is based on the metadata of individual files on which it applies different fi le chunking methods.

Delta Encoding: Delta encoding is another method to fi nd the boundaries of the file, which prevents repetition of objects relative to each other retaining an existing version of the object with the same name. This method eliminates an object completely in some of the cases, and so the necessary presence of the basic versions, for computing a delta, will be challenging. Compared to variable-size and fixed-size chunking, Delta encoding gives a good deduplication ratio in desktop applications such as Word, Excel, Zip, Photos, though the problem in delta encoding creates more fingerprints than other methods which yield more space utilization. Bolosky et al. found a file granularity method to detect duplicate data instead of using chunk granularity; on the other hand, Deep Store is planned to work with not only one chunking method but can accommodate various chunking approaches.

Basic Sliding Window: The other approach is the basic sliding window (BSW) approach: it applies break condition logic and marks the boundary of file. Chunk boundary is computed based on the fingerprint algorithm. File boundary is marked based on break condition. The problem with this approach is the chunk size. The size of the chunk cannot be predicted with this approach, but it is possible to predict the probability of getting a larger chunk or a smaller one. This probability is defined based on the probability of getting a particular fingerprint. A divisor D and the sliding window size define if the probability is bigger or smaller.

Two Threshold Two Divisor (TTTD): Another most frequently used variable-length chunking algorithm is TTTD. The Two Threshold Two Divisor (TTTD) chunking method ascertains that chunks smaller than particular size is not produced. However, it has a drawback that the chunks produced might escape duplicate detection as larger chunks are more likely to be related than smaller ones. To achieve this, the method ignores the chunk boundaries after a minimum size is reached. TTTD applies two techniques where two divisors (D, the regular divisor and D0, the backup divisor) are used. By applying these divisors, TTTD guarantees a minimum and maximum chunk size. A minimum and a maximum size limit is used for splitting a fi le into chunks for searching for duplicates.

TTTD-S algorithm: This is the method proposed by Moh et al. which overcomes the drawback of the TTTD algorithm by introducing TTTD-S algorithm. It computes the average threshold as a new parameter of the main and the backup divisors and uses it as a benchmark for the chunking algorithm. When the average threshold is hit, both primary and secondary divisors shrink to half. The primary and secondary divisors get restored after chunking is completed. This method helps in avoiding a lot of unnecessary computation. The algorithm by Kruus et al. is based on two fundamental assumptions. The first one is during long runs, where there is a higher probability of getting a large set of unknown data; the second one is during long runs where breaking of large chunks into smaller chunks improves efficiency.

Content or Application Aware-Based Chunking: The chunking methods discussed previously do not consider file format of the file of the underlying data, such as file characteristics. If the chunking method understands the data stream of the file (format of the file), the deduplication method can provide the best redundancy detection ratio compared than the fixed and other variable-size chunking methods. The fixed and the variable-size chunking methods set the chunk boundaries based on the parameter or some predefined condition, whereas the file-type-aware chunking understands the structure of the original file data and defines the boundaries based on the file type such as audio, video, or document. The

advantage of the content-aware chunking method is to facilitate more space savings because this algorithm is aware of the file format and sets the boundaries more natural than other algorithm methods. If the file format is known, it also provides the option to change the chunk size depending on the section of the file. Some formats or file section will not change anytime, by this method it can assume how often the file or section is going to change in the future and the size of the chunk accordingly. For example, potential chunk boundaries include the number of pages in a word document or slides in a PowerPoint presentation. Douglis et al. used different approach to handle the content-aware deduplication. In this method the data is considered as an object. Incoming data is converted into the object and the same has been compared with the already stored objects for finding the duplicate data in effectively. Using of the Byte level comparison and the knowledge of the content of the data, the input file is split into large data segments. The splitted data segments are compared with the already stored segments and similar segments are determined. At end the altered bytes are saved. As corroborated by Meister and Brinkmann, the understanding of the compression formats such as TAR or ZIP can have consequential space savings. They also exploited the characteristic of the file format and analyzed the deduplication ratio and the scaling efficiency under different chunking methods. They found another interesting thing that playing with Meta data and the payload we can yield more savings space. The knowledge of the archive structure and the matching of the chunk boundaries with the payload boundaries can result in the invalidation of a lot of small size chunks and can save a lot of space by avoiding duplicate data. However, the mechanism might not be realistic as to grasp all possible file formats and then using them for breaking the files based on their content is a bit unrealistic for any computational method. Another important issue with fi le-type chunking is that the algorithm might not be able to detect duplicate data if it comes across in a different fi le format. This second problem can rather adversely impact the overall performance of this computational approach than improving it. While this approach holds the most promise as far as the space conservation goes, it can utilize more processing power and time if a detailed analysis of all the individual files is done.

3.2 Deduplication Techniques Classification

Deduplication can be divided based on granularity (the unit of compared data), deduplication place, and deduplication time (Table 3.1). The main components of these three classification criteria are chunking, hashing and indexing. Chunking is a process that generates the unit of compared data, called a chunk.

To compare duplicate chunks, hash keys of chunks are computed and compared, and a hash key is saved as an index for future comparison with other chunks. Deduplication is classified based on granularity. The unit of compared data can be at the file level or subfile level, which are further subdivided into fixed-size blocks, variable-sized chunks, packet payload or byte streams in a packet payload. The smaller the granularity used, the larger number of indexes created, but the more redundant data are detected and removed.

Methods based on granularity	Place	Time
File-level deduplication	Server-based deduplication	Inline deduplication
Fixed-size block deduplication	Client-based deduplication	Offline deduplication
Variable-sized block deduplication	Redundancy elimination (end-to-end RE, network-wide RE)	

Table 3.1 Deduplication classification

For place of deduplication, deduplication is divided into server-based and client-based deduplication for end-to-end systems. Server-based deduplication traditionally runs on high-capacity servers, whereas client-based deduplication runs on clients that normally have limited capacity. Deduplication can occur on the network side; this is known as redundancy elimination (RE). The main goal of RE techniques is to save bandwidth and reduce latency by reducing repeating transfers through the network links. RE is further subdivided into end-to-end RE, where deduplication runs at end points on a network, and network-wide RE (or in-network deduplication), where deduplication runs on network routers.

In terms of deduplication time, deduplication is divided into inline and offline deduplication. With inline deduplication, deduplication is performed before data are stored on disks, whereas offline deduplication involving performing deduplication after data are stored. Thus, inline deduplication does not require extra storage space but incurs latency overhead within a write path. Conversely, offline deduplication does not have latency overhead but requires extra storage space and more disk bandwidth because data saved in temporary storage are loaded for deduplication and deduplicated chunks are saved again to more permanent storage. Inline deduplication mainly focuses on latency-sensitive primary workloads, whereas offline deduplication concentrates on throughput-sensitive secondary workloads. Thus, inline deduplication studies tend to show trade-offs between storage space savings and fast running time.

First explain chunk index caches and bloom filters that are used to identify redundant data based on indexes and small arrays, respectively. We then go into detail about classified deduplication techniques, discussing each one by one, in the order of granularity, place and time. Note that a deduplication technique can belong to multiple categories, such as a combination of variable-sized block deduplication, server-based deduplication and inline deduplication.

3.3 Deduplication Techniques by Time

3.3.1 Inline Deduplication

Inline deduplication is a deduplication technique that removes redundancies before data are stored on disk. Inline deduplication can be applied to primary workloads like email, user directories, databases and secondary workloads like archives and backups. Figure 3.3 elaborates how inline deduplication works for primary workloads (latency sensitive) as well

as secondary workloads (throughput sensitive). For primary workloads, as shown in Fig. 3a, deduplication runs on a direct write and read path. When a user or client writes data, deduplication intercepts the data and checks for redundancies. Only unique data and indexes are saved to storage along with cache. Applications using primary workloads are highly latency sensitive; thus, deduplication typically uses in-memory cache to reduce disk I/O requests. Figure 3b shows how deduplication works for secondary workloads. In these workloads, deduplication runs when data are archived or backed up on a backup server. The backup server does not maintain additional storage. Inline deduplication has been proposed to remove redundancies for the primary workload and secondary workload without incurring extra space overhead or requiring more disk bandwidth. However, this approach requires latency overhead in a write path. iDedup exploits temporal locality and spatial locality to maintain fast processing times in a write path. Content address storage (CAS) systems run inline deduplication because blocks are addressed by their fingerprints. A few file systems use inline deduplication for primary storage. Inline deduplication runs deduplication before data are saved to disk storage. iDedup has been proposed for inline deduplication for a primary workload. iDedup exploits spatial locality and temporal locality to achieve high performance (running time). For spatial locality, iDedup performs selective deduplication and mitigates the extra seek time for sequentially read files. For this purpose, iDedup examines blocks at write time and deduplicates full sequences of file blocks if and only if the sequences of blocks are (1) sequential in the file and (2) have duplicates that are sequential on disk. For temporal locality, iDedup maintains dedup-metadata as a Least Recently Used (LRU) cache by which iDedup avoids dedup-metadata I/O.

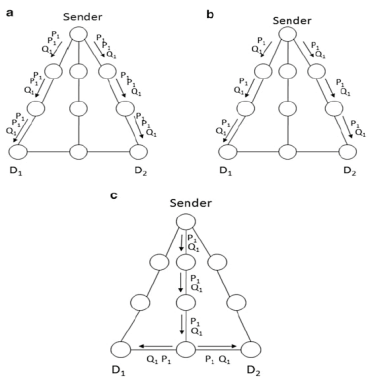


Figure 3.2. Redundant traffic elimination with packet caches on routers. (a) No RE. (b) RE. (c) RE with redundancy-aware routing

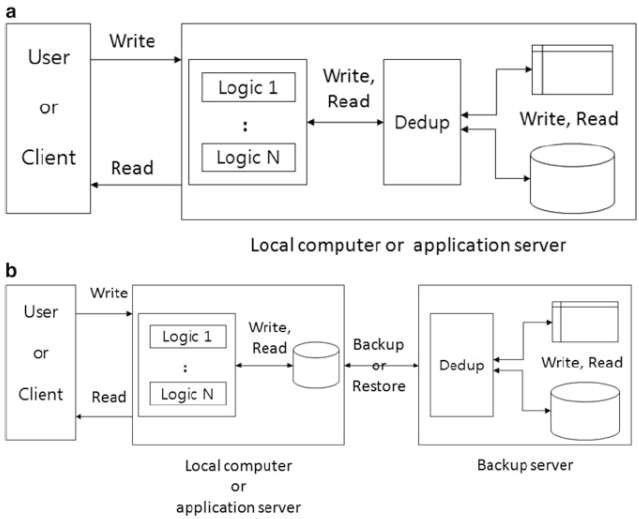


Figure 3.3 Inline deduplication. (a) Inline deduplication for primary workloads. (b) Inline deduplication for secondary workloads

3.3.2 Offline Deduplication

Offline deduplication runs deduplication after data are stored on disk; thus, it does not involve latency overhead in a write path but requires extra storage space. As shown in Figure 3.4, data are saved to storage without deduplication. Offline deduplication runs out of a critical write and read path using already saved data, which does not hurt latency to write and read data. However, offline deduplication has several drawbacks: (1) extra disk space is needed to hold data temporarily before deduplication, (2) deduplication runs on system idle time, so deduplication can be very delayed if the system is running almost all the time, and (3) data on disk are loaded to memory for deduplication, so disk bandwidth is unnecessarily consumed. ChunkStash is a flash-assisted inline deduplication system where chunk metadata (with chunk index as key, and with chunk location and length as value) are saved to flash memory rather than disk. Considering that flash memory is 50 times faster than disk, ChunkStash reduces the penalty of index lookup misses in RAM, which increases inline deduplication throughput. ChunkStash also uses in-memory hash tables using the variant of cuckoo hashing, and compact key signatures rather than full keys are stored in the hash table, which reduces RAM size. HYDRAstor is a grid of storage nodes. It works based on a distributed hash table (DHT) to save blocks to distributed storages. HYDRAstor uses inline deduplication based on immutable and content-addressed and variable-sized blocks, data resilience by erasure coding, load balancing, and preservation of locality of data streams by prefetching. HYDRAstor achieves scalability (by DHT), efficient utilization (by deduplication), fault tolerance (by data resiliency) and system performance (by load balancing, locality and prefetching).

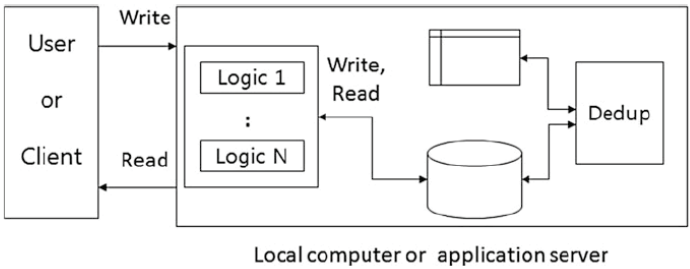


Figure 3.4 Offline deduplication

3.4 Data Deduplication- Multiple Data sets from a Common Storage Pool

Data deduplication operates by segmenting a data set—in a backup environment this is normally a stream of backup data—into blocks and writing those blocks to a disk target. To identify blocks in a transmitted stream, the data deduplication engine creates a digital signature—like a fingerprint—for each data segment and an index of the signatures for a given repository. The index, which can be recreated from the stored data segments, provides the reference list to determine whether blocks already exist in a repository. The index is used to determine which data segments need to be stored and also which need to be copied during a replication operation. When data deduplication software sees a block it has processed before, instead of storing the block again, it inserts a pointer to the original block in the data set's metadata. If the same block shows up multiple times, multiple pointers to it are generated. Variable-length data deduplication technology stores multiple sets of discrete metadata images, each of which represents a different data set but all of which reference blocks contained in a common storage pool.

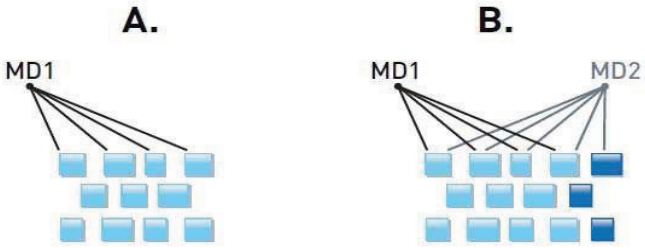


Figure 3.5 Data Deduplication Methodology

When a deduplicated storage pool is first created (A), there is one set of metadata with pointers to the stored blocks. As new data sets are added (B), a separate metadata image

(MD2) is added for each, along with new blocks. Here, MD1 continues to point to the original blocks; MD2 points both to some of the original blocks and to new blocks. For each backup event, the system stores a complete metadata image of the data set, but only new data segments are added to the blockpool.
Since the leverage of the data deduplication technology is highest when there are repeated data segments, the technology is most frequently used today to store backup data. The methodology allows disk to support retention of backup data sets over an extended length of time, and it can be used to recover files or whole data sets from any of multiple backup events. Since it often operates on streams of data created during the backup process, data deduplication was designed to be able to identify recurring data blocks at different locations within a transmitted data set. Because fixed size blocks do not support these requirements well, the Quantum deduplication methodology is built around a system of variable-length data segments.

3.5 Fixed-length Blocks vs variable length data segments

It is possible to look for repeated blocks in transmitted data using fixed-length block divisions, and that approach is currently being used by several backup software suppliers to include deduplication as a feature of the software, and in at least one backup appliance on the market. Fixed block systems are used most often when general purpose hardware is carrying out deduplication because less compute power is required. The trade-off, however, is that the fixed block approach achieves substantially less effective reduction than a variable block approach. The reason is that the primary opportunity for data reduction in a backup environment is in finding duplicate blocks in two transmitted data sets that are made up mostly—but not completely—of the same segments. If we divide a backup data stream into fixed-length blocks, any change to one part of the data set normally creates changes in all the downstream blocks the next time the data set is transmitted. Therefore, two data sets with a small amount of difference are likely to have very few identical blocks (see figure 3.6). Instead of fixed blocks, Quantum's deduplication technology divides the data stream into variable length data segments using a data-dependent methodology that can find the same block boundaries in different locations and contexts. This block-creation process allows the boundaries to "float" within the data stream so that changes in one part of the data set have little or no impact on the boundaries in other locations of the data set. Through this method, duplicate data segments can be found at different locations inside a file, inside different files, inside files created by different applications, and inside files created at different times.

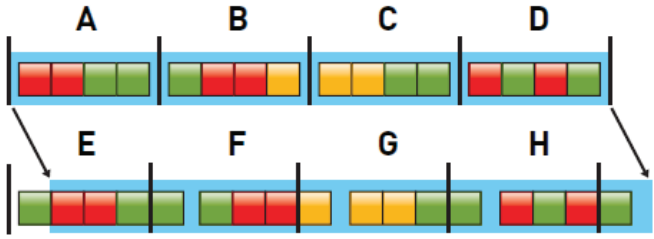
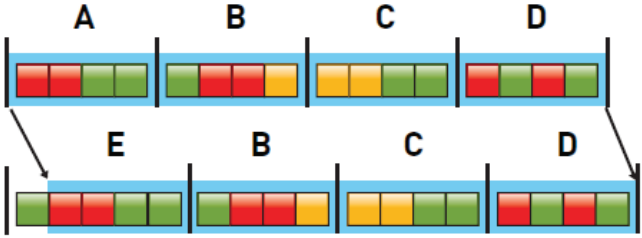


Figure 3.6 Dividing Data Sequences into Fixed or Variable Sized Blocks

Applying fixed block lengths to a data sequence:

The upper line shows the original block division—the lower line shows the blocks after making a single change to Block A (an insertion). In spite of the fact that the shaded sequence of information is identical in the upper and lower lines, all of the blocks have changed content and no duplication is detected. If we stored both sequences, we would have 8 unique blocks.



Applying variable-length segmentation to a data sequence:
Data deduplication utilizes variable-length blocks or data segments when looking at a data sequence. In this case, Block A changes when the new data is added (it is now E), but none of the other blocks are affected. Blocks B, C, and D are all recognized as identical to the same blocks in the first line. If we stored both sequences, we would have only 5 unique blocks.

3.6 Effect of change in Deduplicated Storage Pools

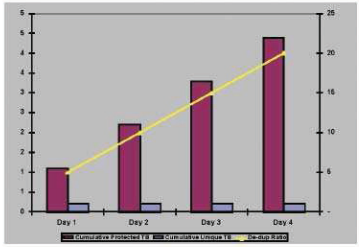
When a data set is processed for the first time by a data deduplication system, the number of repeated data segments within it varies widely depending on the nature of the data (this includes both the types of files and the applications used to create them). The effect can range from negligible benefit to a gain of 50% or more in storage efficiency. However when multiple similar data sets are written to a common deduplication pool—such as a sequence of backup images from a specific disk volume—the benefit is typically very significant because each new write operation only increases the size of the total pool by the number of new data segments that it introduces. In data sets representing conventional business operations, it is common to have a data segment level difference between two backup events in the range of 1 to 5%, although higher change rates are also seen frequently.
The number of new data segments introduced in any given backup event will depend on the data type, the rate of change between backups, whether a fixed-block or a variable-block approach is used, and the amount of data growth from one backup job to the next. The total number of data segments stored over multiple backup events also depends to a very great extent on the retention policies set by the user—the number of backup jobs and length of time they are held on disk. The difference between the amount of space that would be required to store the total number of backup data sets in a conventional disk storage system and the capacity used by the deduplication system is referred to as the deduplication ratio.
Figure 3.7 shows the formula used to derive the data deduplication ratio, and Figure 3.4 shows the ratio for four different backup data sets with different overall compressibility and different change rates. Figure 3.9 also shows the number of backup events required to reach the 20:1 deduplication ratio widely used in the industry as a working average for a variable-length data segment-based data reduction system. In each case, for simplicity we are assuming a full backup of all the primary data for each backup event. With either a daily full model or a weekly full/daily incremental model, the size of the deduplicated storage pool

would be identical since only new data segments are added during each backup event under either model. The deduplication ratio would differ, however, since the space that would have been required for a non deduplicated disk storage system would have been much greater in a daily full model—in other words the storage advantage is greater in a full backup methodology even though the amount of data stored remains essentially the same.

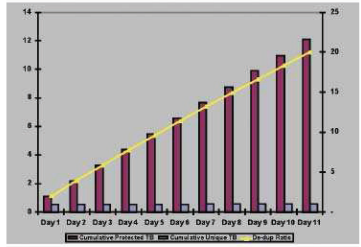
$$\text{Deduplication Ratio} = \frac{\text{Total Data Before Reduction}}{\text{Total Data After Reduction}}$$

Figure 3.7 Deduplication Ratio Formula

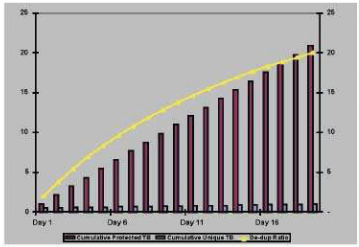
What is clear from the examples is that deduplication has the most powerful effects when it is used for backing up data sets with low or modest change rates between backup events, but even for data sets with high rates of change the advantage can be significant.



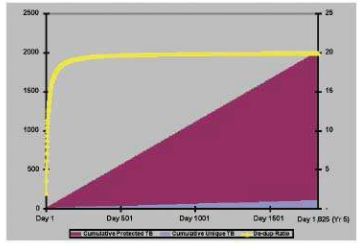
Data Set 1
Compressibility = 5:1
Data Change = 0%
Events to Reach 20:1 Ratio = 4



Data Set 2
Compressibility = 2:1
Data Change = 1%
Events to Reach 20:1 Ratio = 11



Data Set 3
Compressibility = 2:1
Data Change = 5%
Events to Reach 20:1 Ratio = 19



Data Set 4
Compressibility = 2:1
Data Change = 10%
Events to Reach 20:1 Ratio = 1825

Figure 3.8 Effects of Data Change on Deduplication Ratios

3.7 Sharing a common Deduplication Block Pool

Data deduplication systems gain the most leverage when they allow multiple sources and multiple system presentations to write data to a common, deduplicated storage pool.
Quantum's Dxi-Series appliances are an example. Each DXi-Series provides access to a common deduplication storage pool (also known as "blockpool") through multiple presentations that may include a combination of NAS volumes (CIFS or NFS) and virtual tape libraries as well as the Symantec specific OpenStorage (OST) API which writes data to Logical Storage Units (LSUs). Because all the presentations access a common storage pool, redundant data segments are eliminated across all the data sets being written to the appliance. In practical terms, this means that a DXi-Series appliance will recognize and deduplicate blocks that come from different sources and through different interfaces—for example, the same data segments on a print and file server backed up via NAS and on an email server backed up via a VTL.

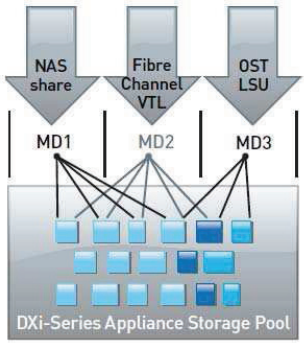


Figure 3.9 Sharing a Deduplication Storage Pool

All the data sets written to the DXi appliance share a common, deduplicated storage pool irrespective of what presentation, interface, or application is used during ingest. One Dxi-Series appliance can support multiple presentations and interfaces simultaneously.

3.8 Data Deduplication Architechture

The data deduplication operation inevitably introduces some amount of overhead and often involves multiple processes at the solution level, including compression (data is normally compressed after it is deduplicated; most "deduplication" processes also include compression when the total data reduction is considered). This means that the choice of where and how deduplication is carried out can affect the speed of a backup process. The deduplication process can be applied to data in a stream (during ingest) or to data-at-rest on disk (post-processing). It can also occur at the destination end of a backup operation or at the source (i.e., at the application server where the backup data is initially processed or on the backup or

media server), or in a hybrid mode where part of the process occurs on the target and part in software on a backup server.
Wherever the data deduplication is carried out, just as in the case of processes like compression or encryption, the fastest performance will in most cases be obtained from purpose-built systems optimized for the specific process. An alternative approach that uses software agents running on general purpose operating platforms to carry out the entire deduplication process can also be effective in some circumstances but it has some disadvantages: since all operations are software based, all protected servers must run the agents, application servers carrying out the process are not designed for the specific data deduplication task, and the server resources will be shared with other operations. For these reasons, the functionality of the software-only approach today generally limits it to very small data sets where system performance is not a priority, and to environments with few servers (since on-going server management overhead is relatively high). Systems that divide deduplication steps between different platforms using a hybrid operating mode are beginning to be offered by vendors. These systems normally carry out segmentation and signature creation on a general purpose system, check the signatures against a central index on an appliance, and then send the unique blocks to the appliance for storage in a central pool of deduplicated data.

3.9 The EMC NetWorker

EMC NetWorker delivers all the qualities that successful organizations, especially small to medium businesses (SMBs) requires, to centralize, automate, and accelerate backup and recovery. With NetWorker, you gain superior manageability, performance, and scalability. It provides fast, reliable, complete data backup and recovery for all systems, databases, and enterprise servers running critical applications. Other key features include:

- ✓ Centralized protection in a heterogeneous environment, with minimal impact to production systems. NetWorker provides comprehensive protection to simplify backup, recovery, and reporting across all operating systems, applications, and topologies
- ✓ Simplifies adoption of new advanced capabilities through integration with next-generation data protection technologies such as data deduplication, backup-to-cloud, CDP, and profile-based server recovery
- ✓ Industry-leading capabilities to protect and recover virtual environments including VMware and Hyper-V to ensure customers the most value from virtualization.
- ✓ Global data deduplication for file systems and applications including Microsoft, SAP, Oracle, Sybase, Lotus, Informix, and DB2.
- ✓ Broad backup-to-disk capabilities, including SAN and NAS backup and with EMC Disk Library and EMC Data Domain
- ✓ Advanced indexing architecture that dramatically increases performance and scalability in backup and recovery operations

- ✓ Enables server-less backup, library sharing, and dynamic drive sharing
- ✓ Online, impact-free backup and granular recovery for leading database, messaging, and ERP applications
- ✓ Fast, reliable disaster recovery management, plus future-proofed Open Tape Format with better recoverability from damaged tape media
- ✓ End-to-end tape media management
- ✓ Policy-based snapshot management

NetWorker enables you to tie protection levels together in a single solution, This "tiered protection" means you can match your investment and protection level to the value of data, all within a single backup application for centralized control and command. While applications and organizational divisions may require different recovery-time and recovery-point objectives (RTO and RPO), successful organizations, especially small to medium businesses (SMBs), can support them all using EMC NetWorker. This results in simpler management and easier introduction of next generation technologies, and enables administrators to more quickly identify and analyze problems for system optimization.

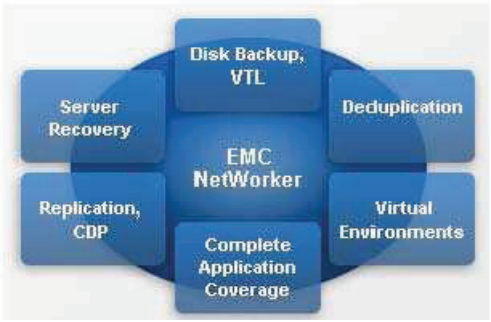


Figure 3.10 EMC NetWorker

Supported Environments:

- ✓ Supports most varieties of UNIX, Windows, Linux, OpenVMS, Mac, and Netware
- ✓ Unparalleled protection for NDMP-compliant- NAS file servers
- ✓ Oracle, DB2, Informix, SQL Server, Sybase, SharePoint, Exchange, Lotus, SAP, MEDITECH
- ✓ Cluster support using EMC AutoStart, Sun Cluster, HP MCSG, OpenVMS Cluster, TruCluster, IBM HACMP, Veritas Cluster, MS Cluster Services
- ✓ Support for extensive list of tape drives and libraries

By standardizing on EMC NetWorker, most organizations, especially small to medium businesses (SMBs).can gain fast, reliable data protection while reducing management overhead in DAS, NAS, and SAN topologies.

3.10 The EMC Data Domain

The EMC Data Domain deduplication storage system can solve many of the challenges that most organizations, especially small to medium businesses (SMBs). experiences with traditional backup, recovery and replication processes. Combining high speed, inline deduplication with local compression, the Data Domain system writes only unique data to disk. Deduplication technology reduces disk capacity requirements and overhead, while increasing accessibility and reliability, and makes a Data Domain system a cost-effective alternative to tape.

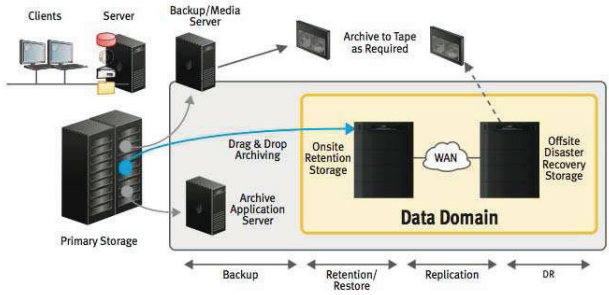


Figure 3.11 EMC Data Domain

Data Domain systems transfer only deduplicated and compressed changes across your IP network, requiring a fraction of the bandwidth, time and cost associated with traditional replication methods. In addition, Data Domain systems make use of technologies that offer advanced data verification and integrity, and leverage CPU advancements to add direct benefit to system throughput and scalability.

3.11 How Data De duplication on Data Domain works

Data Domain deduplication segments the incoming data stream, uniquely identifies the data segments, and then compares the segments to previously stored data. If an incoming data segment is a duplicate of what has already been stored, the segment is not stored again, but a reference is created to it. If the segment is unique, it is stored on disk.

For example, a file or volume that is backed up every week creates a significant amount of duplicate data. Deduplication algorithms analyze the data and can store only the compressed, unique change elements of that file. This process can provide an average of 10-30 times or greater reduction in storage capacity requirements, with average backup retention policies on

normal enterprise data. This means that companies can store 10TB to 30TB of backup data on 1 TB of physical disk capacity, which has huge economic benefits.

3.12 Data Invulnerability Architecture

Data Domain deduplication storage systems focus on data integrity and recoverability as the most important goal. The Data Domain Data Invulnerability Architecture provides continuous fault detection, healing, and write-verification, which ensures backup and archive data are accurately stored, available and recoverable. There are four critical areas of focus:

3.12.1 End-to-End Verification at Backup Time: Data is read after it is written to verify that it is the correct data and that it is reachable through the file system to disk. Most restores happen within a day or two of backups. Systems that verify/correct data integrity slowly over time will be too late for most recoveries.

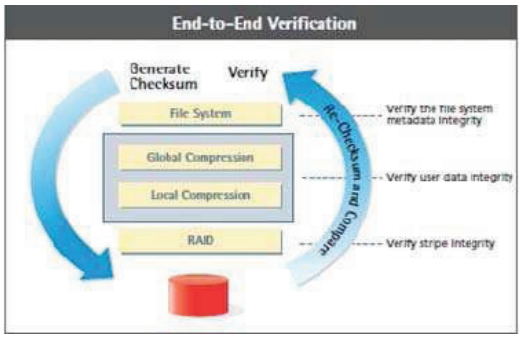


Figure 3.12 End-to-End Verification at Backup Time

3.12.2 Fault Avoidance and Containment: New data never overwrites good data. Data Domain systems use fewer complex data structures and Non-volatile RAM (NVRAM) for fast, safe restart. No partial stripe writes are allowed.

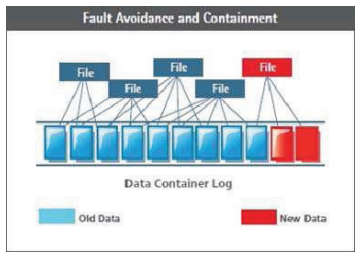


Figure 3.13 Fault Avoidance and Containment

3.12.3 Continuous Fault Detection and Healing: Data Domain RAID-6 provides double disk failure protection and read error correction, on-the-fly error detection and correction, and scrubbing to find/repair grown defects on the disk before they can become a problem.

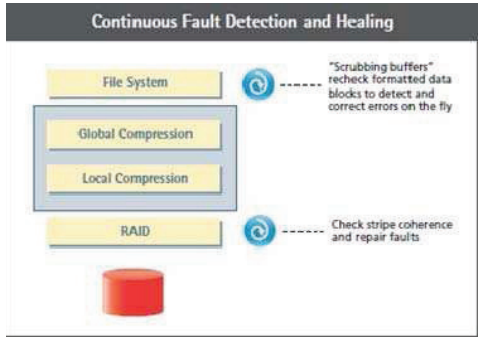


Figure 3.14 Continuous Fault Detection and Healing

3.12.4 Filesystem Recoverability: Data is written in a self-describing format. If necessary, the filesystem can be recreated by scanning the log and rebuilding it from the metadata stored with the data.

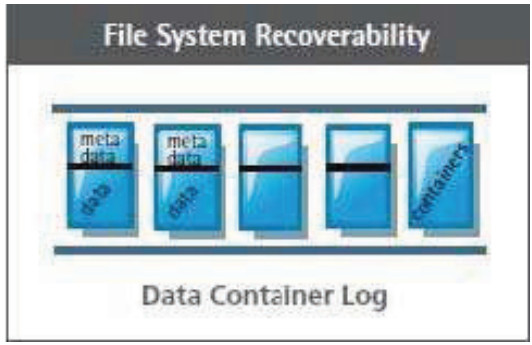


Figure 3.15 Filesystem Recoverability

3.13 EMC Data Domain Stream-Informed Segment Layout Scaling Architecture

Stream Informed Segment Layout (SISL) scaling architecture leverages the continued advancement of CPU performance to add direct benefit to system throughput scalability for inline deduplication. Without careful thought about how to implement it, deduplication can become a disk-intensive activity. The conventional way to increase disk system performance is to increase disk count or use faster, more expensive disks. By adding disks only to increase performance, most organizations, especially small to medium businesses (SMBs) could pay for a lot of unnecessary capacity.

Data Domain systems solve this problem with SISL. It optimizes deduplication throughput and minimizes disk accesses, which allows system throughput to be CPU-centric. Speed will increase directly as new CPUs improve their performance over time. Deduplication provides data reduction that greatly exceeds traditional local compression algorithms. However, to be cost-neutral when compared to tape automation, the deduplication system needs to be CPU-centric and minimize disk accesses, so that it can be built with the minimum number of low-cost, high-capacity disks.

3.14 Solution Benifits

The EMC solution that follows consists of products and services that will enable most organizations, especially small to medium businesses (SMBs) to put information at the center of its business, keep it available and secure around the clock, and streamline processes for greater efficiency. Supporting virtualization across your enterprise, EMC solutions are designed to help you maximize productivity and minimize your total cost of ownership. All EMC products are tested thoroughly in EMC's E-Lab to assure our storage network customers that their multi-vendor environment has been tested and qualified for interoperability, and that it is supported by EMC's industry-leading service organization.

Software

EMC software adds functionality to your solution. It enables your company to meet its operational and business requirements.

EMC NetWorker

EMC NetWorker backup and recovery software centralizes, automates, and accelerates data protection across your IT environment. IT departments can leverage NetWorker to unify all application-, tape-, and disk-based backup and recovery solutions for physical and virtual environments.

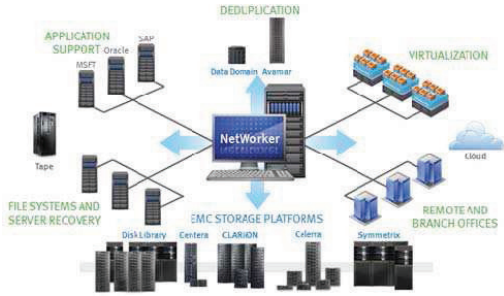


Figure 3.16 EMC NetWorker

Combining traditional backup with next-generation capabilities such as deduplication from industry leading solutions EMC Avamar and EMC Data Domain, NetWorker helps companies safely transition to new technologies that greatly improve data protection. By

implementing NetWorker for backup and recovery, companies can improve data protection, simplify operations, control costs, and deliver higher levels of recovery services than ever before.

EMC NetWorker Module for Databases and Applications

EMC NetWorker Module for Databases and Applications delivers unified protection for DB2, Informix, Lotus, Oracle, and Sybase. This single module provides high performance, online, database-aware protection for mission-critical databases and applications to increase productivity while simplifying licensing and maintenance.

Figure 3.17 EMC NetWorker Module for Databases and Applications

In addition to all standard NetWorker server functionality, the NetWorker Module for Databases and Applications supports deduplication (including Data Domain Boost Software) and NetWorker's advanced VMware capabilities. The NetWorker Module for Databases and Applications ensures maximum recoverability with as little downtime as possible and helps you meet your most demanding SLAs.

EMC NetWorker Module for SAP with Oracle

EMC NetWorker Module for SAP with Oracle ensures availability of mission-critical enterprise resource planning (ERP) data by delivering fast, online backup and recovery for SAP. With NetWorker Module for SAP with Oracle, SAP administrators can centrally manage backup operations and schedules.

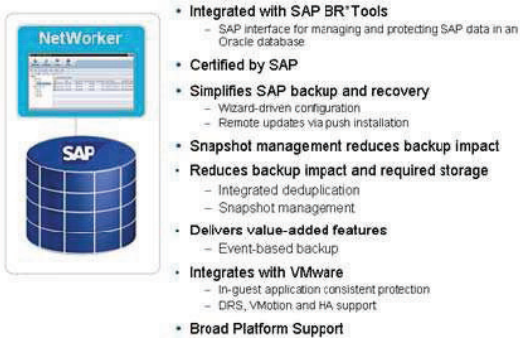


Figure 3.18 EMC NetWorker Module for SAP with Oracle

Featuring integrated deduplication support for Avamar and Data Domain, the NetWorker Module for SAP with Oracle lessens the impact of data protection operations on production servers and networks, and significantly reduces the amount of backup storage required.

Hardware Requirement

With EMC Data Domain systems, most organizations, especially small to medium businesses (SMBs). can realize dramatic cost and bandwidth savings by reducing the amount of disk storage needed to retain and protect data, ensuring data integrity and availability, and maximizing throughput by leveraging CPU technology improvements with the latest Data Domain solutions.

Benefits include:

- ✓ **Fast Inline Deduplication** — the most efficient and economic method of deduplication, it significantly reduces the raw disk capacity needed in the system since only deduplicated data is written to disk.
    - o Up to 3.4 TB/hour of aggregate throughput
    - o Extended retention providing up to .30-1.6 PB of logical storage
    - o 10-30x average data reduction
- ✓ **Fast Time to Disaster Recovery Readiness** — as part of the inline deduplication process, the system does not need to wait to ingest the entire data set before it can begin replicating to the remote site. Data is available for recovery at a DR site faster than with other deduplication products or by physically transporting tapes on trucks.
- ✓ **Easy Integration**:
    - o Supports leading enterprise applications including Oracle, SAP, DB2, SQL, Exchange, VMware, and SharePoint
    - o Simultaneous use of VTL, NDMP, NAS and EMC Data Domain Boost (for use with EMC Avamar, EMC NetWorker, and Symantec OpenStorage)

- ✓ **Multi-site, Cost-Efficient Disaster Recovery** — delivers up to 99% bandwidth reduction, provides flexible replication topologies, multi-site tape consolidation and replication from up to 270 remote sites.
- ✓ **Ultra-safe Storage for Reliable Recovery** — Data Invulnerability Architecture provides continuous recovery verification, fault detection and healing plus dual-disk parity RAID-6.
- ✓ **Operational Simplicity** — Lowers administrative costs, provides power, cooling and space efficiencies for environmentally friendly operations, reduces hardware footprints and supports any combination of backup and archive applications in a single system.

In addition, most organizations, especially small to medium businesses (SMBs) gains the benefit of EMC's long history information infrastructure experience, proven in thousands of customer deployments. EMC's leading-edge technology and full suite of solutions, services and training, offer most organizations, especially small to medium businesses (SMBs) unmatched expertise in this critical area.

CHAPTER 4

EVOLUTION OF CLOUD STORAGE

4.1 Cloud Storage Overview

Cloud computing is a computing paradigm in which hosted services are delivered to the user over a wide area network (WAN) using standard Internet protocols. The term “cloud” was coined from the use of the cloud symbol in network diagrams to represent a section in the Internet. In such diagrams the cloud abstracts the details of that part of the network in order to, in most cases, present a view that is focused on the services provided by the cloud part of the network. Several distinctive features differentiate a cloud service from a traditional hosted service. Firstly, a cloud service is sold on demand, usually with pay-per-use or subscription model. In pay-per-use, the user only pays for how much of the services they use, for example, only the number of bytes transferred. Secondly, a cloud service is elastic – at any given time a user can get as little or as much as they need. The third differentiating feature is that a cloud service is fully managed by the service provider - the user only needs network connectivity to the service and enough local resources to use the service. The local resources vary depending on the type of service but can be as simple as a mobile smart phone. In the cloud business model service level agreements (SLA) make the provider accountable for the quality and reliability of the service. This does not only protect the interests of the client but also clearly spells out exceptions between the parties to the agreement. Examples of services hosted in the cloud include infrastructure services like servers and data storage. The others are platform and software services. Platform services include software and product development tools, whereas software services (software applications) encompass such services as web-based email and database processing.

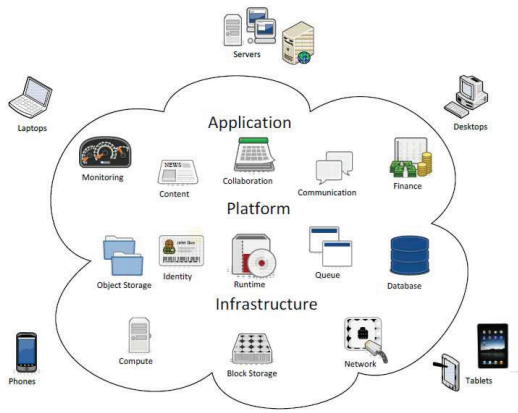


Figure 4.1 An overview of cloud computing

Given the amount of data being generated annually and the need to leverage good storage infrastructure and technologies, data storage emerges as one of the top cloud computing services. In their 2010 digital universe study, IDC indicated that by 2020, the amount of digital information created and replicated will grow to over 35 trillion gigabytes. Further, the report indicates that a significant portion of digital information will be centrally hosted, managed, or stored in public or private repositories that today we call “cloud services”. The 2011 digital universe study predicted that that in 2011 alone, the amount of digital information created and replicated would be more that 1.8 zettabyte (1.8 trillion gigabytes).

Just like Cloud Computing, Cloud Storage has also been increasing in popularity recently due to many of the same reasons as Cloud Computing. Cloud Storage delivers virtualized storage on demand, over a network based on a request for a given quality of service (QoS). There is no need to purchase storage or in some cases even provision it before storing data. You only pay for the amount of storage your data is actually consuming. FilesAnywhere.com was one of the first companies to offer the cloud storage service. Their cloud storage service enabled users to store data on their servers from anywhere at any time, while also being able to retrieve the data from anywhere at any time. FilesAnywhere.com would be a pioneer in the cloud storage business and many companies would follow suit. When virtualized storage is available on demand over a network, organizations are not required to buy or provision storage capacity before storing data. As a result, organizations can save a significant amount of money on storage costs because they typically only pay for the storage that they actually use. When SNIA (STORAGE NETWORKING INDUSTRY ASSOCIATION) recognized the significant changes in the way that organizations use storage, it developed the Cloud Data Management Interface (CDMI) standard for cloud storage vendors and others to use when implementing their own public and private clouds. As the interest in cloud storage grows, SNIA (STORAGE NETWORKING INDUSTRY ASSOCIATION) is also developing a series of education programs. These programs provide advice and recommendations for service providers when deploying clouds, and include use cases for organizations when adopting the technology.

4.2 Evolution of Cloud Storage

Cloud storage is an offering of cloud computing. Figure 4.2 shows the evolution of Cloud Storage based on traditional network storage and hosted storage. Benefit of cloud storage is the access of your data from anywhere. Cloud storage providers provide storage varying from small amount of data to even the entire warehouse of an organization. Subscriber can pay to the cloud storage provider for what they are using and how much they are transferring to the cloud storage. Basically the cloud storage subscriber copies the data into any one of the data server of the cloud storage provider. That copy of data will be made available to all the other data servers of the cloud storage provider featuring redundancy in the availability which ensures that the data of the subscriber is safe even anything goes wrong. Most systems store the same data on servers that use different power supplies.

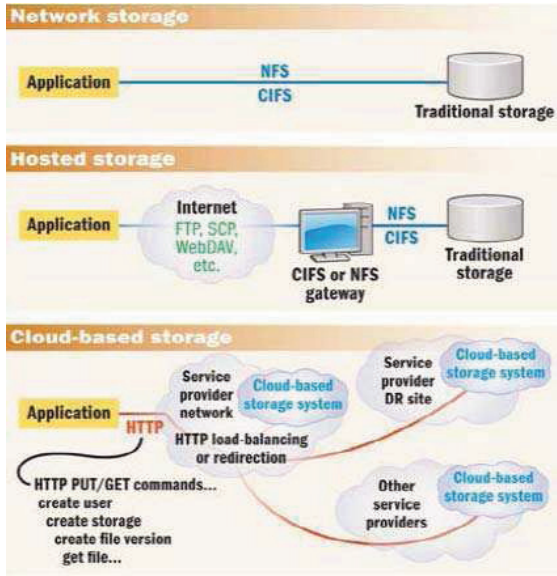


Figure 4.2 Evolution of Cloud Storage

4.3 Benefits of Cloud storage:

- ✓ No need to invest any capital on storage devices.
- ✓ No need for technical expert to maintain the storage, backup, replication and importantly disaster management.
- ✓ Allowing others to access your data will result with collaborative working style instead of individual work.

4.4 What makes Cloud Storage different?

The difference between the purchase of a dedicated appliance and that of cloud storage is not the functional interface, but merely the fact that the storage is delivered on demand. The customer pays for either what they actually use or in other cases, what they have allocated for use. In the case of block storage, a LUN or virtual volume is the granularity of allocation. For file protocols, a file system is the unit of granularity. In either case, the actual storage space can be thin provisioned and billed for based on actual usage. Data services such as compression and deduplication can be used to further reduce the actual space consumed. The management of this storage is typically done out of band of these standard Data Storage interfaces, either through an API, or more commonly, though an administrative browser based user interface. This interface may be used to invoke other data services as well, such as snapshot and cloning.

4.5 The Requirement for a Cloud Storage Standard

Because it is relatively easy for cloud storage solutions to improve business processes, interest in and adoption of cloud storage solutions are growing. Until now, there have been few standards that simplify and allow interoperability across disparate cloud solutions—standards that organizations need if they want to enjoy the benefits of an open, competitive marketplace in cloud storage. Because the variety of use cases rarely share a common interface to cloud storage, SNIA formed a technical working group (TWG) with over 75 members to develop a standard for cloud storage. In June 2009, it published a use cases and reference model, and in July 2010, the TWG published the first draft standard, the Cloud Data Management Interface (CDMI), with the intention of gaining future ISO and ANSI certification. Currently, the TWG is working on cross-platform reference architecture to encourage adoption of this new standard.

4.6 Cloud Storage—an Abstract Model

The simplest way to describe cloud storage is this: A cloud represents a “fuzzy” container for data, and the user doesn’t really care how the cloud provider implements, operates, or manages the cloud. A client, through the medium of a network, makes requests to the cloud storage to securely store and subsequently retrieve data at an agreed level of service. Although seemingly abstract and complex, cloud storage is actually rather simple. Regardless of data type, cloud storage is a pool of resources that are provided in small increments with the appearance of infinite capacity. In other words, cloud storage is virtualized storage on demand and is more formally called 'Data Storage as a Service' (DaaS). DaaS is defined as “Delivery over a network of appropriately configured virtual storage and related data services, based on a request for a given service level.” (See Figure 4.3)

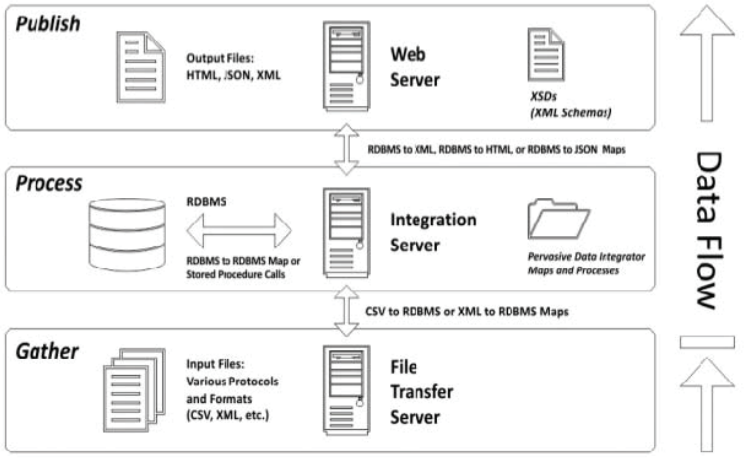


Figure 4.3 Data Storage as a Service

4.7 Cloud Storage—the Reality

Today's IT environment is composed of various products that are intended to store, protect, secure, and make available the information used by businesses and business processes. These products encompass elements used in both the data path and control path between the user and the eventual location of that information. As vendors and suppliers of cloud services have delivered early implementations to users, they have tended to supply a multitude of interfaces that have been re-purposed for DaaS, such as block-based access via iSCSI; POSIX interfaces (NFS, CIFS, and WebDAV); object-based CRUD (Create, Read, Update, Delete) interfaces over HTTP; and a plethora of other proprietary interfaces for database or table access (see Figure 3 – Existing Interface Standards for Data Storage). Compared to the simplicity of the abstract cloud model, the existing cloud storage model is rather complex because there are so many interfaces that are required to meet the different demands of end users for accessing storage.

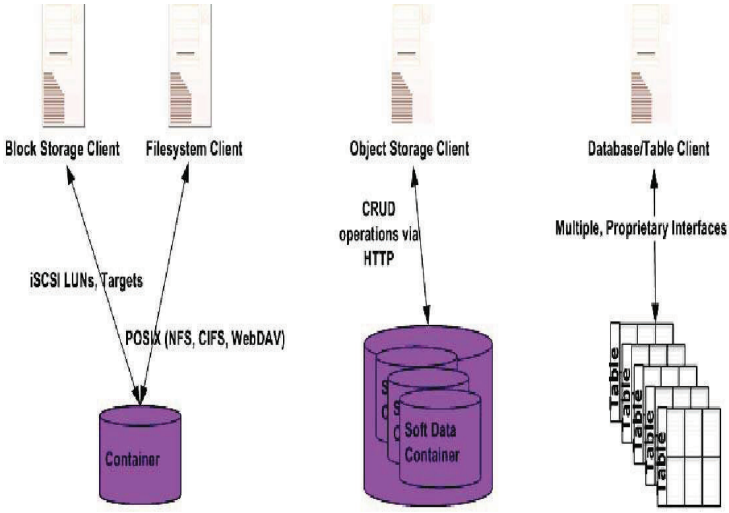


Figure 4.4 Existing Interface Standards for Data Storage

Standards exist and are emerging for interoperability between these elements; however, what is missing is a comprehensive description of where interoperability is needed and where standards can be best applied. All of these interfaces use parts of SNIA's Resource Domain model, which sets out these elements and describes a logical view of their functions and capabilities using a descriptive taxonomy. The purpose of this model is to retain the simplicity of the abstract cloud model by using metadata to drive the underlying services, so that users do not have to manage the service themselves. With a model in place, it's possible to identify existing standards, where appropriate, and identify areas where new standards might be needed. Although a variety of vendors offer "open" licenses for cloud storage interfaces and sets of preexisting libraries that provide similar functionality, no vendor wants a competitor to control the specification of the interface. In addition, multiple "standards" will proliferate, locking users into proprietary architectures. The SNIA's response has been to develop, using this Resource Domain model, the CDMI, an extensible standard that accommodates vendors' requirements and ensures consistency and interoperability for users.

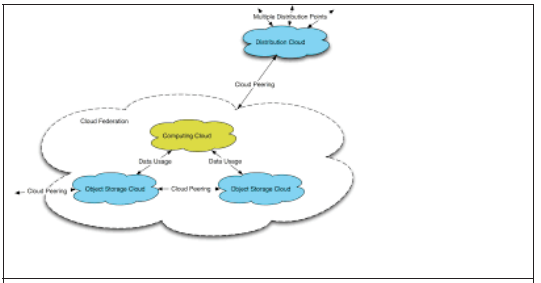


Figure 4.5 Cloud Data Management Interface

4.8 The Complete Picture- Cloud Storage Reference Model

The appeal of cloud storage is due to some of the same attributes that define other cloud services: pay as you go, the illusion of infinite capacity (elasticity), and the simplicity of use/management. It is therefore important that any interface for cloud storage support these attributes, while allowing for a multitude of business cases and offerings, long into the future.
The model created and published by the Storage Networking Industry Association (SNIA) shows multiple types of cloud data storage interfaces are able to support both legacy and new applications. All of the interfaces allow storage to be provided on demand, drawn from a pool of resources. The capacity is drawn from a pool of storage capacity provided by storage services. The data services are applied to individual data elements as determined by the data system metadata. Metadata specifies the data requirements on the basis of individual data elements or on groups of data elements (containers).
As shown in Fig. 4.5, Cloud Data Management Interface (CDMI) is the functional interface that applications will use to create, retrieve, update and delete data elements from the cloud. As part of this interface the client will be able to discover the capabilities of the cloud storage offering and use this interface to manage containers and the data that is placed in them. In addition, metadata can be set on containers and their contained data elements through this interface. It is expected that the interface will be able to be implemented by the majority of existing cloud storage offerings today. This can be done with an adapter to their existing proprietary interface, or by implementing the interface directly. In addition, existing client libraries such as extensible Access Method (XAM) can be adapted to this interface as show in Fig. 4.6.
This interface is also used by administrative and management applications to manage containers, accounts, security access, monitoring/billing information and even for storage that is accessible by other protocols. The capabilities of the underlying storage and data services are exposed so that clients can understand the offering. Conformant cloud offerings may offer a subset of either interface as long as they expose the limitations in the capabilities part of the interface.

4.9 How CDMI Works

Providing both a data path to the cloud service and a management path for the cloud data, CDMI is the functional interface that applications use to create, retrieve, update, and delete data elements in the cloud. As part of this interface, the client will be able to discover the capabilities of the cloud storage offering and use this interface to manage containers and the data that is placed in them. The semantics of CDMI are straightforward; simple containers and data objects are tagged with metadata—some of which are metadata that describe the data requirements of the object or container. The protocol for accessing the data and metadata is RESTful HTTP, first outlined by Roy Fielding (see Figure 4.6 – CDMI Basic Data Flow).

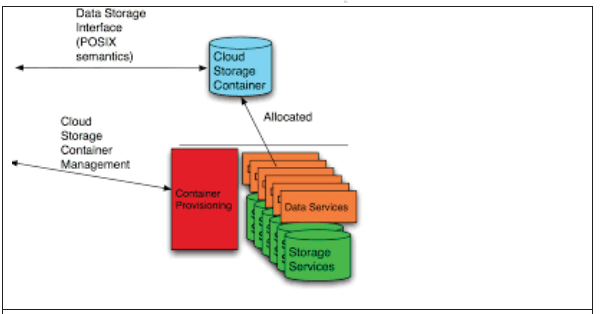


Figure 4.6 CDMI Basic Data Flow

Using RESTful techniques allows CDMI to provide:

- ✓ Unique names – Every object is addressable by a unique identifier.
- ✓ Uniformity – The interface uses only HTTP verbs with other semantics carried in the data model.
- ✓ Simplicity – The complexity is encapsulated in the representations.
- ✓ Statelessness – A lack of persistent client-side connections simplifies implementation.
- ✓ The format of the representations (the data carried by the HTTP requests) is in JavaScript Object Notation (JSON) that allows great flexibility, readability by both humans and machines, and extensibility.
- ✓ This flexibility gives CDMI the following advantages as an interface specification:
- ✓ First, it is well supported by many infrastructures and programming languages.
- ✓ Second, a small learning curve should encourage adoption and an "ecosystem" of support and code around the API.

Vendors and users will find these reasons compelling when developing a cloud using this new, open interface.

4.10 Cloud Storage API

A Cloud Storage Application Programming Interface (API) is a method for access to and utilization of a cloud storage system. The most common of these kinds are REST (REpresentational State Transfer) although there are others, which are based on SOAP (Simple Object Access Protocol). All these APIs are associated with establishing requests for service via the Internet. REST is a concept widely recognized as an approach to "quality" scalable API design.

One of the most important features of REST is that it is a "stateless" architecture. This means that everything needed to complete the request to the storage cloud is contained in the request, so that a session between the requestor and the storage cloud is not required. It is very important because the Internet is highly latent (it has an unpredictable response time and it is generally not fast when compared to a local area network). REST is an approach that has very high affinity to the way the Internet works. Traditional file storage access methods that use NFS (network files system) or CIFS (Common Internet File System) do not work over the Internet, because of latency.

Cloud Storage is for files, which, some refer to as objects, and others call unstructured data. Think about the files stored on your PC, like pictures, spreadsheets and documents. These have an extraordinary variability, thus unstructured. The other kind of data is block or structured data. Think data base data, data that feeds transactional system that require a certain guaranteed or low-latency performance. Cloud Storage is not for this use case. Industrial Design Centre (IDC) estimates that approximately 70% of the machine stored data in the world is unstructured, and this is also the fastest growing data type. So, Cloud Storage is storage for files that is easily accessed via the Internet. This does not mean you cannot access Cloud Storage on a private network or LAN, which may also provide access to a storage cloud by other approaches, like NFS or CIFS. It does mean that the primary and preferred access is by a REST API.

REST APIs are language neutral and therefore can be leveraged very easily by developers using any development language they choose. Resources within the system may be acted on through a URL. So, an API is not a "programming language", but it is the way a programming language is used to access a storage cloud. REST APIs are also about changing the state of resource through representations of those resources. They are not about calling web service methods in a functional sense. The key differences between different Cloud Storage APIs are the URLs defining the resources and the format of the representations. Amazon S3 APIs, Eucalyptus APIs, Rackspace Cloud Files APIs, Mezeo APIs, Nivanix APIs, Simple Cloud API, along with the standards proposed by the Storage Networking Industry Association (SNIA) Cloud Storage Technical Work Group, and more.

4.11 Applications for Cloud Storage

Before investing in cloud storage, providers and users must understand the options for cloud storage and how to create a strong business case. Some obvious and specific use cases that lend themselves to cloud storage include backup, archive, and application data storage.

Backup - In all organizations, business-critical data must be secure, available at short notice, and restorable to a specific time in the past. Traditionally, organizations use backup software and agents on file servers or desktops to back up their data to a mixture of specific disk systems and tape. Using cloud storage, organizations would use the same or similar backup software and agents, but instead, would back up their data to the cloud, which should have sufficient capacity and moderate latency to meet backup and recovery objectives.

When moving backup applications to the cloud, considerations include:

- ✓ Cost – Is cloud storage less expensive over time than existing alternatives?
- ✓ Capacity – Can the cloud handle the required daily, weekly, and monthly capacities and provide enough capacity for the extended periods that backup data is often held?
- ✓ Latency – Is latency low enough to meet backup and recovery objectives, but not so low as to make cloud storage too expensive?
- ✓ Security and privacy – Is the data secured from tampering or third-party access? Does the legal jurisdiction where the service is provided meet privacy requirements?
- ✓ Cloud storage can offer many advantages, not only lower costs, but also reduced chargeback's to business units.
- ✓ With cloud storage, organizations shift the burden of meeting Service Level Agreements (SLAs) to a service provider.
- ✓ Traditional backup environments require capital investment; the largest users are often the only ones who experience the economies of scale and efficiencies that traditional solutions offer.
- ✓ With cloud storage, users avoid overbuying capacity to lessen their risk of running out of capacity, which provides a "quick fix" but at a higher price.
- ✓ A public backup cloud can help turn capital expenses into operating expenses.
- ✓ A private backup cloud has the potential to leverage a dedicated backup cloud in an off-site location, improving the security of disaster recovery.
- ✓ A hybrid backup cloud allows managing backups to local, public, and/or private clouds to meet the varying requirements of cost, availability, latency, and security.

Archive - Organizations are increasingly being forced to retain larger volumes of data for longer periods, from decades to a century or more. The traditional approach is to back up long-term retention data to an external storage media, often tape, and keep it stored off site. Using cloud storage, backup data may be sent to a cloud archive that provides low-cost, high-capacity archive storage. When moving long-term retention data to the cloud, considerations include:

- ✓ Cost – Is cloud storage less expensive over time than long-term media such as tape?
- ✓ Capacity and duration – Can the cloud handle the volume and proposed data retention requirements?
- ✓ Refresh – Since technology rarely lasts a century, can the data be easily retrieved and moved to another storage medium in the future?
- ✓ Security and privacy – Is the data secured from tampering or third-party access? Does the legal jurisdiction where the service is provided meet privacy requirements?

Cloud storage can offer many advantages, not only lower costs, but also reduced chargeback's to business units.

- ✓ Latency is less important than cost, given that archive data is not normally required immediately.
- ✓ Cloud storage allows users to shift the burden of meeting compliance to a service provider, whereas a traditional environment requires users to have long-term archive expertise and to comply with regulatory, compliance, and legal requirements.
- ✓ A private backup cloud has the potential to leverage a dedicated archive cloud that is off site, thereby improving security and making it unnecessary to move tapes.
- ✓ A hybrid backup cloud allows different models for different regions of the world or business units to comply with regulatory, compliance, or legal requirements.

4.12 Application Data Storage

Business-critical applications and supporting applications require temporary and permanent data storage, which is normally supplied by internal local disk, external local disk, NAS, or SAN. Apart from the challenges posed by backing up user data from a wide variety of disparate sources, application data has stricter access and latency requirements.
When moving application data to the cloud, some considerations include:

- ✓ Cost – Is cloud storage less expensive than internal local disk, external local disk, SAN or NAS?
- ✓ Latency – Does cloud storage provide the kind of latency that applications and users expect?
- ✓ Accessibility – Is the data available from multiple locations?
- ✓ Security and privacy – Is the data secured from tampering or third-party access? Does the legal jurisdiction where the service is provided meet privacy requirements?
- ✓ Cloud storage can offer many advantages, including lower costs that result in reduced chargeback's to business units. Additional advantages include:
- ✓ Backup is more easily resolved when the primary application data is already in the cloud.
- ✓ Users shift the burden of meeting SLAs to a service provider.
- ✓ A private cloud has the potential to significantly reduce data and backup management costs.
- ✓ A hybrid backup cloud allows data to be close to the application where latency is critical.
- ✓ Cloud solutions can provide superior geographic latency, data protection, and recovery levels for certain distributed applications.

4.13 Other Applications for Cloud Storage

The examples above show some common themes when evaluating a cloud solution, but there are also many other applications for cloud storage. When assessing the viability of a cloud solution, consider the following business uses that also apply:

- ✓ Utility model – Does the cloud provider recognize the key role of storage in cloud
- ✓ Provisioning?
- ✓ Non-intrusive/non-disruptive change – Is it easy and cost effective to change existing

- ✓ Infrastructures to cloud storage, and does the cloud provider offer vendor-neutral support for various platforms?
- ✓ Rapid, flexible provisioning – Can the cloud solution quickly and non-disruptively expand,
- ✓ Shrink, provision, and de-provision storage on demand?
- ✓ Universal access – Can the cloud solution stores any type of application data and provides access through any standard network access protocols?
- ✓ Autonomic – Does the cloud solution provide always-on operation, zero- to low-touch
- ✓ Administration, fully automated management, and data mobility?
- ✓ Secure and protected data – Does the cloud solution provide controlled access to data with adequate privacy and security? Is the data protection policy widely enforced?
- ✓ Self-service – Does the cloud solution provide on-demand provisioning of capacity with pre assigned service levels and application- and OS-consistent restores?

In summary, cloud storage will prove its worth if it supports business as usual, but better.

4.14 File Storage in the Cloud

Following sections will introduce the key concepts of cloud storage, show their various benefits and discuss some of the current concerns about cloud storage.

4.14.1 General Architecture

Most cloud storage providers generally follow three-layer architecture. In the figure below, you can see an illustration of the general architecture and some of the characteristics that are tied to this architecture in current cloud storage.

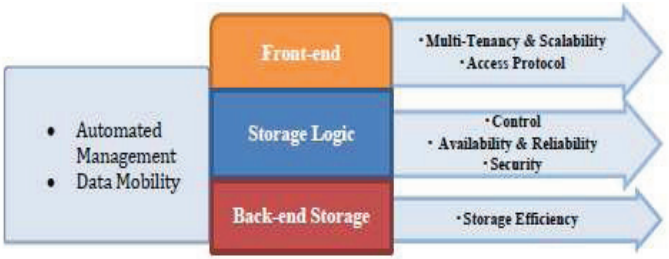


Figure 4.7 Generic cloud storage architecture

The front end is in charge of the communication between the clients and the servers. There will be different APIs to access the actual storage. This layer is also about achieving results such as multi-tenancy. In addition, it provides the means for different types of scalability through various methods.
The storage logic layer handles a variety of features, and is in charge of certain administrative procedures such as ensuring a high level of availability and reliability for instance. It is also a form of security perimeter. Furthermore, it acts like a controller for cloud storage.
The back-end focuses on the actual implementation of the physical storage of data with protocols such as the GFS (Google File System). It involves the use of various ways to increase storage efficiency and in a way to drive the infrastructure costs down.

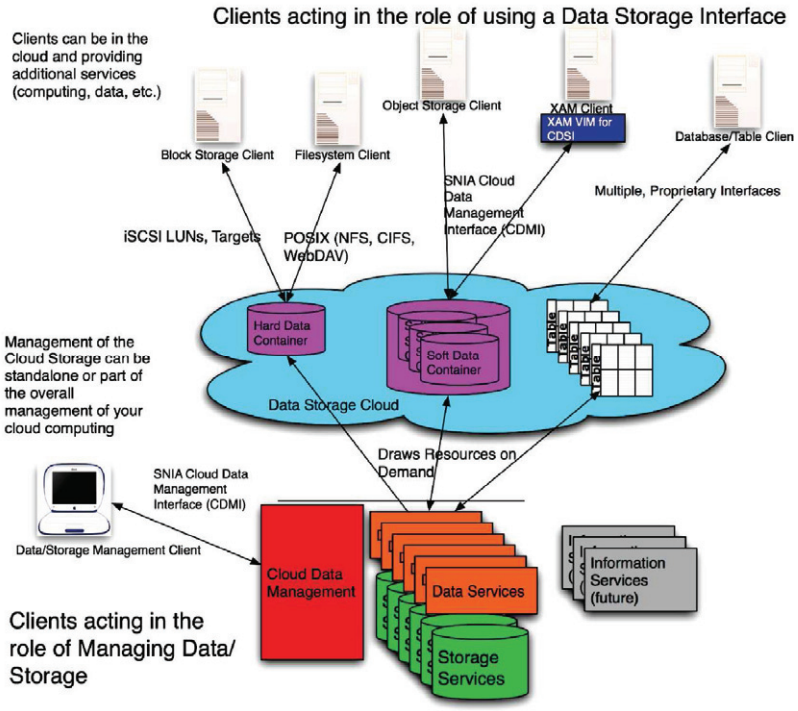


Figure 4.8 The Cloud Storage Reference Model

4.14.2 Defining Characteristics

Multi-tenancy, which refers to the ability for a single instance of services to serve multiple clients or tenants, also applies to several different layers of the cloud storage stack and this allows numerous clients to subscribe to the same cloud computing capabilities while retaining privacy and security over their sensitive data.

Automated Management is an important quality of the cloud storage. Generally, cost can be divided into two categories: the cost of the physical storage infrastructure itself and the cost of managing it. The management cost is hidden but is really a substantial component of the overall cost in the long run. The cloud storage must be able to add new storage and automatically configure itself to accommodate it and to find errors automatically. Automated management is relatively critical to cloud storage because what cloud computing is selling is essentially convenience.

Consistency in performance around the globe is one of the core reasons to choose cloud storage over traditional file hosting. With traditional file hosting, files are typically stored on one server hence clients who are far away from that server will suffer from bad performance. With cloud storage, there are 2 levels of geographical scalability. Firstly, the file is distributed around multiple servers in the region where your original data is stored at. Secondly, there are on-demand CDNs (content delivery networks). These are networks that have servers distributed globally to allow fast content delivery to clients anywhere in the world. By using CDNs, cloud storage can also achieve the same high level of consistency in performance all around the world and also make your data more mobile because it is available and highly accessible at all parts of the world.

Unique access methods are also one of the main differences between cloud storage and traditional storage. Many cloud storage providers now implement multiple access methods but the most prevalent one is still the Web-Service API. These are implemented by following the REST (Representational State Transfer) architecture. The architecture is used to develop protocols over the layer of HTTP to harness HTTP as a transport utility. By following this architecture, APIs are stateless and therefore relatively efficient. Bigger cloud storage providers such as Amazon (S3) and Microsoft Azure are both currently using this approach. There are also other forms of access methods such as file based APIs such as NFS and FTP and these two APIs are adopted by IBM Smart Business Storage Cloud.

High Reliability is one of the cornerstones of cloud storage. One might think that with the technological advances today, hard-disk failures and mass information losses are no longer common. On the contrary, hardware failures are inevitable and could be devastating if backups were not adequate. Cloud providers generally use two different approaches to ensure reliability.

Replication: Big cloud service providers generally have the same information stored on multiple machines. In the case of Google, their cloud back-end storage is typically split into huge clusters and entirely broken into chunks of 64mb each. Each of these chunks is uniquely identifiable and they are replicated to multiple servers in their data centers. Furthermore, these machines are run on different power supplies. That way, even if one power supply fails, clients will still have access to their data.

Reconstruction: Some service providers also use data-reconstruction algorithms to help with lost or damaged data. One of these algorithms is IDA (Information Dispersal Algorithm). This algorithm is able to construct a full set of data from multiple parts of the data that has been distributed before-hand. For example, if the data is divided into 4 parts, it can still be reconstructed if one site holding one part of the data is lost. Different ratios are possible to implement as well. E.g. 20 parts will allow 8 failed sites. These pieces of data are usually distributed at different geographical locations to reduce the chances of all parts of the data being lost at one time.

Good cost-to-storage ratio is another characteristic of cloud storage that is worth mentioning. To reduce cost, more data must be stored with the same hardware resources. One

common way to do this is to use data-reduction algorithms to reduce the resources data take up. There are notably 2 different approaches to this: compression- the encoding of the data in another more economical representation to achieve data reduction, de-duplication- the removal of any identical copies of data found through the scanning of data signatures.

High levels of security are essential to cloud storage, in particular, when storing sensitive data on the cloud.

4.14.3 Concerns about Cloud Storage

The integration of cloud computing technologies to existing IT infrastructure and the performance and latency of cloud storage are two concerns that are specific to cloud storage.

4.14.3.1 Integration

Before utilizing the cloud storage, an organization will need to integrate the cloud storage into their existing work-flow or other forms of offline storage facilities. The fact is normal file servers and cloud storage services do not use the same file access protocols. Servers use block protocol access to their storage, but cloud storage services generally only provides web protocol access such as REST-based APIs, SOAP-based APIs which are APIs designed on-top of the HTTP protocol to provide access with better efficiency. Each of the major providers has their separate set of APIs to handle the operations. This complicates things a little.
Mature organizations generally have more complicated existing file storage workflows. A considerable amount of time, money and attention will have to be spent to integrate the use of cloud storage into their existing workflows. On the other hand, a younger organization with a less complex infrastructure will not face the same problem because it will certainly be easier to integrate cloud storage into a workflow that is not so developed yet.

4.14.3.2 Performance and Latency

Cloud storage may be used by organizations for periodic backups of massive amounts of data. These back-up operations will involve sending data to a geographically distant location. This will inevitably be slower compared to offline storage solutions. While cloud storage is more convenient to use, is immediately scalable for organizations and more reliable, but unfortunately speed-wise it still trails behind offline storage solutions.
In general, cloud storage today is targeted at less performance demanding operations. Organizations should generally leave the operations having a stringent requirement for performance outside of cloud-storage. These include real-time transactions in banks for example.

4.14.4 Cloud Storage Providers

In this section, we will look at some examples of Cloud Storage services, with services from Amazon. Here are some services that AWS has to offer:

4.14.4.1 Amazon Elastic Block System (EBS)

Amazon Elastic Block Store (Amazon EBS) provides persistent block level storage volumes for use with Amazon EC2 instances in the AWS cloud. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from Component failure, offering high availability and durability. Amazon EBS volumes offer the consistent and low-latency performance needed to run your workloads. With Amazon EBS, you can scale your usage up or down within minutes—all while paying a low price for only what you provision.

Benefits

Reliable, secure storage- Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure. Amazon EBS encryption provides seamless support for data at rest security and data in motion security between EC2 instances and EBS volumes. Amazon's flexible access control policies allow you to specify who can access which EBS volumes. Access control plus encryption offers a strong defence-in-depth security strategy for your data.

Consistent and low-latency performance-Amazon EBS General Purpose (SSD) volumes and Amazon EBS Provisioned IOPS (SSD) volumes deliver low-latency through SSD technology and consistent I/O performance scaled to the needs of your application. Stripe multiple volumes together to achieve even higher I/O performance.

Backup, restore, and innovate -Backup your data by taking point-in-time snapshots of your Amazon EBS volumes. Boost the agility of your business by using Amazon EBS snapshots to create new EC2 instances.

Quickly scale up, easily scale down- Increase or decrease block storage and performance within minutes, enjoying the freedom to adjust as your needs evolve. Commission thousands of volumes simultaneously.

Geographic flexibility- Amazon EBS provides the ability to copy snapshots across AWS regions, enabling geographical expansion, data center migration, and disaster recovery.

EBS Features

Amazon EBS Snapshots -Amazon EBS provides the ability to save point-in-time snapshots of your volumes to Amazon S3. Amazon EBS Snapshots are stored incrementally: only the blocks that have changed after your last snapshot are saved, and you are billed only for the changed blocks. If you have a device with 100 GB of data but only 5 GB has changed after

your last snapshot, a subsequent snapshot consumes only 5 additional GB and you are billed only for the additional 5 GB of snapshot storage, even though both the earlier and later snapshots appear complete.

When you delete a snapshot, you remove only the data not needed by any other snapshot. All active snapshots contain all the information needed to restore the volume to the instant at which that snapshot was taken. The time to restore changed data to the working volume is the same for all snapshots. Snapshots can be used to instantiate multiple new volumes, expand the size of a volume, or move volumes across Availability Zones. When a new volume is created, you may choose to create it based on an existing Amazon EBS snapshot. In that scenario, the new volume begins as an exact replica of the snapshot.

The following are key features of Amazon EBS Snapshots:

Immediate access to Amazon EBS volume data - After a volume is created from a snapshot, there is no need to wait for all of the data to transfer from Amazon S3 to your Amazon EBS volume before your attached instance can start accessing the volume. Amazon EBS Snapshots implement lazy loading, so that you can begin using them right away.
Resizing Amazon EBS volumes - When you create a new Amazon EBS volume based on a snapshot, you may specify a larger size for the new volume. Make certain that your file system or application supports resizing a device.
Sharing Amazon EBS Snapshots - Amazon EBS Snapshots' shareability makes it easy for you to share data with your co-workers or others in the AWS community. Authorized users can create their own Amazon EBS volumes based on your Amazon EBS shared snapshots; your original snapshot remains intact. If you choose, you can also make your data available publicly to all AWS users.

Copying Amazon EBS Snapshots across AWS regions - Amazon EBS's ability to copy snapshots across AWS regions makes it easier to leverage multiple AWS regions for geographical expansion, data center migration and disaster recovery. You can copy any snapshot accessible to you: snapshots you created; snapshots shared with you; and snapshots from the AWS Marketplace, VM Import/Export, and AWS Storage Gateway.

Amazon EBS-Optimized Instances -For an additional low, hourly fee, customers can launch certain Amazon EC2 instance types as EBS-optimized instances. EBS-optimized instances enable EC2 instances to fully use the IOPS provisioned on an EBS volume.

EBS-optimized instances deliver dedicated throughput between Amazon EC2 and Amazon EBS, with options between 500 and 4,000 Megabits per second (Mbps) depending on the instance type used. The dedicated throughput minimizes contention between Amazon EBS I/O and other traffic from your EC2 instance, providing the best performance for your EBS volumes. EBS-optimized instances are designed for use with all Amazon EBS volume types.
Amazon EBS Availability and Durability -Amazon EBS volumes are designed to be highly available and reliable. At no additional charge to you, Amazon EBS volume data is replicated across multiple servers in an Availability Zone to prevent the loss of data from the failure of any single component. Amazon EBS volumes are designed for an annual failure rate (AFR) of between 0.1% - 0.2%, where failure refers to a complete or partial loss of the volume, depending on the size and performance of the volume. This makes EBS volumes 20 times more reliable than typical commodity disk drives, which fail with an AFR of around 4%. EBS also supports a snapshot feature, which is a good way to take point-in-time backups of your data.

Amazon EBS Encryption and AWS Identity and Access Management -Amazon EBS encryption offers seamless encryption of EBS data volumes and snapshots, eliminating the need to build and manage a secure key management infrastructure. EBS encryption enables data at rest security by encrypting your data volumes and snapshots using Amazon-managed keys or keys you create and manage using the AWS Key Management Service (KMS). In addition, the encryption occurs on the servers that host EC2 instances, providing encryption of data as it moves between EC2 instances and EBS data volumes. Access to Amazon EBS volumes is integrated with AWS Identity and Access Management (IAM). IAM enables access control to your Amazon EBS volumes. For more information, see AWS Identity and Access Management.

Functionality

AWS IAM allows you to:

Manage IAM users and their access – You can create users in IAM, assign them individual security credentials (in other words, access keys, passwords, and multi-factor authentication devices), or request temporary security credentials to provide users access to AWS services and resources. You can manage permissions in order to control which operations a user can perform.

Manage IAM roles and their permissions – You can create roles in IAM and manage permissions to control which operations can be performed by the entity, or AWS service, that assumes the role. You can also define which entity is allowed to assume the role.

Manage federated users and their permissions – You can enable identity federation to allow existing identities (e.g. users) in your enterprise to access the AWS Management Console, to call AWS APIs, and to access resources, without the need to create an IAM user for each identity.

4.14.4.2 Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) provides developers and IT teams with safe, secure, highly-scalable object storage. Amazon S3 provides a simple web-services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. Amazon S3 can be used alone or together with Amazon EC2/EBS, Amazon Glacier, and third-party storage repositories and gateways to provide cost-effective object storage for a wide variety of use cases including cloud applications, content distribution, backup and archiving, disaster recovery, and big data analytics. Amazon S3 stores data as objects within resources called buckets. You can store as many objects as you want within a bucket, and write, read, and delete objects in your bucket. Objects can be up to 5 terabytes in size. You can control access to the bucket (for example, which can create, delete, and retrieve objects in the bucket), view access logs for the bucket and its objects, and choose the AWS region where Amazon S3 will store the bucket and its contents.

Use Cases

Backup - Amazon S3 offers a highly durable, scalable, and secure solution for backing up and archiving your critical data. You can use Amazon S3's versioning capability to provide even further protection for your stored data. You can also define lifecycle rules to archive sets of Amazon S3 objects to Amazon Glacier, an extremely low-cost storage service.

Content Storage and Distribution - Amazon S3 provides highly durable and available storage for a variety of content. It allows you to offload your entire storage infrastructure into the cloud, where you can take advantage of Amazon S3's scalability and pay-as-you-go pricing to handle your growing storage needs. You can distribute your content directly from Amazon S3 or use Amazon S3 as an origin store for delivery of content to your Amazon CloudFront edge locations.

Big Data Analytics - Whether you're storing pharmaceutical or financial data, or multimedia files such as photos and videos, Amazon S3 is the ideal big data object store. AWS offers a comprehensive portfolio of services to help you manage big data by reducing costs, scaling to meet demand, and increasing the speed of innovation.

Static Website Hosting - You can host your entire static website on Amazon S3 for a low-cost, highly available hosting solution that scales automatically to meet traffic demands. With Amazon S3, you can reliably serve your traffic and handle unexpected peaks without worrying about scaling your infrastructure.

Cloud-native Application Data - Amazon S3 provides high performance, highly available storage that makes it easy to scale and maintain cost-effective mobile and Internet-based apps that run fast. With Amazon S3, you can add any amount of content and access it from anywhere, so you can deploy applications faster and reach more customers.

Disaster Recovery - Amazon S3's highly durable, secure, global infrastructure offers a robust disaster recovery solution designed to provide superior data protection. Whether you're looking for disaster recovery in the cloud or from your corporate data center to Amazon S3, AWS has the right solution for you.

Key Features

Security and Access Management - Amazon S3 provides several mechanisms to control and monitor who can access your data as well as how, when, and where they can access it.

Lifecycle Management - Amazon S3 provides a number of capabilities to manage the lifecycle of your data, including automated archival using the lower-cost Amazon Glacier.

Versioning - Amazon S3 allows you to enable versioning so you can preserve, retrieve, and restore every version of every object stored in an Amazon S3 bucket.

Encryption - You can securely upload/download your data to Amazon S3 via SSL-encrypted endpoints. Amazon S3 also provides multiple options for encryption of data at rest, and allows you to manage your own keys or have Amazon S3 manage them for you.
Cost Monitoring and Controls - Amazon S3 has several features for managing and controlling your costs, including bucket tagging to manage cost allocation and integration with Amazon Cloud Watch to receive billing alerts.

Choice of AWS Region - Amazon S3 is available globally in multiple AWS regions. You can choose the region where a bucket is stored to optimize for latency, minimize costs, or address regulatory requirements.

Programmatic Access Using the AWS SDKs - Amazon S3 is supported by the AWS SDKs for Java, PHP, .NET, Python, Node.js, Ruby, and the AWS Mobile SDK. The SDK libraries wrap the underlying REST API, simplifying your programming tasks.

Transfer Data to and from Amazon S3 with Ease AWS - supports several methods for uploading and retrieving data in Amazon S3 including the public Internet, AWS Direct Connect, and the AWS Import/Export service. And AWS Storage Gateway automatically backs up on-premises data to Amazon S3.

Flexible Storage Options - Amazon S3 is designed for 99.999999999% durability and 99.99% availability of objects over a given year. There is also a low-cost Reduced Redundancy Storage option for less critical data and Amazon Glacier for archiving cold data at the lowest possible cost.

4.14.4.3 Amazon Import/Export

AWS Import/Export accelerates moving large amounts of data into and out of the AWS cloud using portable storage devices for transport. AWS Import/Export transfers your data directly onto and off of storage devices using Amazon's high-speed internal network and bypassing the Internet. For significant data sets, AWS Import/Export is often faster than Internet transfer and more cost effective than upgrading your connectivity. AWS Import/Export supports data transfer into and out of Amazon S3 buckets in the US East (N. Virginia), US West (Oregon), US West (Northern California), EU (Ireland), and Asia Pacific (Singapore) regions.

Common Use Cases for AWS Import/Export

Data Cloud Migration - If you have data you need to migrate into the AWS cloud for the first time, AWS Import/Export is often much faster than transferring that data via the Internet. Content Distribution - Send data to your customers on portable storage devices.

Direct Data Interchange - If you regularly receive content on portable storage devices from your business associates, you can have them send it directly to AWS for import into Amazon S3 or Amazon EBS or Amazon Glacier.

Offsite Backup - Send full or incremental backups to Amazon S3 and Amazon Glacier for reliable and redundant offsite storage.

Disaster Recovery - In the event you need to quickly retrieve a large backup stored in Amazon S3 or Amazon Glacier, use AWS Import/Export to transfer the data to a portable storage device and deliver it to your site.

4.14.4.4 Amazon Storage Gateway

AWS Storage Gateway is a service connecting an on-premises software appliance with cloud-based storage to provide seamless and secure integration between an organization's on-premises IT environment and AWS's storage infrastructure. The service allows you to securely store data in the AWS cloud for scalable and cost- effective storage. The AWS Storage Gateway supports industry-standard storage protocols that work with your existing applications. It provides low-latency performance by maintaining frequently accessed data on-premises while securely storing all of your data encrypted in Amazon S3 or Amazon Glacier.

The AWS Storage Gateway supports three configurations:

Gateway-Cached Volumes: You can store your primary data in Amazon S3, and retain your frequently accessed data locally. Gateway-Cached volumes provide substantial cost savings

on primary storage, minimize the need to scale your storage on-premises, and retain low-latency access to your frequently accessed data.

Gateway-Stored Volumes: In the event you need low-latency access to your entire data set, you can configure your on-premises data gateway to store your primary data locally, and asynchronously back up point-in-time snapshots of this data to Amazon S3. Gateway-Stored volumes provide durable and inexpensive off-site backups that you can recover locally or from Amazon EC2 if, for example, you need replacement capacity for disaster recovery.

Gateway-Virtual Tape Library (Gateway-VTL): With Gateway-VTL you can have a limitless collection of virtual tapes. Each virtual tape can be stored in a Virtual Tape Library backed by Amazon S3 or a Virtual Tape Shelf backed by Amazon Glacier. The Virtual Tape Library exposes an industry standard iSCSI interface which provides your backup application with on-line access to the virtual tapes. When you no longer require immediate or frequent access to data contained on a virtual tape, you can use your backup application to move it from its Virtual Tape Library to your Virtual Tape Shelf in order to further reduce your storage costs.

Benefits

Secure - The AWS Storage Gateway securely transfers your data to AWS over SSL and stores data encrypted at rest in Amazon S3 and Amazon Glacier using Advanced Encryption Standard (AES) 256, a secure symmetric-key encryption standard using 256-bit encryption keys.

Durably backed by Amazon S3 and Amazon Glacier - The AWS Storage Gateway durably stores your on-premises application data by uploading it to Amazon S3 and Amazon Glacier. Amazon S3 and Amazon Glacier redundantly store data in multiple facilities and on multiple devices within each facility. Amazon S3 and Amazon Glacier also perform regular, systematic data integrity checks and are built to be automatically self-healing.

Compatible -There is no need to re-architect your on-premises applications. Gateway-Cached volumes and Gateway-Stored volumes expose a standard iSCSI block disk device interface and Gateway-VTL presents a standard iSCSI virtual tape library interface.

Cost-Effective - By making it easy for your on-premises applications to store data on Amazon S3 or Amazon Glacier, AWS Storage Gateway reduces the cost, maintenance, and scaling challenges associated with managing primary, backup and archive storage environments. You pay only for what you use with no long-term commitments.

Designed for use with other Amazon Web Services - Gateway-Stored volumes and Gateway-Cached volumes are designed to seamlessly integrate with Amazon S3, Amazon EBS, and Amazon EC2 by enabling you to store point-in-time snapshots of your on-premises application data in Amazon S3 as Amazon EBS snapshots for future recovery on-premises or in Amazon EC2. This integration allows you to easily mirror data from your on-premises applications to applications running on Amazon EC2 in disaster recovery (DR) and on-

demand compute capacity cases. Gateway-VTL integrates with Amazon Glacier and allows you to cost effectively and durably store your archive and long-term backup data.

Optimized for Network Efficiency -The AWS Storage Gateway efficiently uses your internet bandwidth to speed up the upload of your on-premises application data to AWS. The AWS Storage Gateway only uploads data that has changed, minimizing the amount of data sent over the internet. You can also use AWS Direct Connect to further increase throughput and reduce your network costs by establishing a dedicated network connection between your on-premises gateway and AWS.

Common Use Cases

Backup -The AWS Storage Gateway enables your existing on-premise to cloud backup applications to store primary backups on Amazon S3's scalable, reliable, secure, and cost-effective storage service. You can create Gateway-Cached storage volumes and mount them as iSCSI devices to your on-premises backup application servers. All data is securely transferred to AWS over SSL and stored encrypted in Amazon S3 using AES 256-bit encryption. Using Gateway-Cached volumes provides an attractive alternative to the traditional choice of maintaining and scaling costly storage hardware on-premises.

For scenarios where you want to keep your primary data or backups on-premises, you can use Gateway-Stored volumes to keep this data locally, and backup this data off-site to Amazon S3. Gateway-Stored volumes provide an attractive alternative to dealing with the longer recovery times and operational burden of managing off-site tape storage for backups.

Use cases for backup and recovery

Databases -AWS storage solutions deliver highly scalable, durable, and reliable cloud storage for backup, and are designed to support mission-critical databases, including Oracle and SAP. With an easy to use web interface, Amazon S3 is designed to deliver flexibility, agility, geo-redundancy, and robust data protection.

Remote/branch office - Amazon S3 provides scalable and cost-effective storage on-demand for your organization's data. Using AWS Storage Gateway you can easily back up data from one location to another for quick recovery.

Media content - AWS storage solutions are designed to deliver capacity and flexibility to make backup and recovery easy for you. Amazon S3 provides scalable and cost-effective storage, on-demand, and eliminates all the heavy lifting of deploying infrastructure.

Disaster Recovery and Resilience - The AWS Storage Gateway, together with EC2, can mirror your entire production environment for disaster recovery (DR). Planning for business continuity in the event of a power outage, fire, flood, or other disaster can be challenging. It requires investments in redundant infrastructure and staff across multiple data center and costly storage replication solutions. AWS Storage Gateway and Amazon EC2 together

provide a simple cloud-hosted DR solution. Using Amazon EC2, you can configure virtual machine images of your DR application servers and only pay for these servers when you need them. In the event your on-premises infrastructure goes down, you simply launch the Amazon EC2 compute instances you need and attach them to copies of your on-premises data. The AWS Storage Gateway addresses the challenges of replicating data for DR by enabling you to create Gateway-Cached volumes that store your data in Amazon S3. By storing your data using the AWS Storage Gateway, you will be prepared for DR if you lose your on-premises application or storage.

Benefits of Using AWS for Disaster Recovery

First Performance -First disk based storage and retrieval of files.

No Tape - Eliminate costs associated with transporting, storing, and retrieving tape media and associated tape backup software.

Compliance - Fast retrieval of files allows you to avoid fines for missing compliance deadlines.

Elasticity - Add any amount of data, quickly. Easily expire and delete without handling media.

Secure - Secure and durable cloud disaster recovery platform with industry-recognized certifications and audits.

Partners - AWS solution provider and system integration partners to help with your deployment.

Corporate File Sharing - Managing on-premises storage for departmental file shares and home directories typically results in high capital and maintenance costs, under-utilized hardware, and restrictive user quotas. The AWS Storage Gateway addresses these on-premises scaling and maintenance issues by enabling you to seamlessly store your corporate file shares on Amazon S3, while keeping a copy of your frequently accessed files on-premises. This minimizes the need to scale your on-premises file storage infrastructure, while still providing low-latency access to your frequently accessed data. Using the AWS Storage Gateway, you can create Gateway-Cached storage volumes up to 32 TB in size and mount them as iSCSI devices from your on-premises file servers. You can then expose these volumes as Common Internet File System (CIFS) shares or Network File System (NFS) mount points to your client machines. The AWS Storage Gateway durably stores files written to these shares or mount points in Amazon S3, while maintaining a cache of recently written and recently read files locally on your on-premises storage hardware for low-latency access. Since you only pay for the storage you actually use, you can scale your storage on-demand and avoid the costs of under-utilized hardware.

Data Mirroring to Cloud-Based Compute Resources

If you want to leverage Amazon EC2's on-demand compute capacity for additional capacity during peak periods, for new projects, or as a more cost-effective way to run your normal workloads, you can use the AWS Storage Gateway to mirror your volume data to Amazon EC2 instances. If you're running development and User Acceptance Testing (UAT) environments in Amazon EC2 to take advantage of AWS's on-demand compute capacity, you can use the AWS Storage Gateway to ensure these environments have ongoing access to the latest data from your production systems on-premises.

Use Cases for Gateway–Virtual Tape Library

Magnetic Tape Replacement for Archiving and Long-Term Backup

Using Gateway-VTL, you can store data requiring long term retention and infrequent access without changing your existing backup applications and tape-based processes. Although magnetic tape-based storage can be cost-effective when operated at scale, it can be a drain on resources as one (or more) tape libraries need to be maintained (often in geographically distinct locations) requiring specialized personnel, and taking up valuable space in data centers. In addition, the tapes themselves must be carefully stored and managed, which can include periodically copying data from old tapes onto new ones to ensure that your data can still be read as tape technology standards evolve.

Tape's low cost potential also requires accurate capacity planning, a process that is usually error-prone, especially when storage growth is unpredictable, as it often is. Over provisioning capacity can result in underutilization and higher costs, while under provisioning can trigger expensive hardware upgrades far earlier than planned. Even when capacity planning is accurate, periodic hardware upgrades are still common as older tape libraries are less efficient and therefore costlier to operate. Archiving valuable data using a tape-based solution also requires costly, multi-site, redundant data centers and offsite vaulting to guarantee durability. This approach also requires manual handling of tape media which increases the risk of data loss.

By using Gateway-VTL, you can eliminate these challenges associated with owning and operating on-premises physical tape infrastructure by storing your archive and long-term backup data on a limitless collection of virtual tapes. Your virtual tapes can be stored in a Virtual Tape Library backed by Amazon S3 or a Virtual Tape Shelf backed by Amazon Glacier. The Virtual Tape Library provides your backup application with on-line access to the virtual tapes. When you no longer require immediate or frequent access to data contained on a virtual tape, you can use your backup application to move it from its Virtual Tape Library to your Virtual Tape Shelf in order to further reduce your storage costs.
Gateway-VTL allows you to eliminate the need for large upfront capital expense and expensive multi-year support commitments. With Gateway-VTL you pay only for the

capacity you use and scale as your needs grow. With the Gateway-VTL solution you also don't need to worry about transporting storage media to offsite facilities and manual handling of tape media. The Gateway-VTL solution reduces your costs and simplifies your data management process while improving the durability of your archive and long-term backup solution.

4.14.4.5 Amazon Glacier

Amazon Glacier is an extremely low-cost cloud archive storage service that provides secure and durable storage for data archiving and online backup. In order to keep costs low, Amazon Glacier is optimized for data that is infrequently accessed and for which retrieval times of several hours are suitable. With Amazon Glacier, customers can reliably store large or small amounts of data for as little as $0.01 per gigabyte per month, a significant savings compared to on-premises solutions. Companies typically over-pay for data archiving. First, they're forced to make an expensive upfront payment for their archiving solution (which does not include the on-going cost for operational expenses such as power, facilities, staffing, and maintenance). Second, since companies have to guess what their capacity requirements will be, they understandably over-provision to make sure they have enough capacity for data redundancy and unexpected growth. This set of circumstances results in under- utilized capacity and wasted money. With Amazon Glacier, you pay only for what you use. Amazon Glacier changes the game for data archiving and cloud backup as you pay nothing upfront, pay a very low price for storage, and can scale your usage up or down as needed, while AWS handles all of the operational heavy lifting required to do data retention well. It only takes a few clicks in the AWS Management Console to set up Amazon Glacier and then you can upload any amount of data you choose.

Benefits

Low Cost - Starting at $0.01 per gigabyte per month, Amazon Glacier allows you to archive large amounts of data at a very low cost. You pay for what you need, with no minimum commitments or up-front fees.

Secure- Amazon Glacier supports data transfer over SSL and automatically encrypts your data at rest. You can also control access to your data using AWS Identity and Access Management (IAM). For customers who must comply with regulatory standards such as PCI and HIPAA, Amazon Glacier's data protection features can be used as part of an overall strategy to achieve compliance. The various data security and reliability features offered by Amazon Glacier are described in detail below.

Encryption by default -Amazon Glacier automatically encrypts data at rest using Advanced Encryption Standard (AES) 256-bit symmetric keys and supports secure transfer of your data over Secure Sockets Layer (SSL).

Immutable archives - Data stored in Amazon Glacier is immutable, meaning that after an

archive is created it cannot be updated. This ensures that data such as compliance and regulatory records cannot be altered after they have been archived.

Flexible access control with IAM policies - Amazon Glacier supports Identity and Access Management (IAM) policies, which enables organizations with multiple employees to create and manage multiple users under a single AWS account. With IAM policies, you create fine-grained policies to control to your Amazon Glacier vaults. You can write IAM policies to selectively grant or revoke certain permissions and actions on each Amazon Glacier vault.

Mandatory request signing - Amazon Glacier requires all requests to be signed for authentication protection. To sign a request, you calculate a digital signature using a cryptographic hash function that returns a hash value that you include in the request as your signature. After receiving your request, Amazon Glacier recalculates the signature using the same hash function and input that you used to sign the request before processing the request.
#
Durable - Amazon Glacier provides a highly durable storage infrastructure designed for online backup and archival. Your data is redundantly stored across multiple facilities and multiple devices in each facility. Amazon Glacier provides a highly durable storage infrastructure designed for long-term data archival storage. It is designed to provide average annual durability of 99.999999999% for an archive. The service redundantly stores data in multiple facilities and on multiple devices within each facility. To increase durability, Amazon Glacier synchronously stores your data across multiple facilities before confirming a successful upload.

To prevent corruption of data packets over the wire, Amazon Glacier uploads the checksum of the data during data upload. It compares the received checksum with the checksum of the received data to detect bit flips over the wire. Similarly, it validates data authenticity with checksums during data retrieval. Unlike traditional systems, that can require laborious data verification and manual repair, Amazon Glacier performs regular, systematic data integrity checks and is built to be automatically self-healing.

Simple -Amazon Glacier allows you to offload the administrative burden of operating storage infrastructure to AWS. Data uploaded to Amazon Glacier remains stored for as long as needed with no additional effort from you.

Flexible - Amazon Glacier scales to meet your storage needs. There is no limit to how much data you can store, and you can choose to store your data in the AWS region that supports your regulatory and business criteria.

Integrated - Through Amazon S3 lifecycle policies, you can optimize your storage costs by moving infrequently accessed objects from Amazon S3 to Amazon Glacier (or vice-versa).

4.14.4.6 Other cloud storage providers

Some popular Cloud Storage providers include Microsoft Azure, JustCloud, zipCloud and livedrive. The providers I have just mentioned differs slightly from what a simple user might want, to simply store some of their files on their hardisk on Cloud. These providers allow us to access the files via methods that can be used in sites or web development. New users may not be so familiar with Cloud and may want to test Cloud out before joining the community. In that case, we recommend Amazon Web Services, that is both user friendly, a year with limited free usage, as well as many functionalities that help you start up.

4.15 Data Deduplication in cloud

To process and protect huge amounts of data efficiently, it is important to employ strategies such as data deduplication to improve both storage capacity and network bandwidth utilization. Further, for most organizations, especially small to medium businesses (SMBs), hosting such services in a public cloud proves to be more economical and efficient. In data deduplication, duplicate data is detected and only one copy of the data is stored, along with references to the unique copy of data, thus removing redundant data. Data deduplication can be performed at three levels, file level, block level (also called chunk level) and byte level, with chunk level being the most popular and widely deployed. For each of these deduplication types, files, data blocks or bytes are hashed and compared for redundancy detection. In general, there are four main steps in chunk level data deduplication, chunking, fingerprinting, index lookup and writing.

(i) **Chunking -** during the chunking stage, data is split into chunks of non-overlapping data blocks. The size of the data block can be either fixed or variable depending on the chunking method used. The Fixed Size Chunking (FSC) method is used in the case of fixed data blocks, whereas the common method used to produce variable sized chunks is Content Defined Chunking (CDC).

(ii) **Fingerprinting -** using a cryptographic hash function (e.g., SHA-1), a fingerprint is calculated for each chunk produced from the chunking phase.

(iii) **Index lookup –** A lookup table (chunk index) is created containing the fingerprints for each unique data chunk. A lookup operation is performed for each fingerprint generated in step (ii) to determine whether or not the chunk is unique. If the fingerprint is not found in the lookup table, it implies that the data chunk is unique. The fingerprint is thus inserted in the table and the chunk is written to the data store in step (iv).

(iv) **Writing –** all unique data chunks are written to the data store. Each chunk stored on the storage system using chunk-based deduplication has a unique fingerprint in the chunk index. To determine whether a given chunk is a duplicate or not, the fingerprint of the incoming data chunk is first looked up in the chunk index. Existence of a matching fingerprint (i.e., hash) in the index indicates that an identical copy of the incoming chunk already exists (i.e., has been

stored earlier) and the system therefore only needs to store a reference to the existing data. If there is no match, the incoming chunk is unique and is stored on the system and its fingerprint inserted in the chunk index accordingly. Deduplication can be performed either 'inline' as the data is entering the storage system/device in real time, or as a 'post-process' after the data has already been stored. Inline data duplication uses less storage space as the duplicate data is detected in real time before the data is stored.

Deduplication based cloud backup emerges as a suitable way of backing up huge amounts of data due to the advantages offered by both cloud computing and data deduplication. The customer can leverage the storage infrastructure offered by the cloud provider, whereas deduplication makes it possible for cloud provider to reduce storage capacity and network bandwidth requirements and optimize storage and networks. Backups benefit even more from deduplication because of the significant redundancy that exists between successive full backups of the same dataset.

While deduplication based cloud backup presents a good backup solution, it has its own shortcomings, with the major one being that of throughput. There are two important metrics that can be used in evaluating the performance of a backup system the backup window (BW) and recovery time objective (RTO). The backup window is the period of time in which backups are allowed to run and complete on a system, whereas RTO is the amount of time between when a disaster happened and the time the business functions are restored. In this thesis, we address two main problems faced by deduplication based cloud backup systems. Firstly, the chunk index and associated fingerprint lookups present a bottleneck to the throughput and scalability of the system.
And secondly, the constrained network bandwidth has a negative impact on the RTO.

(i) **Throughput and scalability of chunk index (fingerprint store) and fingerprint lookup:** during the backup operation, duplicate data is determined by first consulting the chunk index. For a lager data set, it's not possible to store the whole index in RAM, forcing the index lookup to go to the disk and incurring disk I/O penalties. The system incurs longer latency as a result of the costly disk I/O. Furthermore, in a public cloud environment, the system has to handle hundreds of thousands of concurrent backup clients, thereby putting additional pressure on the throughput and scalability of the backup system.

(ii) **The constrained bandwidth challenge:** backup is not the primary goal of a backup system but the means to the goal, which is the ability to restore the backed up data in a timely manner when it's needed. A recent study indicates that 58% of SMB cannot tolerate more than 4 hours of down time before they start experiencing negative effects on the business. A recent survey indicates that 87% of enterprises rank the ability to recover data in a quick and effective way to be very important. However, low bandwidth WAN links present a challenge to cloud backup services that are expected to provide fast data restorations. It is therefore important to devise and employ methods of restoring data that increase the effective throughput of the data restore process.

4.16 System Architechture

The main aim in designing this scheme is to design an improved technique for storage in Cloud computing .The overall architecture of the technique is shown in Figure 1 Overall architecture is divided into four layers,

Interface Layer - Interface layer provides the user interface to select the file for deduplication .It also provides interface to select the type of deduplication and also to specify the split length.

Chunk Layer - Based on the split length different Segments of file are created. For these chunks hash values are computed by using MD5 algorithm.

Algorithm to segment

File Algorithm1: File Segment
Input: File selected
Split Size
Output: Files divided into segments based on split size
Procedure: SegmentFile /*Enter the split size.*/ /*Read the input file and create the byte stream.*/
while (bytes != -1) {Divide the file into number of bytes specified by split length and create the new file with .txt extension}
end while /*Different segments of the file is created.*/
end

Deduplication Layer - It involves the detection of duplicate chunks by comparing hash values generated in the chunk layer. The chunks are compressed after eliminating the duplicate values.

Storage layer - After eliminating the duplicate values the compressed file is stored in the Amazon S3 bucket by using Amazon APIs. The compressed file is uploaded to Amazon storage.

Algorithm to upload file
Algorithm 2: Upload File
Input: File
Output: File uploaded to Cloud
Procedure: Upload
begin upload
/*Select file to upload*/
/*Select Amazon S3 client*/
/* provide AWS credential properties that is secret key and access key of Amazon client.*/
/*upload the file to specific bucket using S3.putobject( )*/
end

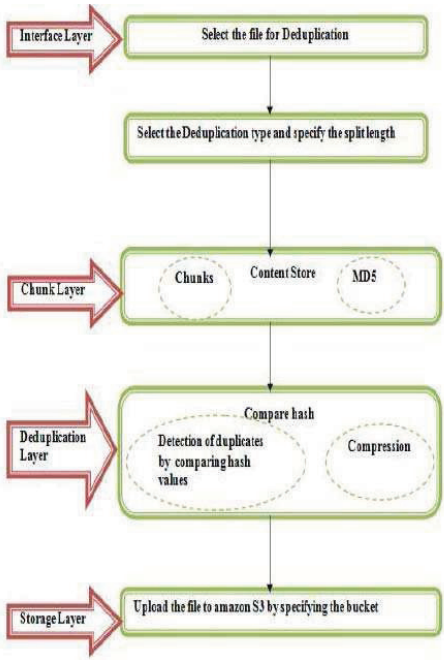


Figure 4.9 Overall architecture

CHAPTER 5

CLOUD DATABASE

5.1 Database in Cloud

Database is a key component in most computing infrastructures. Database allows users to store data in an organized manner and retrieve them easily. In this section, we will discuss a new type of database that is gaining popularity especially in cloud computing, the non-relational database (NOSQL), and compare it with the relational database, from a cloud perspective. We will also look at some common database architectures that cloud computing providers employ and issues regarding programming environments that come with the different types of database.

5.2 Non-Relational

Non-relational databases are commonly referred to by the term "NOSQL" (pronounced “No S Q L”). They are made up of individual tables and these tables cannot have defined relationships between them, unlike in relational databases. For example, in the database schema as shown in Figure 5.1, one can retrieve the account balance of a specific Customer given the Customer's name through table joins using SQL due to the Primary Key/Foreign Key relationship. In a non-relational database for the same schema, without the relationship, the developer has to use application code to obtain the Customer's account number and then access the Account table and match the account number obtained previously to retrieve the balance.

Table: Customer

Fields: name *PRIMARY KEY*, address, accountNumber

FOREIGN KEY accountNumber *REFERENCES* Account.number

Table: Account

Fields: number *PRIMARY KEY*, pinCode, balance

Figure 5.1 - Primary Key/Foreign Key relationship

5.3 Relational vs. Non-Relational

Many cloud computing providers offer users both relational and non-relational databases. Both types of database are scalable in the cloud and can be highly available. In terms of speed, we have done up our own series of benchmarking tests for some relational and non-relational databases cloud computing services. Our findings indicate that relational cloud databases perform create, update and delete operations faster than non-relational cloud databases. However, for read operations, non-relational cloud databases do perform better than relational cloud databases. Usage-wise, both relational and non-relational cloud

databases are as easy to use as the other. This is because cloud computing providers take on most of the burden of database administration, especially for relational databases, as relational databases usually come with heavy database administration workload compared to non-relational databases.

5.4 Architectures

Cloud databases providers often let users choose from multiple database architectures. Since these different architectures have different levels of database consistency, latency and costs, you need to understand the architectures to have a better idea of which service suits your application's needs. We will discuss two different architectures which are being used by major cloud service providers here - the Master/Slave architecture and the architecture based on the Paxos algorithm.

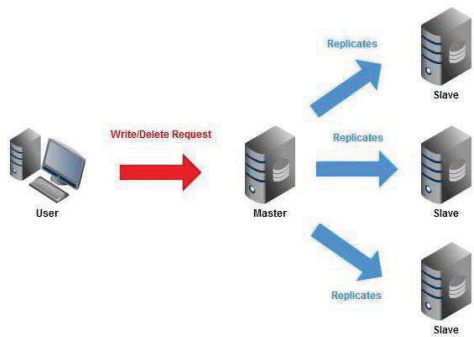


Figure 5.2 The Master/Slave database architecture

In the Master/Slave database architecture (see Figure 5.2); a database server acts as the Master. When the user sends in a write/delete request to his database, the request goes to the Master database server. The Master database server checks against and updates its own database and then asynchronously replicates the update in other Slave database servers.

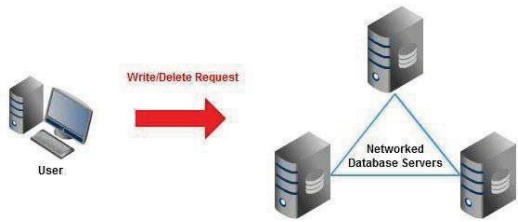


Figure 5.3 Paxos architecture

For the Paxos architecture (see Figure 5.3), when the user sends a write/delete request, this request goes to a network of several database servers. The different database servers will check the requests against their own databases and states and then communicate with each other to affirm the request.

There are numerous pros and cons of using one database architecture over the other.

- ✓ Master/Slave architectures uses lesser write/delete CPU time: Databases built on the Paxos architecture use more write CPU time than databases built on the Master/Slave architecture due to the servers needing to communicate with each other to affirm the write/delete request, unlike in the Master/Slave database architecture, where the Master itself affirms the request and sends the affirmed changes to the Slaves.

- ✓ Master/Slave architecture has lower write/delete latency: It is higher for the databases built on the Paxos architecture as the affirming of requests between the various servers takes time.

- ✓ Master/Slave architectures have stronger query consistency: Query consistency of databases built on Paxos architecture is "eventual" as they require time to process certain tricky requests among the data centres - a read request might come in before the processing of a previous write/delete request can be completed, resulting in the read request not getting the most up-to-date results.

- ✓ Paxos architectures have higher availability and reliability: Databases built on Paxos architecture do not suffer from downtimes like their counterparts built on Master/Slave architecture. For example, if the Master data centre for a certain database built on Master/Slave architecture goes down for maintenance, write/delete requests will not be processed. However for databases built on the Paxos architecture, the user's database can still be updated even if a data centre goes down, as long as there are other data centres that remain operational, since any of the data centres can process the write/delete request.

The overview of the two different architectures should now give you a better understanding of why some databases offered by cloud computing providers are more costly than others or why there is higher consistency for some types of database over the others for example. Moreover, the two types of architectures covered can also serve as examples for you to make use of infrastructures offered by cloud computing providers to model and build your very own cloud database architecture. For example in Figure 5.4, the Amazon Read Replica instances act as the slave databases in Master/Slave architecture.

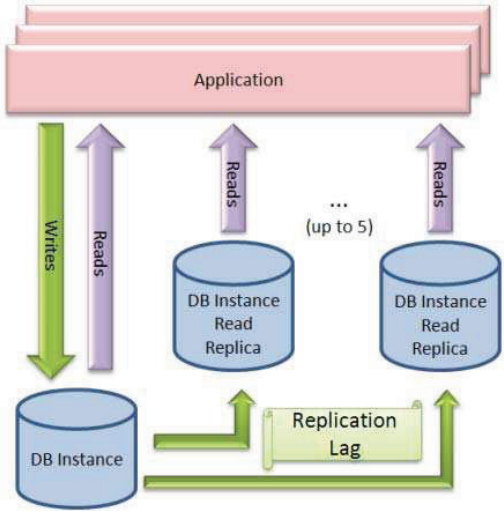


Figure 5.4 Build a Master/Slave database using Amazon RDS and its Read Replica complement

5.5 Examples of Cloud-based Database

From the discussion in previous part, we can see the differences between Relational Database and Non-Relational Database, as well as the differences between different architecture of Cloud Database: Master/Slave architecture and Paxos architecture. In this part, we will take a look at some specific examples of Cloud-based Database systems including: Amazon Relational Database Service (Relational database), Amazon Dynamo DB (NoSQL database), Google Datastore (NoSQL database) and Google Cloud SQL (Relational database).

5.5.1 Amazon RDS

Amazon Relational Database Service (Amazon RDS) is a web service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost- efficient and resizable capacity while managing time-consuming database management tasks, freeing you up to focus on your applications and business. Amazon RDS gives you access to the capabilities of a familiar MySQL, Oracle, SQL Server, PostgreSQL or Amazon Aurora relational database management system. This means that the code, applications, and tools you already use today with your existing databases can be used with Amazon RDS. Amazon RDS automatically patches the database software and backs up your database, storing the backups for a user-defined retention period and enabling point-in-time recovery. You benefit from the flexibility of being able to scale the compute resources or storage capacity associated with your Database Instance (DB Instance) via a single API call.

Amazon RDS DB Instances can be provisioned with either standard storage or Provisioned IOPS storage. Amazon RDS Provisioned IOPS is a storage option designed to deliver fast, predictable, and consistent I/O performance, and is optimized for I/O- intensive, transactional (OLTP) database workloads. In addition, Amazon RDS makes it easy to use replication to enhance availability and reliability for production workloads. Using the Multi-AZ deployment option you can run mission critical workloads with high availability and built-in automated fail-over from your primary database to a synchronously replicated secondary database in case of a failure. Amazon RDS for MySQL also enables you to scale out beyond the capacity of a single database deployment for read-heavy database workloads. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Service Highlights

Simple to Deploy Database Web Service- Amazon RDS makes it easy to go from project conception to deployment. Use the AWS Management Console or simple API calls to access the capabilities of a production-ready relational database in minutes without worrying about infrastructure provisioning or installing and maintaining database software.

Managed-Amazon RDS handles time-consuming database management tasks, such as backups, patch management, and replication, allowing you to pursue higher value application development or database refinements.

Compatible-With Amazon RDS, you get native access to a relational database management system. This facilitates compatibility with your existing tools and applications. In addition, Amazon RDS gives you optional control over which supported engine version powers your DB Instance via DB Engine Version Management.

Fast, Predictable Performance- Database Instances using Amazon RDS's MySQL, Oracle, SQL Server, and Oracle engines can be provisioned with General Purpose (SSD) Storage, Provisioned IOPS (SSD) Storage, or Magnetic Storage.

Amazon RDS General Purpose (SSD) Storage delivers a consistent baseline of 3 IOPS per provisioned GB and provides the ability to burst up to 3,000 IOPS. Amazon RDS Provisioned IOPS (SSD) Storage is a high-performance storage option designed to deliver fast, predictable, and consistent performance for I/O intensive transactional database workloads. You can provision from 1,000 IOPS to 30,000 IOPS per DB Instance. (Maximum realized IOPS will vary by engine type.) Magnetic Storage (formerly known as Amazon RDS Standard storage) is useful for small database workloads where data is accessed less frequently.

Database Instances using the Amazon Aurora engine employ a fault-tolerant, self-healing SSD-backed virtualized storage layer purpose-built for database workloads.

Scalable Database in the Cloud- You can scale the compute and storage resources available to your database to meet your application needs using the Amazon RDS API or the AWS Management Console. If you are using Amazon RDS Provisioned IOPS storage with Amazon RDS for MySQL, Oracle, or PostgreSQL, you can provision and scale the storage up to 3TB and IOPS to up to 30,000. Note that maximum realized IOPS will vary by engine type. In addition, for the MySQL, PostgreSQL, and Amazon Aurora database engines, you can also associate one or more Read Replicas with your database instance deployment, enabling you to scale beyond the capacity of a single database instance for read-heavy workloads.

The Amazon Aurora database engine allows you to scale your storage up to 64TB. You can associate up to 15 Amazon Aurora Replicas with your database instance deployment. Amazon Aurora Replicas share the same underlying storage as the source instance, lowering costs and avoiding the need to copy data to the replica nodes.

Reliable-Amazon RDS has multiple features that enhance reliability for critical production databases, including automated backups, DB snapshots, automatic host replacement, and Multi-AZ deployments. Amazon RDS runs on the same highly reliable infrastructure used by other Amazon Web Services.

For the Amazon Aurora engine, Amazon RDS uses RDS Multi-AZ technology to automate failover to one of up to 15 Aurora Replicas you have created in any of three Availability Zones.

Designed for use with other Amazon Web Services- Amazon RDS is tightly integrated with other Amazon Web Services. For example, an application running in Amazon EC2 will experience low-latency database access to an Amazon RDS DB Instance in the same region.

Secure- Amazon RDS provides a number of mechanisms to secure your DB Instances. Amazon RDS allows you to encrypt your databases using keys you manage through AWS Key Management Service (KMS). On a database instance running with Amazon RDS encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots.

Amazon RDS supports Transparent Data Encryption in SQL Server and Oracle. Transparent Data Encryption in Oracle is integrated with AWS CloudHSM, which allows you to securely generate, store, and manage your cryptographic keys in single-tenant Hardware Security Module (HSM) appliances within the AWS cloud.

Amazon RDS includes web service interfaces to configure firewall settings that control network access to your database. Amazon RDS allows you to run your DB Instances in Amazon Virtual Private Cloud (Amazon VPC). Amazon VPC enables you to isolate your DB Instances by specifying the IP range you wish to use, and connect to your existing IT infrastructure through industry-standard encrypted IPsec VPN.

Inexpensive- You pay very low rates and only for the resources you actually consume. In addition, you benefit from the option of On-Demand pricing with no up-front or long-term commitments, or even lower hourly rates via our reserved pricing option.

- ✓ On-Demand DB Instances let you pay for compute capacity by the hour with no long-term commitments. This frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs.

- ✓ Reserved DB Instances give you the option to reserve capacity within a data center and in turn receive a significant discount on the hourly charge for instances that are covered by the reservation. You can choose between are three RI payment options -- No Upfront, Partial Upfront, All Upfront -- which enable you to balance the amount you pay upfront with your effective hourly price and receive a significant discount over On-Demand prices.

5.5.2 Amazon Aurora

Amazon Aurora is a MySQL-compatible, relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost effectiveness of open source databases. Amazon Aurora provides up to five times better performance than MySQL at a price point one tenth that of a commercial database while delivering similar performance and availability. Amazon Aurora joins MySQL, Oracle, Microsoft SQL Server, and PostgreSQL as the fifth database engine available to customers through Amazon RDS. Amazon RDS handles routine database tasks such as provisioning, patching, backup, recovery, failure detection, and repair.

5.5.3 Amazon DynamoDB

Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale. It is a fully managed database and supports both document and key-value data models. Its flexible data model and reliable performance make it a great fit for mobile, web, gaming, ad-tech, the Internet of things (IoT), and many other applications.

Benefits

Fast, Consistent Performance- Amazon DynamoDB is designed to deliver consistent, fast performance at any scale for all applications. Average service-side latencies are typically single-digit milliseconds. As our data volumes grow and application performance demands increase, Amazon DynamoDBmatic partitioning and solid state drive (SSD) technologies to meet your throughput requirements and deliver low latencies at any scale.

Highly Scalable- When creating a table, simply specify how much request capacity you require. If your throughput requirements change, simply update your table's request capacity using the AWS Management Console or the Amazon DynamoDB API. Amazon DynamoDB manages all the scaling behind the scenes, and you are still able to achieve your prior throughput levels while scaling is underway.

Flexible- Amazon DynamoDB supports both document and key-value data structures, giving you the flexibility to design the best architecture that is optimal for your application.

Fine-grained Access Control- Amazon DynamoDB integrates with AWS Identity and Access Management (IAM) for fine-grained access control for users within your organization. You can assign unique security credentials to each user and control each user's access to services and resources.

Fully Managed- Amazon DynamoDB is a fully managed cloud NoSQL database service—you simply create a database table, set your throughput, and let the service handle the rest. You no longer need to worry about database management tasks such as hardware or software provisioning, setup and configuration, software patching, operating a reliable, distributed database cluster, or partitioning data over multiple instances as you scale.

5.5.4 Amazon Redshift

Amazon Redshift is a fast, fully managed, petabytes-scale data warehouse solution that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools. You can start small with no commitments or upfront costs and scale to a petabyte or more. Amazon Redshift delivers fast query performance by using columnar storage technology to improve I/O efficiency and parallelizing queries across multiple nodes. Amazon Redshift uses standard PostgreSQL JDBC and ODBC drivers, allowing you to use a wide range of familiar SQL clients. Data load speed scales linearly with cluster size, with integrations to Amazon S3, Amazon DynamoDB, Amazon Elastic MapReduce, Amazon Kinesis, or any SSH-enabled host. We've automated most of the common administrative tasks associated with provisioning, configuring and monitoring a data warehouse. Backups to Amazon S3 are continuous, incremental, and automatic. Restores are fast; you can start querying in minutes while your data is spooled down in the background. Enabling disaster recovery across regions takes just a few clicks. Security is built-in. You can encrypt data at rest and in transit using hardware accelerated AES-256 and SSL, isolate your clusters using Amazon VPC, and even manage your keys using hardware security modules (HSMs). All API calls, connection attempts, queries, and changes to the cluster are logged and auditable.

Features and Benefits of Amazon Redshift

Amazon Redshift delivers fast query performance by using columnar storage technology to improve I/O efficiency and parallelizing queries across multiple nodes. Amazon Redshift has custom JDBC and ODBC drivers that you can download from the Connect Client tab of our Console, allowing you to use a wide range of familiar SQL clients. You can also use standard PostgreSQL JDBC and ODBC drivers. Data load speed scales linearly with cluster size, with integrations to Amazon S3, Amazon DynamoDB, Amazon Elastic MapReduce, Amazon Kinesis or any SSH-enabled host.

- ✓ Amazon Redshift's data warehouse architecture allows you to automate most of the common administrative tasks associated with provisioning, configuring and monitoring a cloud data warehouse. Backups to Amazon S3 are continuous, incremental and automatic. Restores are fast; you can start querying in minutes while your data is spooled down in the background. Enabling disaster recovery across regions takes just a few clicks.
- ✓ Security is built-in. You can encrypt data at rest and in transit using hardware-accelerated AES-256 and SSL, isolate your clusters using Amazon VPC and even manage your keys using AWS Key Management Service (KMS) and hardware security modules (HSMs). All API calls, connection attempts, queries and changes to the cluster are logged and auditable. You can use AWS CloudTrail to audit Redshift API calls.

Optimized for Data Warehousing

Amazon Redshift uses a variety of innovations to obtain very high query performance on datasets ranging in size from a hundred gigabytes to a petabyte or more. It uses columnar storage, data compression, and zone maps to reduce the amount of I/O needed to perform queries. Amazon Redshift has a massively parallel processing (MPP) data warehouse architecture, parallelizing and distributing SQL operations to take advantage of all available resources. The underlying hardware is designed for high performance data processing, using local attached storage to maximize throughput between the CPUs and drives, and a 10GigE mesh network to maximize throughput between nodes.

Scalable-With a few clicks of the AWS Management Console or a simple API call, you can easily change the number or type of nodes in your cloud data warehouse as your performance or capacity needs change. Dense Storage (DS) nodes allow you to create very large data warehouses using hard disk drives (HDDs) for a very low price point. Dense Compute (DC) nodes allow you to create very high performance data warehouses using fast CPUs, large amounts of RAM and solid-state disks (SSDs). Amazon Redshift enables you to start with as little as a single 160GB DC1.Large node and scale up all the way to a petabyte or more of compressed user data using 16TB DS2.8XLarge nodes. While resizing, Amazon Redshift places your existing cluster into read-only mode, provisions a new cluster of your chosen size, and then copies data from your old cluster to your new one in parallel. You can continue running queries against your old cluster while the new one is being provisioned. Once your data has been copied to your new cluster, Amazon Redshift will automatically redirect queries to your new cluster and remove the old cluster.

Cheap

No Up-Front Costs- You pay only for the resources you provision. You can choose On-Demand pricing with no up-front costs or long-term commitments, or obtain significantly discounted rates with Reserved Instance pricing. On-Demand pricing starts at just $0.25/hour per 160GB DC1.Large node or $0.85/hour per 2TB DS2.XLarge node. With Partial Upfront Reserved Instances, you can lower your effective price to $0.10/hour per DC1.Large node ($5,500/TB/year) or $0.228/hour per DS2.XLarge node ($999/TB/year). To see more details, visit the Amazon Redshift Pricing page.

Simple

Get Started in Minutes- With a few clicks in the AWS Management Console or simple API calls, you can create a cluster, specifying its size, underlying node type, and security profile. Amazon Redshift will provision your nodes, configure the connections between them, and secure the cluster. Your data warehouse should be up and running in minutes.

Fully Managed-Amazon Redshift handles all the work needed to manage, monitor, and scale your data warehouse, from monitoring cluster health and taking backups to applying patches and upgrades. You can easily resize your cluster as your performance and capacity needs change. By handling all these time-consuming, labour-intensive tasks, Amazon Redshift frees you up to focus on your data and business.

Fault Tolerant-Amazon Redshift has multiple features that enhance the reliability of your data warehouse cluster. All data written to a node in your cluster is automatically replicated to other nodes within the cluster and all data is continuously backed up to Amazon S3. Amazon Redshift continuously monitors the health of the cluster and automatically re-replicates data from failed drives and replaces nodes as necessary.

Automated Backups-Amazon Redshift's automated snapshot feature continuously backs up new data on the cluster to Amazon S3. Snapshots are continuous, incremental and automatic. Amazon Redshift stores your snapshots for a user-defined period, which can be from one to thirty-five days. You can take your own snapshots at any time, which leverage all existing system snapshots and are retained until you explicitly delete them. Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery. Once you delete a cluster, your system snapshots are removed but your user snapshots are available until you explicitly delete them.

Fast Restores- You can use any system or user snapshot to restore your cluster using the AWS Management Console or the Amazon Redshift APIs. Your cluster is available as soon as the system metadata has been restored and you can start running queries while user data is spooled down in the background.

Secure Encryption- With just a couple of parameter settings, you can set up Amazon Redshift to use SSL to secure data in transit and hardware-accelerated AES-256 encryption for data at rest. If you choose to enable encryption of data at rest, all data written to disk will be encrypted as well as any backups. By default, Amazon Redshift takes care of key management but you can choose to manage your keys using your own hardware security modules (HSMs), AWS CloudHSM, or AWS Key Management Service.

Network Isolation- Amazon Redshift enables you to configure firewall rules to control network access to your data warehouse cluster. You can run Amazon Redshift inside Amazon Virtual Private Cloud (Amazon VPC) to isolate your data warehouse cluster in your own virtual network and connect it to your existing IT infrastructure using industry-standard encrypted IPsec VPN.

Audit and Compliance- Amazon Redshift integrates with AWS CloudTrail to enable you to audit all Redshift API calls. Amazon Redshift also logs all SQL operations, including connection attempts, queries and changes to your database. You can access these logs using SQL queries against system tables or choose to have them downloaded to a secure location on Amazon S3. Amazon Redshift is compliant with SOC1, SOC2, SOC3 and PCI DSS Level 1 requirements.

Compatible SQL- Amazon Redshift is a SQL data warehouse solution and uses industry standard ODBC and JDBC connections. You can download our custom JDBC and ODBC drivers from the Connect Client tab of our Console. Many popular software vendors have certified Amazon Redshift with their offerings to enable you to continue to use the tools you do today.

Integrated- Amazon Redshift is integrated with other AWS services and has built in commands to load data in parallel to each node from Amazon S3, Amazon DynamoDB or your EC2 and on-premise servers using SSH. AWS Data Pipeline, Amazon Kinesis, and AWS Lambda integrate with Amazon Redshift as a data target.

5.5.5 Google Datastore

Google Datastore (App Engine Datastore) is the main data storage service for Google App Engine applications. It's a NoSQL database system, built on top of Google's own Big table database structure. As a NoSQL database system, Google Datastore is a schema-less database. It stores data in data objects known as entities. Each entity is categorized into some categories known as its kind (for query purpose), and it keeps a key(which is not mutable) to identifies itself from other entities of the same kind. Each entity has one or more properties, which is a named value of some supported data types.

Google Datastore offers two data storage options: High Replication Datastore (HRD), which makes use of Paxos architecture to enhance reliability and availability, and Master/Slave Datastore, which makes use of Master-slave architecture to ensure strong consistency for

database operations. As a Cloud database service, in particularly a NoSQL database system, Google Datastore uses a distributed architecture to help increase the scalability for the database system. It can scale easily to extremely large data sets, while still maintaining good performance.

5.5.6 Google Cloud SQL

Google Cloud SQL is a web services from Google that provides relational database service for Application deployed on Google App Engine. This is a new feature from Google App Engine and it's currently in limited preview phase. Google Cloud SQL supports MySQL database, with feature to import or export from existing MySQL database into and out of the cloud. As Google Cloud SQL is designed to ensure reliability and availability, it supports replication of data in different availability regions.

Currently, Google Cloud SQL only supports Java-based and Python-based application. To use it, developers need to use JDBC (Java Database connectivity) to connect to the database if their application is a Java-based application or DB-API if their application is a Python-based application.

Similar to Amazon Relational Database service, the underlying database system in Google Cloud SQL is fully managed by Google, so user can be saved from redundant and tiring tasks such as patch management for the database. On top of it, a rich GUI is provided to help user to managing, monitoring and configuring their database system easily.

5.6 Considerations when choosing a cloud-based database

With different architectures of cloud-based database, there are a number of considerations users should consider when choosing to use a cloud-based database system:

5.6.1 Portability

Moving to a Cloud-based database system means the user needs to transfer their existing data from their current database to the cloud. Especially with organizations who currently use traditional relational database and have lots of existing data, portability is really needed. For these organization, choosing some relational database systems on the cloud, such as Amazon Relational Database services (which support Oracle and MySQL database, with import, export feature), or Google Cloud SQL (which is currently in limited preview phase, which also support import, export existing data) is a sensible solution.

In addition, the migration possibility of database from one cloud-computing provider to another, or even from a cloud-computing provider to your own server, matters. There might be unexpected circumstances that occur, forcing the user to drop the current cloud-computing provider and moving to another one. Therefore, before actually settling on a particular

database from a particular cloud-computing provider, the user needs to consider if they can easily port their application and its database code after they have implemented it.

5.6.2 Reliability and Availability

For database that requires high reliability and availability, a cloud-based database that offer replication of data is really important.

For example, Amazon Relation Database Services offer a feature that could help to ensure reliability for the data: Multi AZ deployment. When user enable this and run their instance as a Multi AZ instance, Amazon RDS will automatically create and manage a "standby" replica in a different Availability Zone. Database updates are performed to both primary and standby database at the same time. The standby database cannot be used to serve read traffic, but it can be used to replace the primary one in case of database maintenance or database instance failure. It helps to ensure reliability and availability for the database system in case of any incident.

Google Cloud SQL is also designed to cater for database with high replication application, because it is designed with inherent support for replication of data in different availability regions. Google Datastore (a NoSQL db system) offers the model high-replication Datastore (HRD), using Paxos architecture to increase reliability and availability for the database sytem. However, with database that does not require high replication of data, using database service with these feature could badly affects the performance of the application.

5.6.3 Scalability

Scalability is one of the main reasons why companies should consider using cloud-based database system, because most cloud-based database systems are designed to offer users with easier scalability than traditional database systems.

For users with some existing database systems, who just want to improve their database performance, and take advantage of a cloud solution, but at the same time requires complex transaction operations (such as join query) and complex relations among data in their database, the solution of moving their existing database system to a cloud service like Amazon Relational Database Service or Google Cloud SQL is a great option to consider.

However, for applications that really demands performance and scalability instead of complex database operation, or the data stored is not well-structured, any relational database system, due to the innate nature of relational database, will not perform as good as NoSQL database system for extremely large amount of data. Therefore, a NoSQL database solution on the Cloud such as DynamoDB (from Amazon), or Google Datastore (used for Google App Engine), is much more suitable. For Amazon DynamoDB, all the users need to do is specify the level of traffic they wish to serve, and Amazon will take care of all the works of scaling up the system to ensure the application could serve the desired traffic level.

5.6.4 Programming Environment

Aside from deciding whether to use a relational or non-relational database or what architecture the database should be built upon, the user should also be concerned with the programming environments that come with the database. This is because the programming environment contributes to the perceived speed of database operations from the client side and the migration possibility of your database.

Different databases can be accessed by only certain programming languages and their APIs. For example, when using Google App Engine's non-relational database (Google Datastore), user can only use Java, Python, or Go to access it, even though Google said that it plans to support more language in the future.

However, if using MySQL hosted on Amazon Web Services, the user is able to use a myriad of programming languages such as C#, Visual Basic and Java. The runtime of programs coded in these various languages differ, impacting the end user's experience, because the information interchange of client-server and server-database depend mostly on the programming environment.

INTERVIEW QUESTIONNAIRES

1. What are the benefits of fibre channel sans?

Fibre Channel SANs are the de facto standard for storage networking in the corporate data center because they provide exceptional reliability, scalability, consolidation, and performance. Fibre Channel SANs provide significant advantages over direct-attached storage through improved storage utilization, higher data availability, reduced management costs, and highly scalable capacity and performance.

2. What Environment is most suitable for fibre channel sans?

Typically, Fibre Channel SANs are most suitable for large data centers running business-critical data, as well as applications that require high-bandwidth performance such as medical imaging, streaming media, and large databases. Fibre Channel SAN solutions can easily scale to meet the most demanding performance and availability requirements.

3. What Customer problems do fibre channel sans solves?

The increased performance of Fibre Channel enables a highly effective backup and recovery approach, including LAN-free and server-free backup models. The result is a faster, more scalable, and more reliable backup and recovery solution. By providing flexible connectivity options and resource sharing, Fibre Channel SANs also greatly reduce the number of physical devices and disparate systems that must be purchased and managed, which can dramatically lower capital expenditures. Heterogeneous SAN management provides a single point of control for all devices on the SAN, lowering costs and freeing personnel to do other tasks.

4. How long has fibre channel been around?

Development started in 1988, ANSI standard approval occurred in 1994, and large deployments began in 1998. Fibre Channel is a mature, safe, and widely deployed solution for high-speed (1 GB, 2 GB, 4 GB) communications and is the foundation for the majority of SAN installations throughout the world.

5. What is the future of fibre channel sans?

Fibre Channel is a well-established, widely deployed technology with a proven track record and a very large installed base, particularly in high-performance, business-critical data center environments. Fibre Channel SANs continue to grow and will be enhanced for a long time to

come. The reduced costs of Fibre Channel components, the availability of SAN kits, and the next generation of Fibre Channel (4 GB) are helping to fuel that growth. In addition, the fibre channel roadmap includes plans to double performance every three years.

6. What are the benefits of 4 GB fibre channel?

Benefits include twice the performance with little or no price increase, investment protection with backward compatibility to 2 GB, higher reliability due to fewer SAN components (switch and HBA ports) required, and the ability to replicate, back up, and restore data more quickly. 4 GB Fibre Channel systems are ideally suited for applications that need to quickly transfer large amounts of data such as remote replication across a SAN, streaming video on demand, modeling and rendering, and large databases. 4 GB technology is shipping today.

7. How is fibre channel different from iSCSI?

Fibre Channel and iSCSI each have a distinct place in the IT infrastructure as SAN alternatives to DAS. Fibre Channel generally provides high performance and high availability for business-critical applications, usually in the corporate data center. In contrast, iSCSI is generally used to provide SANs for business applications in smaller regional or departmental data centers.

8. When should i deploy fibre channel instead of ISCSI?

For environments consisting of high-end servers that require high bandwidth or data center environments with business-critical data, Fibre Channel is a better fit than iSCSI. For environments consisting of many midrange or low-end servers, an IP SAN solution often delivers the most appropriate price/performance.

9. Name some of the SAN topologies?

Point-to-point, arbitrated loop, and switched fabric topologies

10.What's the need for separate network for storage why LAN cannot be used?

LAN hardware and operating systems are geared to user traffic, and LANs are tuned for a fast user response to messaging requests. With a SAN, the storage units can be secured separately from the servers and totally apart from the user network enhancing storage access in data blocks (bulk data transfers), advantageous for server-less backups.

11. What are the advantages of RAID?

"Redundant Array of Inexpensive Disks"
Depending on how we configure the array, we can have the
- data mirrored [RAID 1] (duplicate copies on separate drives)
- striped [RAID 0] (interleaved across several drives), or
- parity protected [RAID 5](extra data written to identify errors).
These can be used in combination to deliver the balance of performance and reliability that the user requires.

12. Define RAID? Which one you feel is good choice?

RAID (Redundant array of Independent Disks) is a technology to achieve redundancy with faster I/O. There are Many Levels of RAID to meet different needs of the customer which are: R0, R1, R3, R4, R5, R10, and R6. Generally customer chooses R5 to achieve better redundancy and speed and it is cost effective.

R0 – Striped set without parity/[Non-Redundant Array].

Provides improved performance and additional storage but no fault tolerance. Any disk failure destroys the array, which becomes more likely with more disks in the array. A single disk failure destroys the entire array because when data is written to a RAID 0 drive, the data is broken into fragments. The number of fragments is dictated by the number of disks in the drive. The fragments are written to their respective disks simultaneously on the same sector. This allows smaller sections of the entire chunk of data to be read off the drive in parallel, giving this type of arrangement huge bandwidth. RAID 0 does not implement error checking so any error is unrecoverable. More disks in the array mean higher bandwidth, but greater risk of data loss.

R1 - Mirrored set without parity.

Provides fault tolerance from disk errors and failure of all but one of the drives. Increased read performance occurs when using a multi-threaded operating system that supports split seeks, very small performance reduction when writing. Array continues to operate so long as at least one drive is functioning. Using RAID 1 with a separate controller for each disk is sometimes called duplexing.

R3 - Striped set with dedicated parity/Bit interleaved parity.

This mechanism provides an improved performance and fault tolerance similar to RAID 5, but with a dedicated parity disk rather than rotated parity stripes. The single parity disk is a bottle-neck for writing since every write requires updating the parity data. One minor benefit is the dedicated parity disk allows the parity drive to fail and operation will continue without parity or performance penalty.

R4 - Block level parity.

Identical to RAID 3, but does block-level striping instead of byte-level striping. In this setup, files can be distributed between multiple disks. Each disk operates independently which allows I/O requests to be performed in parallel, though data transfer speeds can suffer due to the type of parity.
The error detection is achieved through dedicated parity and is stored in a separate, single disk unit.

R5 - Striped set with distributed parity.

Distributed parity requires all drives but one to be present to operate; drive failure requires replacement, but the array is not destroyed by a single drive failure. Upon drive failure, any subsequent reads can be calculated from the distributed parity such that the drive failure is masked from the end user. The array will have data loss in the event of a second drive failure and is vulnerable until the data that was on the failed drive is rebuilt onto a replacement drive.

R6 - Striped set with dual distributed Parity.

Provides fault tolerance from two drive failures; array continues to operate with up to two failed drives. This makes larger RAID groups more practical, especially for high availability systems. This becomes increasingly important because large-capacity drives lengthen the time needed to recover from the failure of a single drive.
Single parity RAID levels are vulnerable to data loss until the failed drive is rebuilt: the larger the drive, the longer the rebuild will take. Dual parity gives time to rebuild the array without the data being at risk if one drive, but no more, fails before the rebuild is complete.

13. What are the difference between RAID 0+1 and RAID 1+0?

RAID 0+1 (Mirrored Stripped)

In this RAID level all the data is saved on stripped volumes which are in turn mirrored, so any disk failure saves the data loss but it makes whole stripe unavailable. The key difference from RAID 1+0 is that RAID 0+1 creates a second striped set to mirror a primary striped set. The array continues to operate with one or more drives failed in the same mirror set, but if drives fail on both sides of the mirror the data on the RAID system is lost. In this RAID level if one disk is failed full mirror is marked as inactive and data is saved only one stripped volume.

RAID 1+0 (Stripped Mirrored)

In this RAID level all the data is saved on mirrored volumes which are in turn stripped, so any disk failure saves data loss. The key difference from RAID 0+1 is that RAID 1+0 creates a striped set from a series of mirrored drives. In a failed disk situation RAID 1+0 performs better because all the remaining disks continue to be used.
The array can sustain multiple drive losses so long as no mirror loses both its drives. This RAID level is most preferred for high performance and high data protection because rebuilding of RAID 1+0 is less time consuming in comparison to RAID 0+1.

14. When JBOD's are used?

"Just a Bunch of Disks"

It is a collection of disks that share a common connection to the server, but don't include the mirroring, striping, or parity facilities that RAID systems do, but these capabilities are available with host-based software.

15. Differentiate RAID & JBOD?

RAID: "Redundant Array of Inexpensive Disks"

Fault-tolerant grouping of disks that server sees as a single disk volume
Combination of parity-checking, mirroring, striping
Self-contained, manageable unit of storage

JBOD: "Just a Bunch of Disks"
Drives independently attached to the I/O channel
Scalable, but requires server to manage multiple volumes
Do not provide protection in case of drive failure

16. What is a HBA?

Host bus adapters (HBAs) are needed to connect the server (host) to the storage.

17. What are the advantages of SAN?

Massively extended scalability
Greatly enhanced device connectivity
Storage consolidation
LAN-free backup
Server-less (active-fabric) backup
Server clustering
Heterogeneous data sharing
Disaster recovery - Remote mirroring
While answering people do NOT portray clearly what they mean & what advantages each of

them have, which are cost effective & which are to be used for the client's requirements.

18. What is the difference b/w SAN and NAS?

The basic difference between SAN and NAS, SAN is Fabric based and NAS is Ethernet based.
SAN - Storage Area Network
It accesses data on block level and produces space to host in form of disk.
NAS - Network attached Storage

It accesses data on file level and produces space to host in form of shared network folder.

19. What is a typical storage area network consists of - if we consider it for implementation in a small business setup?

If we consider any small business following are essentials components of SAN
- Fabric Switch
- FC Controllers
- JBOD's

20. Can you briefly explain each of these Storage area components?

Fabric Switch: It's a device which interconnects multiple network devices .There are switches starting from 16 port to 32 ports which connect 16 or 32 machine nodes etc. vendors who manufacture these kind of switches are Brocade, McData.

21. FC Controllers: These are Data transfer media they will sit on PCI slots of Server; you can configure Arrays and volumes on it.

JBOD: Just Bunch of Disks is Storage Box, it consists of Enclosure where set of hard-drives are hosted in many combinations such SCSI drives, SAS, FC, and SATA.

22. What is the most critical component in SAN?

Each component has its own criticality with respect to business needs of a company.

23. How is a SAN managed?

There are many management software's used for managing SAN's to name a few
- Santricity
- IBM Tivoli Storage Manager.
- CA Unicenter.
- Veritas Volume manger.

24. Which one is the Default ID for SCSI HBA?

Generally the default ID for SCSI HBA is 7.
SCSI- Small Computer System Interface
HBA - Host Bus Adaptor

25. What is the highest and lowest priority of SCSI?

There are 16 different ID's which can be assigned to SCSI device 7, 6, 5, 4, 3, 2, 1, 0, 15, 14, 13, 12, 11, 10, 9, 8.
Highest priority of SCSI is ID 7 and lowest ID is 8.

26. How do you install device drivers for the HBA first time during OS installation?

In some scenarios you are supposed to install Operating System on the drives connected thru SCSI HBA or SCSI RAID Controllers, but most of the OS will not be updated with drivers for those controllers, that time you need to supply drivers externally, if you are installing windows, you need to press F6 during the installation of OS and provide the driver disk or CD which came along with HBA.

If you are installing Linux you need to type "linux dd" for installing any driver.

27. What is Array?

Array is a group of Independent physical disks to configure any Volumes or RAID volumes.

28. Can you describe at-least 3 troubleshooting scenarios which you have come across in detail?

SCENARIO 1: How do you find/debug when there is error while working SCSI devices?

In our daily SAN troubleshooting there are many management and configuration tools we use them to see when there is a failure with target device or initiator device.
Some time it is even hard to troubleshoot some of the things such as media errors in the drives, or some of the drives taking long time to spin-up.
In such cases these utilities will not come to help. To debug this kind of information most of the controller will be implemented with 3-pin serial debug port. With serial port debug connector cable you can collect the debug information with hyper terminal software.

SCENARIO 2: I am having an issue with a controller its taking lot of time to boot and detect all the drives connected how can I solve this?

There are many possibilities that might cause this problem. One of the reason might be you are using bad drives that cannot be repaired. In those cases you replace the disks with working ones. Another reason might be slots you connected your controller to a slot which might not be supported. Try to connect with other types of slots. One more probable reason is if you have flashed the firmware for different OEM's on the same hardware. To get rid of this the flash utilities will be having option to erase all the previous and EEPROM and boot block entry option. Use that option to rectify the problem.

SCENARIO 3: I am using tape drive series 700X, even the vendor information on the Tape drive says 700X, but the POST information while booting the server is showing as 500X what could be the problem?

First you should make sure your hardware is of which series, you can find out this in the product website. Generally you can see this because in most of the testing companies they use same hardware to test different series of same hardware type. What they do is they flash the different series firmware. You can always flash back to exact hardware type.

29. Which are the SAN topologies?

SAN can be connected in 3 types which are mentioned below:
Point to Point topology
FC Arbitrated Loop (FC :Fibre Channel)
Switched Fabric

30. Which are the 4 types of SAN architecture types

a. Core-edge
b. Full-Mesh
c. Partial-Mesh
d. Cascade

31.Which command is used in linux to know the driver version of any hardware device?

dmesg

32. How many minimum drives are required to create R5 (RAID 5)?

You need to have at least 3 disk drives to create R5.

33. Can you name some of the states of RAID array?

There are states of RAID arrays that represent the status of the RAID arrays which are given

below
a. Online
b. Degraded
c. Rebuilding
d. Failed

34. Name the features of SCSI-3 standard?

QAS: Quick arbitration and selection
Domain Validation
CRC: Cyclic redundancy check

35. Can we assign a hot spare to R0 (RAID 0) array?

No, since R0 is not redundant array, failure of any disks results in failure of the entire array so we cannot rebuild the hot spare for the R0 array.

36. Can you name some of the available tape media types?

There are many types of tape media available to back up the data some of them are DLT: digital linear tape - technology for tape backup/archive of networks and servers; DLT technology addresses midrange to high-end tape backup requirements. LTO: linear tape open; a new standard tape format developed by HP, IBM, and Seagate. AIT: advanced intelligent tape; a helical scan technology developed by Sony for tape backup/archive of networks and servers, specifically addressing midrange to high-end backup requirements.

37. What is HA?

HA High Availability is a technology to achieve failover with very less latency. It's a practical requirement of data centers these days when customers expect the servers to be running 24 hours on all 7 days around the whole 365 days a year - usually referred as 24x7x365. So to achieve this, a redundant infrastructure is created to make sure if one database server or if one app server fails there is a replica Database or Appserver ready to take-over the operations. End customer never experiences any outage when there is a HA network infrastructure.

38. What is virtualization?

Virtualization is logical representation of physical devices. It is the technique of managing and presenting storage devices and resources functionally, regardless of their physical layout or location. Virtualization is the pooling of physical storage from multiple network storage devices into what appears to be a single storage device that is managed from a central console. Storage virtualization is commonly used in a storage area network (SAN). The

management of storage devices can be tedious and time-consuming. Storage virtualization helps the storage administrator perform the tasks of backup, archiving, and recovery more easily, and in less time, by disguising the actual complexity of the SAN.

39. Describe in brief the composition of FC Frame?

Start of the Frame locator
Frame header (includes destination id and source id, 24 bytes/6 words)
Data Payload (encapsulate SCSI instruction can be 0-2112 bytes in length)
CRC (error checking, 4 bytes)
End of Frame (1 byte)

40. What is storage virtualization?

Storage virtualization is amalgamation of multiple n/w storage devices into single storage unit.

41. What are the protocols used in physical/datalink and network layer of SAN?

a) Ethernet
b) SCSI
c) Fibre Channel

42. What are the types of disk array used in SAN?

a) JBOD
b) RAID

43.What are different types of protocols used in transportation and session layers of SAN?

a) Fibre Channel Protocol (FCP)
b) Internet SCSI (iSCSI)
c) Fibre Channel IP (FCIP)

44.What is the type of Encoding used in Fibre Channel?

8b/10b, as the encoding technique is able to detect all most all the bit errors

45.How many classes of service are available in Fibre Channel?

7 Classes of service are available in Fibre Channel

Class-1: Dedicated connection between two communicators with acknowledgement of frame

delivery.
In class 1 service, a dedicated connection source and destination is established through the fabric for the duration of the transmission. It provides acknowledged service. This class of service ensures that the frames are received by the destination device in the same order in which they are sent, and reserves full bandwidth for the connection between the two devices. It does not provide for a good utilization of the available bandwidth, since it is blocking another possible contender for the same device. Because of this blocking and necessary dedicated connection, class 1 is rarely used.

Class-2: connection less but provides acknowledgement
Class 2 is a connectionless, acknowledged service. Class 2 makes better use of available bandwidth since it allows the fabric to multiplex several messages on a frame-by-frame basis. As frames travel through the fabric they can take different routes, so class 2 service does not guarantee in-order delivery. Class 2 relies on upper layer protocols to take care of frame sequence. The use of acknowledgments reduces available bandwidth, which needs to be considered in large-scale busy networks.

Class-3: connection less and provides no notification of delivery
There is no dedicated connection in class 3 and the received frames are not acknowledged. Class 3 is also called datagram connectionless service. It optimizes the use of fabric resources, but it is now upper layer protocol to ensure that all frames are received in the proper order, and to request to the source device the retransmission of missing frames. Class 3 is a commonly used class of service in Fibre Channel networks.

Class-4: allows fractional bandwidth for virtual circuits
Class 4 is a connection-oriented service like class 1, but the main difference is that it allocates only a fraction of available bandwidth of path through the fabric that connects two N_Ports. Virtual Circuits (VCs) are established between two N_Ports with guaranteed Quality of Service (QoS), including bandwidth and latency. Like class 1, class 4 guarantees in-order delivery frame delivery and provides acknowledgment of delivered frames, but now the fabric is responsible for multiplexing frames of different VCs. Class 4 service is mainly intended for multimedia applications such as video and for applications that allocate an established bandwidth by department within the enterprise. Class 4 was added in the FC-PH-2 standard.

Class -5: Class 5 is called isochronous service, and it is intended for applications that require immediate delivery of the data as it arrives, with no buffering. It is not clearly defined yet. It is not included in the FC-PH documents.

Class-6: Provides multicast, dedicated connection with acknowledgment
Class 6 is a variant of class 1, known as multicast class of service. It provides dedicated connections for a reliable multicast. An N_Port may request a class 6 connection for one or more destinations. A multicast server in the fabric will establish the connections and get acknowledgment from the destination ports, and send it back to the originator. Once a connection is established, it should be retained and guaranteed by the fabric until the initiator

ends the connection. Class 6 was designed for applications like audio and video requiring multicast functionality. It appears in the FC-PH-3 standard.

Class-F: used for switch to switch communication in the fabric.
Class F service is defined in the FC-SW and FC-SW-2 standard for use by switches communicating through ISLs. It is a connectionless service with notification of non-delivery between E_Ports used for control, coordination, and configuration of the fabric. Class F is similar to class 2; the main difference is that Class 2 deals with N_Ports sending data frames, while Class F is used by E_ports for control and management of the fabric.

46.What are the main constrains of SCSI in storage networking?

a)Deployment distance (max. of 25 mts)
b)Number of devices that can be interconnected (16)

47.What is a Fabric?

Interconnection of Fibre Channel Switches

48.What are the services provided by Fabric to all the nodes?

a)Fabric Login
b)SNS
c)Fabric Address Notification
d)Registered state change notification
e)Broadcast Servers

49.What is the difference between LUN and WWN?

LUN: unique number that is assigned to each storage device or partition of the storage that the storage can support.
WWN: 64bit address that is hard coded into a fibre channel HBA and this is used to identify individual port (N_Port or F_Port) in the fabric.

50.What are the different topologies in Fibre Channel?

a)Point-to-Point
b)Arbitrary Loop
c)Switched Fabric Loop

51.What are the layers of Fibre Channel Protocol?

a)FC Physical Media
b)FC Encoder and Decoder

c)FC Framing and Flow control
d)FC Common Services
e)FC Upper Level Protocol Mapping

52.What is zoning?

Fabric management service that can be used to create logical subsets of devices within a SAN. This enables portioning of resources for management and access control purpose.

53.What are the two major classification of zoning?

Two types of zoning are
a) Software Zoning
b) Hardware Zoning

54.What are different levels of zoning?

a)Port Level zoning
b)WWN Level zoning
c)Device Level zoning
d)Protocol Level zoning
e)LUN Level zoning

55.What are the 3 prominent characteristics of SAS Protocol?

a)Native Command Queuing (NCQ)
b)Port Multiplier
c)Port Selector

56.What are the 5 states of Arbitrary Loop in FC?

a)Loop Initialization
b)Loop Monitoring
c)Loop arbitration
d)Open Loop
e)Close Loop

57.How does FC Switch maintain the addresses?

FC Switch uses simple name server (SNS) to maintain the mapping table

58.What is the purpose of disk array?

Probability of unavailability of data stored on the disk array due to single point failure is totally eliminated.

59.What is disk array?

Set of high performance storage disks that can store several terabytes of data. Single disk array can support multiple points of connection to the network.

60.What is virtualization?

A technique of hiding the physical characteristics of computer resources from the way in which other system application or end user interact with those resources. Aggregation, spanning or concatenation of the combined multiple resources into larger resource pools.

61.What is Multipath I/O?

Fault tolerant technique where, there is more than one physical path between the CPU in the computer systems and its main storage devices through the buses, controllers, switches and other bridge devices connecting them.

62.What is RAID?

Technology that groups several physical drives in a computer into an array that you can define as one or more logical drive. Each logical drive appears to the operating system as single drive. This grouping enhances the performance of the logical drive beyond the physical capability of the drives.

63.What is stripe-unit-size?

It is data distribution scheme that complement s the way operating system request data. Granularity at which data is stored on one drive of the array before subsequent data is stored on the next drive of the array. Stripe unit size should be close to the size of the system I/O request.

64.What is LUN Masking?

A method used to create an exclusive storage area and access control. And this can be achieved by storage device control program.

65.What is the smallest unit of information transfer in FC?

Frame

66.How is the capacity of the HDD calculated?

Number of Heads X Number of Cylinders X Sectors per track X Sector Size

67.What is bad block reallocation?

A bad sector is remapped or reallocated to good spare block and this information is stored in the internal table on the hard disk drive. The bad blocks are identified during the media test of the HDD as well as during various types of read write operations performed during the I/O tests. Apart from the new generation of HDD comes with a technology called BGMS (background media scan) which continuously scans the HDD media for defects and maps them when the drive is idle (this is performed after the HDD is attached to the system).

68.What are two types of recording techniques on the tapes?

a)Linear Recording
b)Helical Scan Recording.

69.What is snapshot?

A snapshot of data object contains an image of data at a particular point of time.

70.What is HSM?

Hierarchical storage management - An application that attempts to match the priority of data with the cost of storage.

71. What is hot-swapping?

Devices are allowed to be removed and inserted into a system without turning off the system.

72. What is Hot-Sparing?

A spare device is available to be inserted into the subsystem operation without having to remove and replace a device.

73. What are different types of backup system?

a) Offline
b) Online
c) Near Line

74. What is the different between mirroring, Routing and multipathing?

Redundancy Functions Relationships Role
Mirroring Generates 2 ios to 2 storage targets Creates 2 copies of data
Routing Determined by switches independent of SCSI Recreates n/w route after a failure
Multipathing Two initiator to one target Selects the LUN initiator pair to use

75.Name few types of Tape storage?

a) Digital Linear Tape
b) Advanced Intelligent Tape
c) Linear Tape Open

76.What is a sequence in FC?

Group of one or more frames that encompasses one or more “information units” of a upper layer protocol.

Example:
It requires
i) One sequence to transfer the command
ii) One or more sequence to transfer the data
iii) Once sequence to transfer the status.

77. What is Exchange in FC?

Exchange is to establish a relationship between 2 N_PORTs and then these two ports transfer data via one or more sequence within this relationship.
Example: Exchange exist to transfer the command, data and the status of one SCSI task

78. Why do we need Login in FC?

Port Login: To exchange service parameters between N_Ports and N_Ports
Process Login: To establish the SCSI operating environment between two N_PORTS
Fabric Login: Similar to port login, FLOGI is an extended link service command that sets up a session between two participants. With FLOGU a session is created between an N_Ports or N_Ports and the switch.

79. What are the different types of clusters?

a) High availability clusters
b) High Performance Clusters
c) Load Balancing Clusters.

80. What are three levels of management in storage?

a) Storage Level Management
b) Network Level Management
c) Enterprise Level Management

81. What are the key activities in SAN management?

a) Monitoring
b) Configuring
c) Controlling
d) Troubleshooting
e) Diagnosing

82. What is the difference between HBA and NIC?

HBA => Host bus adapters are used in storage based traffic while NIC (Network Interface Cards are used in IP based LAN traffic.

83. What is the measuring unit of data activity?

Gigabits per second (Gb/ps)

84. What are the basic storage policies?

a) Security and authentication
b) Capacity, Content and quota management
c) Quality of Service

85. What is bypass circuitry?

A circuit that automatically removes the storage device from the data path (FC device out of FC AL loop) when signaling is lost (this signal is called port by-pass signal).

86. How many connections are possible in Fabric topology?

2^{24} (24 bit address to the port), and the largest possible fabric will have 239 interconnected switches.

87. What is one of the constrain of using storage switch?

Latency

88. What is the difference between NAS and SAN?

NAS
Cables used in the n/w
n/w protocols (TCP/IP, IPx) and file sharing protocols (CIFS & NFS)
Lower TCO
Support heterogeneous clients
Slow

SAN
High-speed connectivity such as FC
Do not use n/w protocols because data request are not made over LAN
Higher TCO
Requires special s/w to provide access to heterogeneous clients
Fast

89. What is Jitter?

Jitter refers to any deviation in timing that a bit stream suffers as it traverses the physical medium and the circuitry on-board the end devices. A certain amount of deviation from the original signaling will occur naturally as serial bit stream propagates over fibre-optic or copper cabling.
Mainly caused by electro-magnetic interference

90. What is BER/Bit error rate?

Probability that a transmitted bit will be erroneously received is the measure of number of bits (erroneous) at the output of the receiver and dividing by the total number of bits in transmission.

91. What is WWPN?

WWPN is the 16bit character that is assigned to the port, SAN volume controller uses it to uniquely identify the fibre channel HBA that is installed in the host system.

92. What is connection allegiance?

Given multiple connections are established, individual command/response pair must flow over the same connection. This connection allegiance ensures that specific read or writes commands are fulfilled without any additional overhead of monitoring multiple connections and to see whether a particular request is completed.

93. What is burst Length?

The burst length is the number of bytes that the SCSI initiator sends to the SCSI target in the FCP_DATA sequence.

94. What is NAS in detail?

NAS or Network Attached Storage

“NAS is used to refer to storage elements that connect to a network and provide file access services to computer systems. A NAS Storage Element consists of an interface or

engine, which implements the file services, and one or more devices, on which data is stored. NAS elements may be attached to any type of network. When attached to SANs, NAS elements may be considered to be members of the SAS (SAN Attached Storage) class of storage elements.

A class of systems that provide file services to host computers. A host system that uses network attached storage uses a file system device driver to access data using file access protocols such as NFS or CIFS. NAS systems interpret these commands and perform the internal file and device I/O operations necessary to execute them.

Though the NAS does speed up bulk transfers, it does not offload the LAN like a SAN does. Most storage devices cannot just plug into gigabit Ethernet and be shared - this requires a specialized file server the variety of supported devices is more limited. NAS has various protocols established for such needed features as discovery, access control, and name services.

95. Briefly list the advantages of SAN?

SANs fully exploit high-performance, high connectivity network technologies
SANs expand easily to keep pace with fast growing storage needs
SANs allow any server to access any data
SANs help centralize management of storage resources
SANs reduce total cost of ownership (TCO).

iSCSI fundamentals

iSCSI is a protocol defined by the Internet Engineering Task Force (IETF) which enables SCSI commands to be encapsulated in TCP/IP traffic, thus allowing access to remote storage over low cost IP networks.

96. What advantages would using an iSCSI Storage Area Network (SAN) give to your organization over using Direct Attached Storage (DAS) or a Fibre Channel SAN?

iSCSI is cost effective, allowing use of low cost Ethernet rather than expensive Fibre architecture.
Traditionally expensive SCSI controllers and SCSI disks no longer need to be used in each server, reducing overall cost.
Many iSCSI arrays enable the use of cheaper SATA disks without losing hardware RAID functionality.
The iSCSI storage protocol is endorsed by Microsoft, IBM and Cisco, therefore it is an industry standard.
Administrative/Maintenance costs are reduced.
Increased utilisation of storage resources.
Expansion of storage space without downtime.

Easy server upgrades without the need for data migration.
Improved data backup/redundancy.

96. What is Amazon EC2 service?

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable (scalable) computing capacity in the cloud. You can use Amazon EC2 to launch as many virtual servers you need. In Amazon EC2 you can configure security and networking, and manage storage.

97. What are the features of Amazon EC2 service?

As Amazon EC2 service is a cloud service so it has all the cloud features. Amazon EC2 provides the following features:
· Virtual computing environment (known as instances)
· Pre-configured templates for your instances (known as Amazon Machine Images – AMIs)
· Amazon Machine Images (AMIs) is package that you need for your server (including the operating system and additional software)
· Amazon EC2 provides various configuration of CPU, memory, storage and networking capacity for your instances (known as instance type)
· Secure login information for your instances using key pairs (AWS stores the public key and you store the private key in a secure place)
· Storage volumes for temporary data that's deleted when you stop or terminate your instance (known as instance store volumes)
· Amazon EC2 provides persistent storage volumes (using Amazon Elastic Block Store – EBS)
· A firewall that enables you to specify the protocols, ports, and source IP ranges that can reach your instances using security groups
· Static IP addresses for dynamic cloud computing (known as Elastic IP address)
· Amazon EC2 provides metadata (known as tags)
· Amazon EC2 provides virtual networks that are logically isolated from the rest of the AWS cloud, and that you can optionally connect to your own network (known as virtual private clouds – VPC)

98. What is Amazon Machine Image and what is the relation between Instance and AMI?

Amazon Web Services provides several ways to access Amazon EC2, like web-based interface, AWS Command Line Interface (CLI) and Amazon Tools for Windows Power shell. First you need to sign up for an AWS account and you can access Amazon EC2. Amazon EC2 provides a Query API. These requests are HTTP or HTTPS requests that use the HTTP verbs GET or POST and a Query parameter named Action.

99. What is Amazon Machine Image (AMI)?

An Amazon Machine Image (AMI) is a template that contains a software configuration (for example, an operating system, an application server, and applications). From an AMI, we launch an instance, which is a copy of the AMI running as a virtual server in the cloud. We can launch multiple instances of an AMI.

100. What is the relation between Instance and AMI?

We can launch different types of instances from a single AMI. An instance type essentially determines the hardware of the host computer used for your instance. Each instance type offers different compute and memory capabilities. After we launch an instance, it looks like a traditional host, and we can interact with it as we would any computer. We have complete control of our instances; we can use sudo to run commands that require root privileges.

101. Explain storage for Amazon EC2 instance.

Amazon EC2 provides many data storage options for your instances. Each option has a unique combination of performance and durability. This storage can be used independently or in combination to suit your requirements.

There are mainly four types of storage provided by AWS. Amazon EBS: Its durable, block-level storage volumes that you can attach to a running Amazon EC2 instance. The Amazon EBS volume persists independently from the running life of an Amazon EC2 instance. After an EBS volume is attached to an instance, you can use it like any other physical hard drive. Amazon EBS encryption feature supports encryption feature.

Amazon EC2 Instance Store: Storage disk that is attached to the host computer is referred to as instance store. Instance storage provides temporary block-level storage for Amazon EC2 instances. The data on an instance store volume persists only during the life of the associated Amazon EC2 instance; if you stop or terminate an instance, any data on instance store volumes is lost. Amazon S3: Amazon S3 provides access to reliable and inexpensive data storage infrastructure. It is designed to make web-scale computing easier by enabling you to store and retrieve any amount of data, at any time, from within Amazon EC2 or anywhere on the web. Adding Storage: Every time you launch an instance from an AMI, a root storage device is created for that instance. The root storage device contains all the information necessary to boot the instance. You can specify storage volumes in addition to the root device volume when you create an AMI or launch an instance using block device mapping.

102. Does deduplication occur inline or post process?

Inline. Why this matters:
With inline processing, deduplication decisions are made on the fly, as data arrives. Duplicate data is never written to media; this approach greatly benefits flash storage, which has a finite

number of write cycles before it degrades. More importantly, inline processing enables immediate replication for data protection. Post-process algorithms write all data at least once to a staging storage cache and perform the deduplication analysis after the fact. This consumes valuable storage space and can more quickly exhaust the life expectancy of flash devices and substantially increase system resource requirements. Many OEMs offering post-processing deduplication recommend not using the feature when consistent and predictable performance is needed, because they are prone to experiencing unexpected performance fluctuations due to dedupe background processing. Be wary of vendors who recommend turning their dedupe off when predictable performance is required from their storage system, because their implementation is not using optimal performance techniques that are available in today's inline deduplication engines. Finally, post-process implementations are not able to meet the system requirements of rapidly changing data sets, such as when many hundreds of desktops are cloned in a VDI environment. With post-process dedupe, the system can run out of available cache storage space almost immediately and will be unable to catch up.

103. What algorithm do you use for hashing and do you require any special hardware?

Support is provided for both hardware assisted and software only approaches. Why this matters: Deduplication algorithms employ content hashing to recognize duplicate blocks. Hashing maps the content of a block to a concise value. SHA-256 is known as a cryptographic hash and will produce a match only if two blocks match exactly. Using SHA-256 is a great approach, but it suffers from being slow on commodity CPUs (typically 100 MB/s of 4 KB blocks can be processed per processor core) and being computationally intensive. It's unlikely that a SHA-256 software computation will meet primary storage expectations without specialized hardware. If a vendor decides to deploy SHA-256, they should also employ an ASIC or FPGA to accelerate hashing.

In the absence of hardware acceleration, a fast (non-cryptographic) hash is the best solution. Fast hashes (e.g., MurmurHash3) can be computed 15-20X faster and performed completely in software. Using a single Xeon core, a typical software SHA-256 calculator might sustain data rates of 180 MB/s, whereas a MurmurHash3 calculator can sustain 3,000 MB/s. Since fast hashes are non-cryptographic, they do require a data read compare to verify matches. For deduplicated data blocks, you are essentially trading a write operation for a read (an operation that is generally faster, particularly on flash storage).

104. Does your product combine deduplication with block level compression?

Yes. Deduplication and compression are complementary data efficiency technologies. Why this matters: Many datasets see 2X data reduction from block-level compression techniques. However, when compression is combined with block-level deduplication, the benefits are magnified by as much as 17X. ESG Labs research shows that this is especially true in environments where large amounts of redundant data are stored (e.g., VM images) as well as in those with highly redundant workflows (e.g., database development and testing). The greatest benefit comes from solutions that combine the two data efficiency technologies.

The key to saving the most when these two technologies are combined is to deduplicate first and then compress. Since compression works file by file and deduplication works across petabytes of data, running deduplication first removes all the duplicates that would have been compressed if compression were run first. This sequencing results in less computational load on storage systems and delivers greater compute savings as well as greater data reduction across broad data streams.

105. **Should we use deduplication, compression, or both?**

Both. Deduplication and compression work independently and are complementary technologies that together can provide data reduction of up to 90 percent. Used alone, either can be effective at reducing disk backup capacity requirements, but an overall data reduction strategy includes both deduplication and compression. If you achieve 2:1 compression of backup data that has already been deduplicated by a 10:1 ratio, the result is a total data reduction ratio of 20:1.

106. **What is best, hardware-based or software-based data reduction?**

There are reasons to consider both, but compression in particular requires processing power, so it impacts backup performance least when it is deployed as a hardware-based solution. Hardware-based data reduction appliances also offer the advantage of compatible pre-configured applications, rather than a piecemeal collection of products that you must purchase, install, configure and manage yourself. Look for data reduction appliances that do not require a separate backup server, which adds complexity and increases the overall cost of the backup infrastructure. An 'all-in-one' backup data reduction appliance takes the guesswork out of implementing multiple solutions.

107. **Tape backup systems include compression; why can't we use hardware compression on my disk backup system?**

Disk compression has traditionally been file-based. Hardware compression works on data blocks, not files. Until recently, the challenges of implementing block-based data compression on disk have been insurmountable. That is now changing with the introduction of data-reducing backup appliances equipped with hardware-based disk compression.

108. **What about backups of remote servers, or backups that are sent offsite for disaster recovery purposes?**

Backup data reduction must address backing up the servers at HQ as well as remote offices, and sending backup data to a disaster recovery (DR) site. Backup should be a single, continuous process, centrally and conveniently managed, which is possible with data reduction appliances.

109. Can the same data reduction appliance for backup data be used for primary data?

Typically, deduplication and compression are applied to primary data in very different ways. Primary storage deduplication works on blocks of data aligned at the disk's boundaries. Windows and Linux file systems align the beginning of each file at the beginning of a block. This means that primary storage deduplication will always identify duplicate blocks within files. Also, databases read and write on fixed-size pages, so duplicate data within a single database or across databases can be detected. Backup applications, on the other hand, create files that are the equivalent of .tar or .zip files in which the blocks are not always aligned the same way, so backup deduplication applications have a very different job to do than primary data deduplication.

110. Why do data reduction solutions vary so much in price?

Just comparing prices won't provide a true apples-to-apples evaluation. A better metric is the cost per terabyte of backup to disk capacity. How many concurrent backup streams are supported? What is the impact on backup performance, if any? The wrong answers to these questions will actually cost you more even at the lower entry price point. BridgeSTOR's Virtual Storage - Advanced Data Reduction (VS-ADR) brings deduplication, compression, and thin provisioning to both application and backup-to-disk storage. VS-ADR is your ticket to vastly improved capacity utilization and "greener" storage. The AOS Backup Exec 2010 Deduplicating Backup Appliance, an all-in-one appliance that includes Symantec Backup Exec 2010 Deduplication Suite, sells for a street price of less than $USD 20,000.

References

[1] TechTarget., "What is Cloud Computing?" Source: http://searchcloudcomputing.techtarget.com/definition/cloud-computing (Accessed on 28th December 2018)

[2] Planets., "The Digital Divide Assessing Organisations' Preparations for Digital Preservation" Source: https://www.planets-project.eu/docs/reports/planets-market-survey-white-paper.pdf (Accessed on 28th December 2018)

[3] Cibecs., "Improving Backup and Restore Performance for Deduplication-based Cloud Backup Services" Source: http://cibecs.com/wpcontent/ uploads/2011/09/Survey-2011_Aug-E.pdf (Accessed on 28th December 2018)

[4] "The 2010 Digital Universe Study: A Digital Universe Decade – Are You Ready?" Source: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-youready.pdf

[5] Wikipedia.,"Cloud computing" Source: http://en.wikipedia.org/wiki/File:Cloud_computing.svg

[6] Roy, S., Bose, R., Sarddar, D. (2015). "A Fog-Based DSS Model for Driving Rule Violation Monitoring Framework on the Internet of Things". IJAST, Vol. 82, pp. 23 – 32.

[7] "The 2011 Digital Universe Study: Extracting Value from Chaos." Source: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaosr.pdf

[8] Tolia, N., Kozuch, M., Satyanarayanan, M., Karp, B., Bressoud, T. and Perrig, A. (2003). "Opportunistic use of content addressable storage for distributed file systems." Proceedings of the 2003 USENIX Annual Technical Conference. pp. 127-140.

[9] "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. "Source: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index gci/Cloud_Index_White_Paper.html. (Accessed on 31 Dec 2018)

[10] Quinlan, S. and Venti, S. D. (2002). "A new approach to archival storage." Proceedings of the first USENIX conference on file and storage technologies, Monterey, CA

[11] Denehy, T. E. and Hsu, W. W. (2003). "Reliable and efficient storage of reference data." Technical Report: RJ10305, IBM Research, Oct 2003.

[12] Andrew, T. (1999). "Efficient algorithms for sorting and synchronization." Source: https://www.samba.org/tridge/phd_thesis.pdf (Accessed on 1st January, 2019)

[13] Sarddar, D., Roy, S., Bose, R. (2014). "An Efficient Edge Servers Selection in Content Delivery Network Using Voronoi Diagram". IJRITCC, Vol. 2, No. 8, pp. 2326 – 2330.

[14] Schouten, E. (2014). "Cloud computing defined: Characteristics & service levels". Source: https://www.ibm.com/blogs/cloud-computing/2014/01/cloud-computingdefined-characteristics-service-levels/ (Accessed on 5th January 2019)

[15] Mell, P., Grance, T. (2009). "The NIST Definition of Cloud Computing". NIST, Info. Tech. Lab. Source: https://www.nist.gov/sites/default/files/documents/itl/cloud/clouddef-v15.pdf (Accessed on 5th January 2019)

.

[16] Won, Y., Kim, R., Ban, J., Hur, J, Oh, S. and Lee, J. (2008). "Prun: eliminating information redundancy for large scale data backup system." Proceedings IEEE international conference computational sciences and its applications (ICCSA'08).

[17] Won, Y., Ban, J., Min, J., Hur, J., Oh, S. and Lee, J. (2008). "Efficient index lookup for de-duplication backup system." Proceedings of IEEE international symposium modeling, analysis and simulation of computers and telecommunication systems (MASCOTS'08), Sept 2008, pp. 1–3.

[18] Kulkarni, P., Douglis, F., LaVoie, J. and Tracey, J. (2004). "Redundancy elimination within large collections of files." Proceedings of the USENIX annual technical conference, pp. 59–72.

[19] Kruus, E., Ungureanu, C. and Dubnicki, C. (2010). "Bimodal content defined chunking for backup streams." Proceedings of the 8th USENIX conference on file and storage technologies. USENIX Association.

[20] Policroniades, C. and Pratt, I. (2004). " Alternatives for detecting redundancy in storage systems data." Proceedings of the annual conference on USENIX annual technical conference. USENIX Association.

[21]] Kavitha, K. (2014). "Study on Cloud Computing Model and its Benefits, Challenges". IJIRCCE, Vol. 2, No. 1, pp. 2423 – 2431.

[22] Kubiatowicz, J. et al. (2000). "Oceanstore: an architecture for global store persistent storage." Proceedings of the 9th international conference on architectural support for programming languages and operating systems." ACM SIGPLAN Notices, Vol. 35, pp. 190 – 201.

[23] Quinlan, S. and Dorwards, S. (2002). "Venti: a new approach to archival storage." Proceedings of USENIX conference on file and storage technologies, (FAST'02), Monterey, CA,USA, Jan2002, pp. 7-7

[24] Rabin, M. (1981). "Fingerprinting by random polynomials." Center for Research in Computing Technology, Aiken Computation Laboratory, University.

[25] Lillibridge, M., Eshghi, K., Bhagwat, D., Deolalikar, V., Trezise, G. and Camble, P. (2009)."Sparse indexing: large scale, inline deduplication using sampling and locality." Proceedings of the 7th USENIX conference on file and storage technologies (FAST'09), San Francisco, CA, USA, Feb 2009, pp. 111–124.

[26] Muthitacharoen, A., Chen, B. and Mazi`eres, D. (2001). "A low-bandwidth network file system." SIGOPS Source: https://pdos.csail.mit.edu/papers/lbfs:sosp01/lbfs.pdf (Accessed on 6th January 2019)

[27] Zhu, B., Li, K, and Patterson, H. (2008). "Avoiding the disk bottleneck in the data domain deduplication file system." Proceedings of the 6th USENIX conference on file and storage technologies, (FAST'08), Berkeley, CA, USA, Feb 2008, pp. 269–282.

[28] Liu, C., Lu, Y., Shi, C., Lu, G., Du, D. and Wang, D. (2008). "ADMAD: application-driven metadata aware de-duplication archival storage system." Proceedings of 5th IEEE international workshop storage network architecture and parallel I/Os, (SNAPI'08), Baltimore, MD,US, Sep 2008, pp. 29–35.

[29] Mogul, J., Douglis, F., Feldmann, A. and Krishnamurthy, B. (1997). "Potential benefits of delta encoding and data compression for HTTP." Proceedings of ACM SIGCOMM'97 Conference on Applications, technologies,architectures,and protocols for computer communication , (SIGCOMM'97), Cannes,France,Sep 1997, pp. 181– 194.

[30] Bolosky, W.J,, Corbin, S., Goebel, D., Douceur, J.R. (2000). "Single instance storage in windows 2000." Proceedings of fourth USENIX windows systems Symposium, (SNAPI'08), Seattle,Washington,US, Aug 2000, pp. 13–24.

[31] Bose, R., Roy, S. "Synthesizing information security measures in the context of traditional IT infrastructure, and in the spheres of Cloud and IoT environments." Research India Publications, Delhi, India, ISBN: 978-93-86138-10-1.

[32] Thein, N.L, and Thwel, T. T. (2012). "An efficient Indexing Mechanism for data de-duplication." Proceedings of the 2009 international conference on the current trends in information technology (CTIT), pp. 1–5.

[33] Roy, S., Bose, R., Sarddar, D. (2015). "A novel replica placement strategy using binary item-to-item collaborative filtering for efficient voronoi-based cloud-oriented content delivery network". ICACEA, pp. 603 – 608.

[34] Meister, D. and Brinkmann, A. (2009). "Multi-level comparison of data deduplication in a backup scenario." Proceedings of SYSTOR'09: The Israeli experimental systems conference, May 2009, pp. 1–12.

[35] "Amazon CloudFront Documentation". Source: https://aws.amazon.com/documentation/cloudfront/

[36] "Data Domain LLC. Deduplication FAQ." Source: http: //www. datadomain. com/ resources / faq.html (Accessed on 5th January 2019)

[37] Meyer, D.T. and Bolosky, W.J. (2011). "A study of practical deduplication." Proceedings of 9th USENIX conference on file and storage technologies.(FAST'11) , pp. 85-90

[38] Panzieri, Ozalp, Babaoglu1, Stefano,Ferretti, Vittorio,Ghini, MorenoMarzolla,"Distributed Computing in the 21st Distributed Computing in the 21st Century: Some Aspects of Cloud

Computing" source: https://link.springer.com/chapter/10.1007/978-3-642-24541-1_30 (Accessed on 7th January 2019)

[39] Bose, R. Roy, S. Sarddar, D. (2015). "A Billboard Manager Based Model That Offers Dual Features Supporting Cloud Operating System And Managing Cloud Data Storage". IJHIT, Vol. 8, No. 6, pp. 229 – 236.

[40] Liu, F., Tong, J., Mao, J., Bohn, R., Messina, J., Badger, L., Leaf, D. (2011). "NIST Cloud Computing Reference Architecture". NIST, U.S. Department of Commerce, Special Publication 500 – 292, pp. 1 – 28.

[41] Benjamin, Z. (2008). "Avoiding the Disk Bottleneck in the Data Domain Deduplication File System." Data Domain, Inc. "Proceedings of the 6th USENIX Conference on File and Storage Technologies.(FAST'08), Feb 2008, pp. 269-282

[42] Wikipedia.,"Amazon Elastic Compute Cloud," Source: http://en.wikipedia.org/wiki/File:Cloud_computing.svg (Accessed on 4th January 2019)

[43] Bhagwat, D., Eshghi, K., Darrell, D. E. and Lillibridge, L. M. (2009) "Extreme Binning: Scalable, Parallel Deduplication for Chunk-based File Backup." 2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication

[44] Nurmi, D., Wolski, R., Obertelli, C. G. G., Soman, S., Youseff, L. and Zagorodnov, D. (). "The Eucalyptus Open-source Cloud-computing System." 9th IEEE/ACM International Symposium on Cluster Computing and the Grid.

[45] "Amazon Simple Storage Service." API reference, API version 2006-03-01," 2006. https://docs.aws.amazon.com/AmazonS3/latest/API/Welcome.html (Accessed on 4th January 2019)

[46] Wu, J., Ping, L., Ge, X., Wang, Y. and Fu, J. (2010). "Cloud Storage as the Infrastructure of Cloud Computing." ICICCI pp. 380-383

[47] Neto, M.D. (2014). "A brief history of cloud computing". Source: https://www.ibm.com/blogs/cloud-computing/2014/03/a-brief-history-of-cloudcomputing-3/ ((Accessed on 6th January 2019)

[48] "Amazon: Amazon simple storage service." Source: https://aws.amazon.com/s3/faqs/ (Accessed on 4th January 2019)

[49] Anand, A., Gupta, A., Akella, A., Seshan, S. and Shenker, S. (2008) "Packet caches on routers: the implications of universal redundant traffic elimination." Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication (SIGCOMM'08), New York,USA,Feb 2009, pp.219–230

[50] Anand, A., Sekar, V. and Akella, A. (2009). "SmartRE: an architecture for coordinated network-wide redundancy elimination." Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication (2009).

[51] Bolosky, W., Corbin, S., Goebel, D. and Douceur, J. (2000). "Single instance storage in Windows 2000." Proceedings of the 4th USENIX Windows Systems Symposium.

[52] Bonwick,J. (2009). “ZFS deduplication.” Source: https://blogs.oracle.com/bonwick/entry/zfs_dedup (2009) (Accessed on 4th January 2019)

[53] “Cisco:Wide area application services.” Source: http://www.cisco.com/c/en/us/products/routers/widearea- application-services/index.html (Accessed on 6th Januart 2019)

[54] “Citrix: Cloudbridge.” Source: http:// docs.citrix.com/en-us/legacy-archive/cloudbridge html (Accessed on 6th Januart 2019)

[55] Debnath, B., Sengupta, S. and Li, J. (2010). “ChunkStash: speeding up inline storage deduplication using flash memory.” In: USENIX Annual Technical Conference (2010).

[56] Dong, W., Douglis, F., Li, K., Patterson, R.H., Reddy, S. and Shilane, P. (2011). “Tradeoffs in scalable data routing for deduplication clusters.” Proceedings of the USENIX Conference on File and Storage Technologies (FAST).

[57] Drago, I., Mellia, M., Munafo, M., Sperotto, A., Sadre, R.and Pras, A. (2012) “Inside dropbox: understanding personal cloud storage services.” Proceedings of the 2012 ACM Conference on Internet Measurement Conference (IMC), pp. 481–494.

[58] “Dropbox.” Source:: http://www.dropbox.com (Accessed on 4th Januarty2019)

[59] Dubnicki, C., Gryz, L., Heldt, L., Kaczmarczyk, M., Kilian, W., Strzelczak, P., Szczepkowski, J., Ungureanu, C.and Welnicki, M. (2009). “HYDRAstor: a scalable secondary storage.” Proceedings of the USENIX Conference on File and Storage Technologies (FAST). pp. 197–210.

[60] ElShimi, A., Kalach, R., Kumar, A., Oltean, A., Li, J.and Sengupta, S. (2012). “Primary data deduplication-large scale study and system design.” USENIX Annual Technical Conference.

[61] “EMC: Achieving storage efficiency through EMC celerra data deduplication.” Source: http://china. emc.com/collateral/hardware/white-papers/h6265-achieving-storage-efficiency-celerra-wp.Pdf (2009). (Accessed on 4th January 2019)

[62] “EMC: Avamar.” Source: http://www.emc.com/backup-and-recovery/avamar/avamar.htm (Accessed on 4th January 2019)

[63] “EMC: Centera: Content Addresses Storage System, Data Sheet.” Source: http://www.emc.com/ collateral/hardware/data-sheet/c931-emc-centera-cas-ds.pdf (Accessed on 4th January 2019)

[64] “EMC: Networker.” Source: http: // en.wikipedia.org/wiki/EMC_NetWorker (Accessed on 6th January 2019)

[65] Guo, F.and Efstathopoulos, P. (2011). “Building a high-performance deduplication system.”: USENIX Annual Technical Conference.

[66] Hu, W., Yang, T. and Matthews, J.N. (2010). “The good, the bad and the ugly of consumer cloud storage.” ACM SIGOPS Oper. Syst. Rev., 44(3), pp. 110–115.

[67] “IBM: IBM white paper: IBM storage tank - a distributed storage system.” Source: https://www.usenix. org/legacy/events/fast02/wips/pease.pdf (2002) (Accessed on 1st January 2019)

[68] “JustCloud:” Source: http://www.justcloud.com/ (Accessed on 1st January 2019)

[69] Kim, D. and Choi, B. Y. (2012). “HEDS: hybrid deduplication approach for email servers.” 2012 Fourth International Conference on Ubiquitous and Future Networks (ICUFN).

[70] Kim, D., Song, S., Choi, B.Y. (2013). “ SAFE: structure-aware file and email deduplication for cloudbased storage systems.” Proceedings of the 2nd IEEE International Conference on Cloud Networking.

[71] Li, J., He, L.W., Sengupta, S. and Aiyer, A. (2009). “Multimodal object de-duplication.” Microsoft Corporation (2009). Patent

[72] Bose, R. Roy, S. and Sarddar, D. (2015). “User Satisfied Online IaaS Cloud Billing Architecture with the Help of Billboard Manager”. IJGDC, Vol. 8, No. 2, pp. 61 – 78.

Acronyms

J

K

L

M

W

X

Z