EDSi Enterprise Data Solutions, Inc.

Sun microsystems

StorageTek™

Windows Data Management Client
The WDMC

Archiving Unstructured User Data
A Technical White Paper

May 2009

Introduction

Storage needs have been growing at an explosive rate over the last several years. The amount of digital data that users generate is increasing, the size of digital files is increasing, and the amount of time that data must be kept is increasing with regulatory requirements. This trifecta of growth is creating headaches for IT departments. These patterns are not going to change in the foreseeable future, so IT administrators need to have a comprehensive data management strategy in place in order to stay ahead of the game.

This white paper shows how administrators can reduce their storage equipment costs, operations cost and administration costs by implementing a data management strategy using the Windows Data Management Client (WDMC) and Sun Microsystems' SAM-FS file system. This paper first reviews some of the challenges surrounding storage management, discusses the EDSI/Sun Solution and then presents a case study that focuses on managing unstructured user data.

The Problem

IT Administrators are constantly faced with growing storage needs from the user community. There are several factors that contribute to these storage demands:

- ➢ The number of digital elements is growing

  Although the concept of a paperless office has not really materialized, there is a digital copy of almost every document and data element in the office. From standard office documents, to digital scans, to digital music, pictures and video, the amount of data that users generate is getting bigger and bigger. This applies to industry verticals such as Medical Imaging, Media and Entertainment, Video Surveillance, Seismic & Geographical Studies; as well as to general user data such as home directories and user profiles.
- ➢ The size of digital elements are growing

  Color documents and scans are now commonplace, images are becoming higher and higher in resolution, and video surveillance data is being stored and kept for extended periods of time. Data keeps getting bigger.
- ➢ Data Retention Requirements are Changing

  Regulatory requirements such as Sarbanes-Oxley and HIPAA have extended the amount of time that data must be kept.

  IT Administrators' lack of understanding about the data is also contributing to the data retention problem as no one is knowledgeable enough to determine whether data can be deleted or not.

This growth pattern has evidenced itself both in structured data such as databases, medical image files, video files, and various application data files; as well as unstructured user data found in home directories and file shares.

The Band-Aid

The good news for administrators is that the rapidly increasing size of hard drives and storage systems makes it easy to cover up these problems simply by adding more storage. This strategy addresses the symptom but not the underlying problem. Eventually this strategy will run into roadblocks:

- Backup

  Increasing data volumes lead to increasing backup windows and increasing load on the backup system, network and host machines. IT Departments are forced to purchase more tapes and backup capacity to store data that, in many cases, either doesn't need to be backed up or doesn't need to be in a regular rotation cycle.

- Operations Costs

  A hidden cost for many IT Departments is the power and cooling requirements that are placed on Data Centers and Building Operations Groups. A study recently completed for the U.S. Congress recently concluded:

  > "Left unchecked, data centers could double their energy consumption over the next five years at a cost of $7.4 billion annually, according to a report [to the US Congress] issued by the U.S. Environmental Protection Agency. By 2011 the equivalent of 10 new power plants would be needed to supply 12 gigawatts of electricity unless the energy efficiency of data centers can be improved. That's bad news for the corporate bottom line and the environment. It's also a hit on taxpayers' wallets: federal government data centers alone consume about 10 percent of that electricity."[1]

  Disk-Only Backup and Storage Solutions DO NOT address this issue. The answer for power consumption is to move data that doesn't need to be accessed on a regular basis to tape. Studies abound which show the relative operations cost of Tape vs. Disk. The Clipper Report performed a Case Study where the tape power & cooling costs were 1/25th the cost of a comparable Disk System.[2]

- Equipment Costs

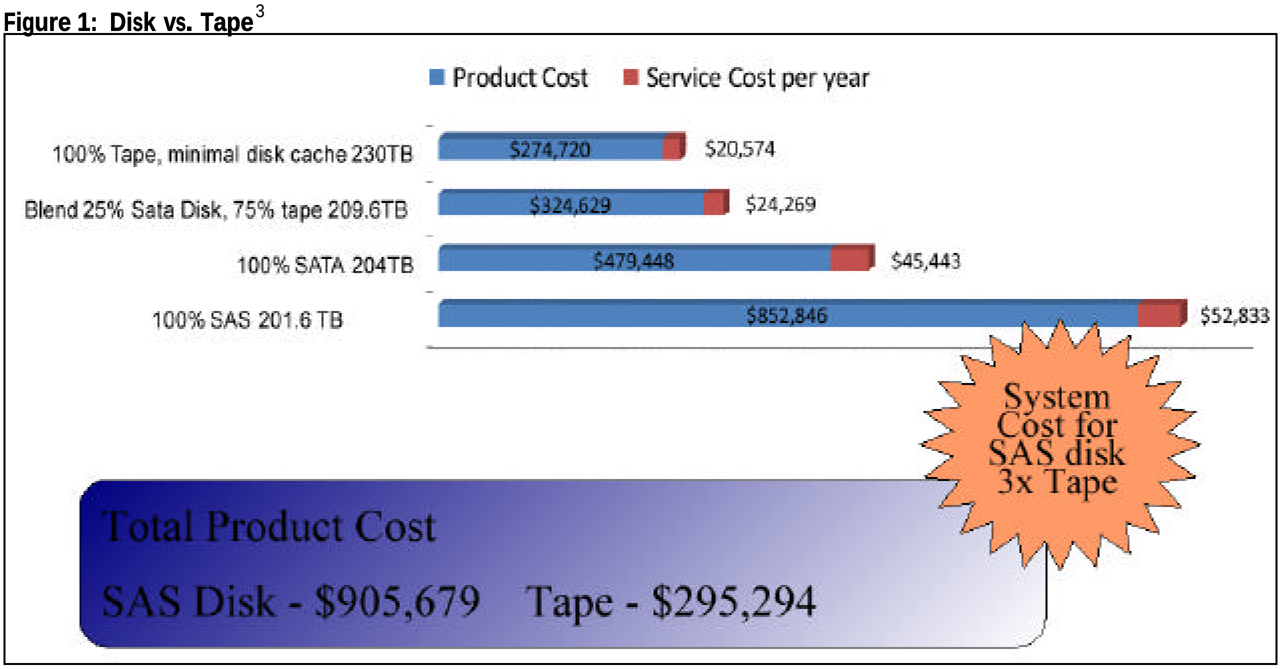
  Although spinning disk is fairly easy to implement, it is much more expensive to acquire than a tape-based solution. Figure 1 on the next page shows a comparison between the list price of a 200 TB Disk-Only solution compared to a comparable Tape-Based Solution. The tape system is one-third the cost of the comparable disk system.

  But the initial acquisition cost is only part of the story. A disk system must be replaced every three to five years, whereas a tape system will last 15-20 years. This means that the disk will be replaced approximately five times over the life of the tape system making the **disk 15 TIMES MORE EXPENSIVE** than the comparable tape system.

[1] Source: "http://www.treehugger.com/files/2007/08/the_problem_is_1.php"

[2] Source: "http://www.sun.com/storage/mainframe/docs/Clipper_Group_Tape_Disk_Costs_TCG2007014.pdf"

Figure 1: Disk vs. Tape[3]



Don’t Say ILM…

All of this points to the adoption of an Information Lifecycle Management System (ILM). For the purpose of this document ILM will be defined as a data archiving process that automatically moves data to the most cost-effective storage media, based on the predetermined policies of accessibility, security and long-term storage. ILM strategies and solutions were introduced as early as 2004 but have been implemented with mixed results. In order to implement an ILM strategy, one must first CLASSIFY all of an organization’s data and then determine an appropriate location (Tier 1 Disk, Tier 2 Disk, Tape) for the data. While it may be possible to classify structured data such as medical images or application files, the classification of unstructured data remains a challenge to deal with. There are tools to help identify data, but a true classification process requires input and analysis from all business units and data producers within an organization. This takes a lot of time, effort and money, requiring buy-in from the highest levels of an organization. The IT administrator may be tasked with dealing with the storage problem, but he or she may not have authority to design and implement a complete Data Management Strategy.

Sun Microsystems and EDSI have recognized the problems of getting an ILM strategy off the ground and have created a solution that enables IT administrators to manage their data without having to completely classify everything before starting.

[3] Source: Sun Microsystems 2008 Price List

The Windows Data Management Client

The WDMC provides an alternative to expensive ILM Solutions that still delivers many of the cost-saving features to data center administrators. The WDMC allows administrators to archive data based on simple metrics such as Date Last Accessed, Date Last Modified, File Type, File Size, File Location, and other readily available metadata. The WDMC stubs the file and automatically restores them when they are accessed so that users do not even need to be involved in the policy design process. The system sets the "Offline Bit" on files that are archived so users will know their data has been moved to a different tier and that they may need to wait when restaging files. The WDMC also provides tools that the users can use to manipulate their data so they can implement their own workflow and/or help to manage the process.

The WDMC provides a number of key features:

- Release Policy Options:

  The WDMC provides a number of options for selecting data to be released

  - High Watermark/Low Watermark (HW/LW): The WDMC can be used to automatically manage disk free space. When a high watermark is reached, data will be selected for release according to a formula that looks at both the size of the file and its age on disk. The HW/LW method also looks at special exclusions as defined below.
  - File Path: Customers who have a system for classifying their data can create policies based on file location. Certain directory trees can be defined for each of the data classifications created by the customer.
  - File Metadata: Policies can be created that look at metadata such as file type, date last accessed, date last modified or file size.
  - Manual: Users can manually release files and directories using the Explorer plug-in.

- High-Speed Restore:

  Using the stubbing functionality created for the archive component, the WDMC can quickly restore a snapshot of the file system while staging data in the background. This allows the system to be returned to use immediately. The most important data, as determined by the users, is restored first.

- Windows Explorer Plug-In:

  Administrators can optionally choose to install a Windows Explorer Plug-In on clients' computers to give them the ability to stage and release files or entire directories.

- High Performance

  WDMC is multi -threaded and utilizes a lightweight network protocol to efficiently transmit data. The WDMC can take advantage of 10GB Networks and performs well in large environments.

- Compression & Encryption

  The WDMC can compress and store data in a compressed format. The WDMC can encrypt data before sending it over the network and can store data in encrypted format to provide maximum security.

- Encrypted Authentication

  The WDMC utilizes an encrypted protocol for communication. This allows the software to backup and/or archive remote machines such as servers or laptops without the need for a VPN Tunnel or dedicated link.

Features (Cont'd)

- Multiple Versions

  The WDMC will manage multiple versions of a file. The number of versions kept can be controlled by either specifying a data range or maximum number of copies.

- Restricted Time Period

  Control the load on your system by only allowing the software to run during off-hours.

- Self-Service Restore

  Using the WDMC Client, users can be permitted to restore their own files, including older versions of a file. This saves costs by minimizing the workload on the existing IT Staff.

- Centralized Administration

  The WDMC provides an optional Centralized Administration Console that allows administrators to operate and monitor all machines from one place. The Centralized Admin can automatically send a variety of daily reports or alerts, and can configure policies on remote machines even when they're offline.

The next section provides a Case Study on a large company that utilized the WDMC and SAM-FS to create an archive system for unstructured User Profile Data.

Case Study: Archiving 5000 User Profiles at a large Media & Entertainment Company
This study shows how the WDMC was used to archive unused data out of user home directories and user profiles.

The Environment

This system was implemented in a Windows environment that had 5,000 users whose Profiles were stored on one of three central servers. The environment consisted of the following:

- Three (3) Windows 2003 Standard Edition machines running on HP DL385 Servers with two (2) Quad-Core CPUs and 8GB RAM each.
- 30 TB of SAN Storage running on Compaq EVA 5000, split evenly between the 3 servers. Approx 25 TB of storage was in use.
- SAM-FS Backend: Two (2) 5220 Servers running SAM 4.5, 6540 Primary Disk allocated from trays of 1 TB SATA Disks, 8500 Library with five (5) LTO4 Drives. The SAM-FS System is shared between multiple environments.
- 10 Million Files consuming 20 TB of data over three servers
  - Server 1: 3.8M Files, 5.4 TB
  - Server 2: 2.99M Files, 6.4 TB
  - Server 3: 3.27M Files, 7.8 TB
- NetBackup used for system backups. Full backups performed every weekend and incrementals nightly.
- Norton Anti-Virus used to protect servers, including Real-Time Protection. McAfee Anti-Virus on Windows Clients.

Configuration

The System was configured as follows:

- WDMC Software installed on each of the three servers.
- A dedicated 10GB Ethernet Network was installed on each server and on the SAM-FS Server.
- WDMC Configured to run 10 simultaneous threads on backup and scans.
- Backup and Release Policies: System was configured to backup all data into SAM-FS and then release data matching release criteria:
  - No HW/LW
  - Files were initially set to be archived after two years of inactivity
  - Exclusions for EXE's, DB Files, Temporary Internet Files & Cookies
- Run-Time Exclusions set for 7am – 11:45pm local time.

Results & Performance

The WDMC took a few nights to perform the initial backup. Due to the limited number of tape drives and the potential for large-scale restores, a very limited release policy was selected. Even the initial release policy of two years caused 25% of the data to be freed up. NetBackup and Norton anti-virus were configured to ignore files that were offline (released to SAM-FS).

After monitoring the system for several months, it was determined that very few files were being restored by the users on an ad-hoc basis. Several directories were manually restored for projects that needed the specific data. This was accomplished using the Explorer Plug-In in one instance and by the system Administrator in another. Policies were gradually changed to select more files for release. That is an ongoing process.

Important Notes

A number of items were discovered during the process of archiving User Profile data. A couple of points are listed below:

- User profile data contains a number of temporary files such as Temporary Internet Files. These should not be backed up, and the system has a capability to exclude all Temporary Internet files.
- It is very important to set user expectations with respect to offline files. The WDMC provides a number of features to help with that:

  The Offline Bit

  Windows has a feature called the Offline Bit. When the offline bit is set on a file, the icon changes and shows a black clock in the lower left hand corner of the icon to indicate that a user may have to wait for a file to be restored. The WDMC sets the Offline Bit for all files that have been released into SAM-FS. Instructing users on the meaning of the offline bit is very important.

  User Control

  The WDMC provides a Windows Explorer Plug-In that allows users to manually re-stage files or entire directories. This optional component would be installed on each client machine. This proved to be very valuable for certain users.

Conclusion

The trend towards increasing storage demands is not going to change in the near future. Although it requires work to implement, IT departments need to move toward a data management strategy that puts the right data on the right storage for the right amount of time. The WDMC with SAM-FS will provide organizations with the infrastructure needed to accomplish these goals, even with highly unstructured user data.

For More Information:

Enterprise Data Solutions, Inc.
1717 Superior Ave.
Cleveland, OH 44114
www.edsi.us.com
866-302-EDSI (3374)

Patrick Connor
patrick.connor@edsi.us.com