N94- 33803

Volume Serving and Media Management in a Networked, Distributed Client/Server Environment

Ralph H. Herring and Linda L. Tefend

EMASS® Storage Systems
Solutions from E-Systems
P. O. Box 660023
2260 Merritt Drive
Dallas, TX 75266-0023
Phone: (214) 205 - 6478
Fax: (214) 205 - 7200
lindat@Emass.Esy.COM

1. Introduction

Data storage requirements have increased exponentially in the last 10 years. While many things have contributed to this explosive growth, perhaps the biggest single cause is the increase in data processing capability brought on by the wide acceptance and use of supercomputers and large networks of workstations. This added processing power allows work on complex problems such as medical, digital imaging, modeling, and satellite data analysis that could not be tackled in the past. More significantly, added processing capability results in major increases in both the quantity and size of data files to be managed. Prior memory architectures have been out-dated by these changes. This results in a whole new field, the field of mass storage.

Figure 1 shows the major functional blocks of a classic mass storage system. The application program processes data and prepares it for initial storage. Access to the data by the application program is by the file name established when the data was initially stored. Other applications can share the mass storage system by common access to the file names or by use of their own names. These application programs can be on the same computer system (supercomputer, minicomputer, main frame, or workstation) or networked to the parent computer system.

The file server accepts file requests by file name. Because file storage is hierarchical, a file may be on solid state (RAM) memory, magnetic disk, optical disk, or tape. When the file server is asked to retrieve a file, it determines the file/medium relationship. If the file is stored on a medium managed by the volume server, the file server generates a media request to the volume server. After the medium is mounted, the file server receives file data from the storage drive.

The volume server accepts media requests, by media name, from the file server. The volume server maintains the relationship between each medium it manages and the associated media type and location. The volume server works with a variety of sizes and types of media. Although the volume server does not control read/write operations with the storage drives, it knows drive status and can maintain mount statistics and request queues for each drive.

A mass storage system can consist of several robotic and manual archives offering storage for several media types chosen for a variety of reasons (cost, convenience, speed, reliability, etc.) An archive houses the storage drives and delivers media to the drives. An archive recognizes media by external labels, so it has no need to know the information on the media. Storage drives provide the means to store and retrieve individual files. Storage drives interact directly with the file server to pass file data. Several drives can be associated with a single archive.

The E-Systems Modular Automated Storage System (EMASS) is a family of hierarchical mass storage systems providing complete storage/"file space" management. The EMASS volume server provides the flexibility to work with different clients (file servers), different platforms, and different archives with a "mix and match" capability. This volume server implementation encompasses the mass storage functions shown in figure 1. The EMASS design considers all file management programs as clients of the volume server system. System storage capacities are tailored to customer needs ranging from small data centers to large central libraries serving multiple users simultaneously. All EMASS hardware is Commercial-off the Shelf (COTS), selected to provide the performance and reliability needed in current and future mass storage solutions. All interfaces use standard commercial protocols and networks suitable to service multiple hosts. EMASS is designed to efficiently store and retrieve in excess of 10,000 terabytes of data. Current clients include CRAY's YMP Model E based Data Migration Facility (DMF), IBM's RS/6000 based Unitree, and CONVEX based EMASS File Server software.

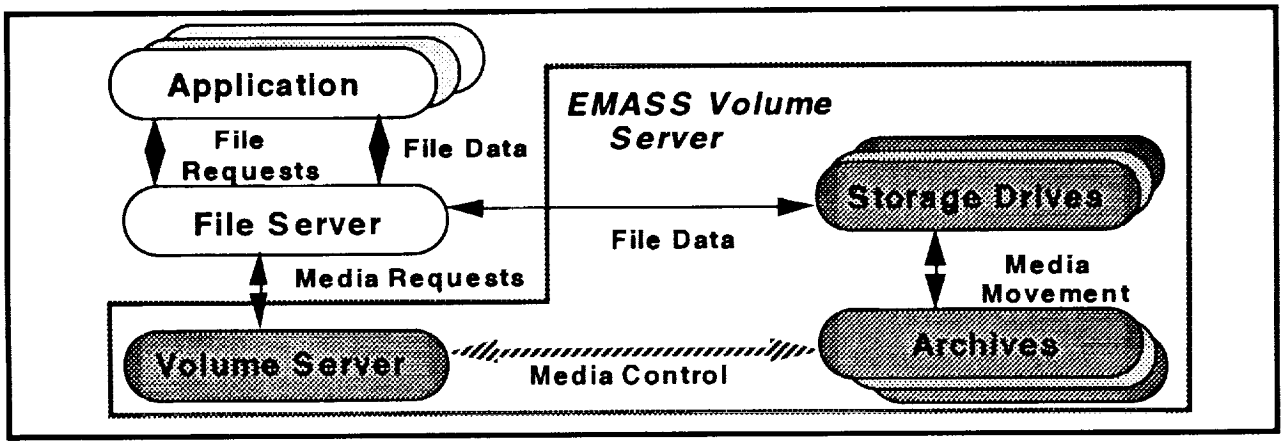


Figure 1 Mass Storage System Components

The VolServ™ software provides the capability to accept client or Graphical User Interface (GUI) commands from the Operator's Console and translate them to the commands needed to control any configured archive. The VolServ system offers advanced features to enhance media handling and particularly media mounting such as: automated media migration, preferred media placement, drive load leveling, registered MediaClass™ groupings, and drive pooling.

2. Mission

Provide Transparent Media and Drive Management The EMASS volume server provides the ability to accept and execute defined commands for media within its domain. The volume server finds and moves media based on logical name. If the request involves a mount, the volume server finds a storage drive compatible with the medium and accomplishes the mount. If the request involves media movement between archives, the volume server manages the move without involving the client. The volume server system can be applied to a large range of configurations with storage options involving data rates, media types, number of storage locations, and number of drives. The volume server system provides data in readily accessible, near immediate storage for purposes such as: history (archival backup), redundancy (data security), overflow (near line recovery of data as needed),buffer (temporary storage for later processing), and data distribution and transfer. Many applications require a single volume server system to provide for multiple networked clients. These clients do not have to be the same type of computer and may or may not share data, drives, or media. The volume server satisfies this mission with a design that can be used equally well for any of the listed purposes.

Minimize Impact of Utilizing Emerging Storage Technologies The EMASS volume server employs an object-oriented implementation to provide the ability to add new archives, drives, and interfaces in a modular manner. Existing applications can be preserved while new ones are added by including their specific control and status interface characteristics. The evolution in storage systems has been so rapid that any other approach would doom a volume server to obsolescence in the near future. The volume server uses COTS archives and drives with close attention to industry standards. The file server interface (volume name) need not change even to add new robotic archives, because this interface incorporates the "transparent" media location capability. Further, the VolServ software is both modular and portable as demonstrated through added archive types and porting to multiple process control computers, including SUN, IBM RS/6000, and CONVEX.

Conform to Industry Standards The IEEE recognized the need for a standardized way to structure memory storage systems through the development and release of its IEEE Mass Storage System Reference Model, Version 4 and the on-going work of Version 5. While this model is not yet released as an industry standard, it is being developed to allow and encourage vendors to develop inter-operable storage components that can be combined to form integrated storage systems and services. This model recognizes separation of the memory architecture into component elements including file management and volume management. The EMASS VolServ software provides the major Physical Volume Library (PVL) functions of centralized management of storage media, control of storage media architectures (PVRs), and automation of mounting and dismounting media into drive devices. The VolServ software is also designed to support multiple independent client systems. The VolServ software conforms to the concepts of the IEEE MSS Reference Model, ensuring it can readily adapt to future innovations in media storage architecture.

3. Library Services

The EMASS VolServ software represents the most complete media and drive management package available in the industry. The VolServ software provides the capability to accept client or Graphical User Interface (GUI) commands and translate them to the appropriate commands to control any supported archive. The VolServ software can service a variety of robotic archives and manual archives as shown in figure 2. Client commands are received through a layered Ethernet™ interface featuring a Remote Procedure Call (RPC) communication path; multiple clients can share the same interface. Operator commands are provided on a series of screens using the OSF/Motif GUI.

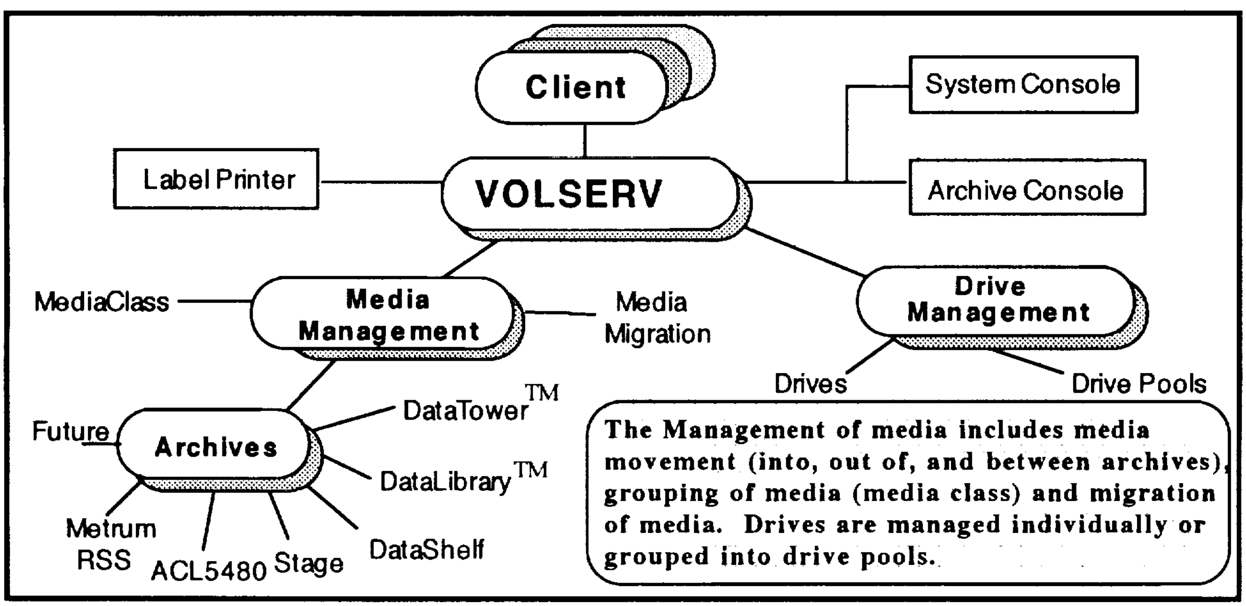


Figure 2 The Volume Server Emphasizes Media and Drive Management

Centralized Media Management The volume server provides a complete media management capability. The VolServ software automatically locates media within its domain. For example, the volume server receives a mount request, locates the requested medium, mounts the medium, and returns to the client the drive on which the medium was mounted. The volume server supports media migration between archives based on media type or media class.

Either source or destination migration archive can be robotic or manual. If multiple archives support the same media type, media can be migrated from one archive to another and, upon reaching the migration threshold of the second archive, to still a third archive.

Centralized Management of Storage Drives The VolServ software allocates storage drives for use by the client and controls placement of media into and out of the drives. The client provides control of read/write activities to/from the medium and releases the medium from the drive when finished. In automated archive systems, the use of storage drive types is restricted by the archive architecture. Manual archives can include drives of a variety of types. Each Archive Manager console has a screen which supports mounting and dismounting media. The VolServ software identifies which medium and drive to use.

Categories of Storage The volume server supports three categories of media storage. These are:

- Media within a robotic archive and available for near immediate data recovery by the client. The volume server assures media contained within an archive are suitable for mounting on drives associated with the archive.
- Media in a manual archive handled by an archive operator. The volume server provides clear operator instructions via Archive GUI consoles for media mounting on drives considered part of the manual archive and/or movement associated with drives contained in a robotic archive.
- Media which has been checked out and currently belongs to no archive. The VolServ software maintains the history of all checked out media to simplify future check in of the media and return to active control. The check out capability is separate from the "export" capability which removes the media from all databases.

Media Classes The EMASS volume server provides the ability to segregate media both physically and logically. Physical separation is done through archives and is enhanced by the ability to select preferred media placement within an archive. Preferred placement is implemented through the use of media classes. Media classes are a logical segregation of media based on client control and security needs. Media classes can be assigned to span multiple archives that support the associated media type(s). When a media class spans archives, media can be freely moved between these archives and automatic media migration can be used. The volume server provides the capability to define, modify, or delete media classes during initial system configuration or subsequently. Every medium known to the VolServ system must be associated with a media class. Media classes can segregate media by date, backup, inventory, per cent of medium filled with data, type of data or any other organizational need. Media classes figure prominently in the media mount algorithm.

Figure 3 shows a system configuration with four clients, two of which connect directly to the VolServ system and the tape drives. Four MediaClass groupings and two archives are shown. Since any client recording data on media needs access to scratch media, the "Scratch" media class is associated with both archives. Client A needs access only to Seismic data. If the "Seismic" media class is limited to Archive A, Client A needs no connection to drives in Archive B. Client B needs access to all four media classes and needs access to drives in Archive A and in Archive B. Clients C and D receive data through Client B. Neither client desires Seismic data, so they create traffic primarily for Archive B. A growth path could provide drive interfaces between Client C and the drives in Archive B. Media class "Maps" spans both archives and is ideal for migration and drive load balancing.

Membership in a media class is exclusive. Every medium belongs to one media class. A media class supports one media type. Media enter a media class as they are imported into the VolServ system. A default class can be specified (usually the "scratch" media class) for media auto-imported into robotic archives. The class for a medium can be entered via the Import command. The class associated with a medium can be changed via the reclassify command. A medium can, optionally, be reclassified as part of a mount operation.

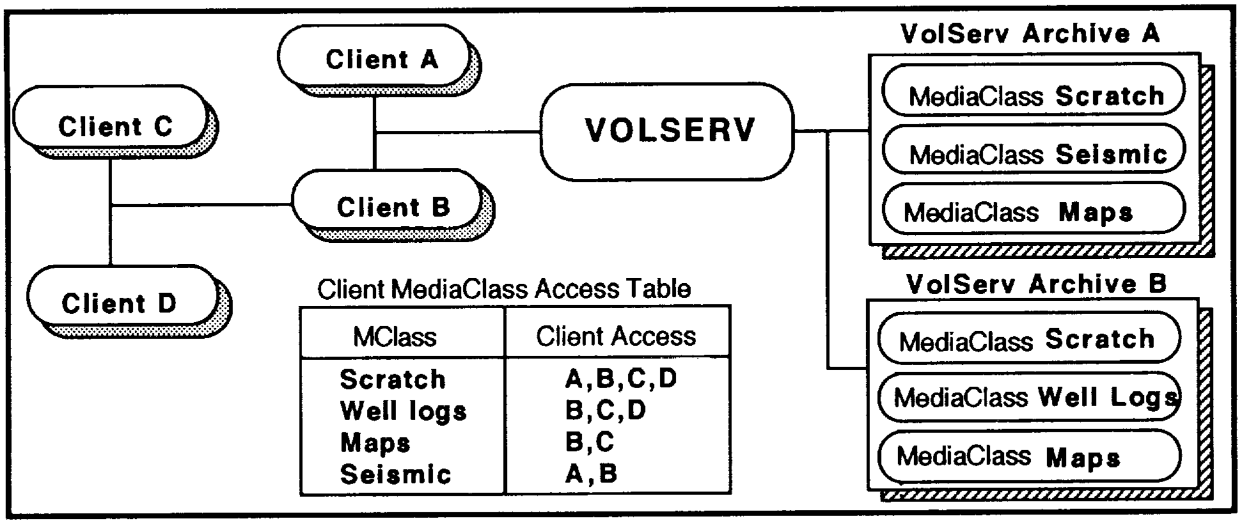


Figure 3 MediaClass Example

Media Migration The EMASS volume server system supports migration between archives based on media type or media class. Migration can be established at three levels: 1) automatic identification of the media to be migrated and their destination archive(s), 2) an operator notification when a user-specified threshold has been reached, and 3) no migration activity. A destination archive must be specified for each migratable media class.

A simple migration hierarchy includes a robotic archive and a manual archive. Scratch media are used and reassigned to a "permanent" media class in the robotic archive. When the media class high threshold is reached, the least-recently-used media are migrated to the manual archive. A medium can be recalled from the manual archive for use in the robotic archive. A medium can be exported (removed from the VolServ system) when it is no longer needed. Media migration can be used to accommodate other purposes:

- Migration from a robotic archive to another robotic archive - useful when one robotic archive provides better performance than another, or because the uses for the media change and different clients have access to one archive and not the other.

- Migration from a manual archive to another manual archive - useful when one manual archive is closer to the robotic archives than the other, because one manual archive has drives while the other has fewer or no drives, or because one archive is an organized DataShelf™ while the other is a "keep-it-for-awhile-longer" stage type archive.

- Migration to balance the load between similar archives with several clients as shown earlier.

- Migration set at a quantity of one (or two media) to provide one set of backups on-line while automatically migrating the previous backups to a manual archive or to a degauss and reuse category.

MediaClass Migration Each archive media class has associated migration parameters including capacity, high threshold, and low threshold. High and low thresholds are specified as a percentage of capacity so do not have to be updated when the capacity is changed. Capacity is the maximum number of media desired in the archive media class. High threshold is used to trigger migration processing. When automatic media migration is executed the VolServ software determines how many media must be removed to reach the low threshold and places those media on the ejection list. (Media can be removed from the eject list via the clear eject command.) An archive operator supports media migration by selecting media to be physically ejected from the eject list on the archive's console. Once ejected, these media appear on the destination archive's enter list. An operator completes migration by physically placing them into the entry port or manual interface for adding media to the destination archive.

MediaType Migration The volume server also provides automatic media migration for media types. Media type migration is conducted one media class at a time. When media type migration is triggered, the media class with the highest migration priority has its fill level lowered to its low threshold. This processing is applied, iteratively, to the media class with the next highest migration priority until the fill level for the media type reaches its low threshold. Depending on migration priorities and thresholds, migration may not be applied to all media classes.

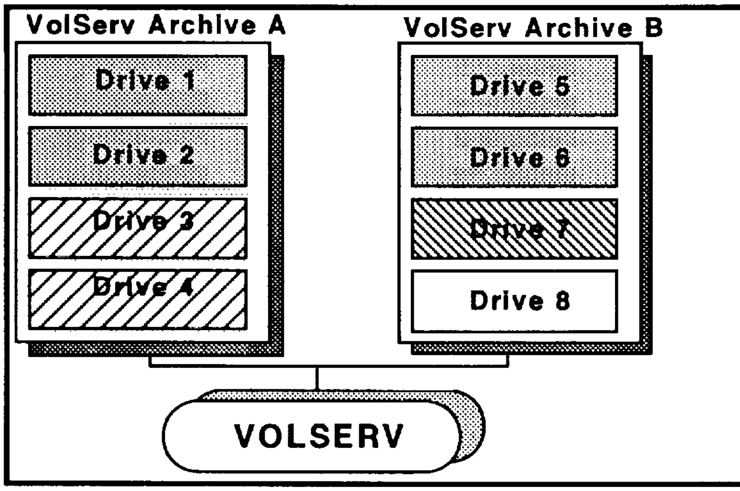
Use of the Low Threshold The VolServ system offers the capability to notify an operator when the number of media in a media class decreases below the low threshold. Low threshold notification can be used when scratch media are used and reassigned to "permanent" media classes or when media are exported, moved, or reclassified down to the low threshold. The client may use this information for inventory management, to keep a minimum number of scratch or in-work tapes, or other purposes. The client can ignore the notification or take an appropriate action such as adding media, reclassifying media, lowering the archive media class capacity, etc.

DrivePools A drive pool is a logical grouping of drives associated with one or more archives. A drive pool can frequently offer more rapid media mounting than the standard mount on a client-specified drive. Drive pools allow the VolServ software to select the best drive to satisfy a mount request. The volume server provides automatic media movement to get a medium in the same archive as the selected drive. A preferred solution is to satisfy a mount request within the archive that contains the medium. This solution is enabled by constructing a non-exclusive drive pool with at least one drive in each archive. If the medium has been relocated to a manual archive, a human is required to mount and dismount the medium. Figure 4 shows an example of drive pool organization for a system configuration with two archives, each with four drives. Drive pool 1 contains all four drives in archive A. This corresponds to the MediaClass grouping of figure 3 where all media for Client A are held in Archive A. When a drive pool contains all drives in an archive, it allows a medium to be mounted immediately on any available drive. Further, if the mount is queued, it is a candidate for the first available drive.

Drive pool 2 has two drives in each archive, ideal when the client has media classes in each archive, for example, Client B of figure 3. A client could include all drives in one archive and some drives in the other if this improves system operation. In this case, only two were chosen to ensure Client B never takes all the resources in Archive A. Drive pool 3 has two drives in Archive A. Pool 3 could be a second pool for Client A or for Client B. Client A may use pool 1 for data capture and hence want access to all four drives. Client A may use pool 3 for a less critical activity (data playback) to ensure playback operations never take all the resources. Pool 4 has only one drive. Drives can be added or deleted from a pool, so this could be a temporary state. A client may have committed to request all mounts by drive pool, but this function is of lower priority. The DataLibrary™ and the manual archives allow drives to mount more than one media type. If Archive B is a DataLibrary, pool 4 may be for D2 medium, while pool 2 is for D2 small. Drive 8 can only be mounted by requesting the specific drive.

Medium and Drive Mounting Options The volume server's goal is to mount a desired medium on any acceptable drive as quickly as possible by offering several options on the way the medium and the drive are chosen. These options take advantage of the media class and drive pool groupings. In each case, the mount will be queued if either the medium or the drive is busy. A sophisticated media-drive pairing algorithm selects media as follows:

- A specific medium - The VolServ software finds and mounts the user-specified medium on the nearest available on-line drive (if given a choice) or on the requested drive.



Pool	1	2	3	4
Drive 1	X	X		
Drive 2	X	X		
Drive 3	X		X	
Drive 4	X		X	
Drive 5		X		
Drive 6		X		
Drive 7				X
Drive 8				

Figure 4 Example Use of Drive Pools

- A list of media - The VolServ software pairs each medium in the user-specified listed of media with an available on-line drive (if given a choice) and selects the medium/drive pair requiring the least robotic movement.
- A MediaClass - The VolServ software exercises its drive-media pairing algorithm, using any medium in the user-specified media class, to find an available mount with the least robotic movement.

The EMASS volume server allows selection of drives in several ways to assist in satisfying clients requests:

- A specific drive - If the user-specified drive is available and in a different archive from the medium, the VolServ software initiates and controls an inter-archive media movement to get the medium to the drive.
- A drive pool - The VolServ software looks for an open drive in the user-specified drive pool that requires the least media movement.
- A drive pool with exclusions - After excluding the specified drives from the drive pool, the VolServ software processes this mount request the same as a mount by drive pool request.

Robot Allocation For most automated archives, a specific medium, a specific drive, and a specific load port can only be reached by a single robot. To move a medium, the supporting robot is scheduled to perform the movement. The DataLibrary has an enhanced capability when two or more robots are used. If a medium and a suitable drive or load port can be accessed by more than one robot, the VolServ software determines the first robot to become available. The goal is to match a medium, a drive, and a robot to minimize movement activity time.

4. Archives

A volume server archive is based on the Physical Volume Repository (PVR) definition in the IEEE Mass Storage Reference Model. A single instance of the EMASS system can be composed of one or more archives of the same or different architectures. Automated archives have self-contained, robotically-accessed media storage and retrieval providing a mechanical interface to the storage drives and providing load/unload ports for entering and ejecting media. Manual archives contain no robotics, so a human operator processes each media request in accordance with commands from the VolServ software to the appropriate archive console. Manual archives can include drives of a variety of types. Each Archive Manager console has a screen which supports mount and dismount of tapes.

The VolServ software supports four types of automated archives and two types of manual archives:

- **DataTower** stores 227 small 19 millimeter (D2) cassettes, has a single robot, supports up to four ER90™ tape drives, and provides 6 Terabytes of storage. Up to four towers can be interconnected as one archive with pass-through ports.

- **DataLibrary** provides expandable storage in increments of 4-foot cassette storage modules(CSMs), uses ER90 tape drives, and provides up to 5,000 Terabytes of storage. Each CSM can store 240 small, 192 medium, or a combination of small and medium 19 millimeter (D2) cassettes. A DataLibrary can be constructed with up to 20 aisles with up to 20 CSMs on each side of the aisle. Each aisle consists of a robot with access to any cassette in the CSMs on the aisle. Internal CSMs, drives, and load ports associated with internal aisles can be accessed by two robots. Drives can be located at either end of an aisle.

- **ACL5480** stores 288 3480 cassettes, uses one or two 3480 tape drives, and provides 58 Gigabytes of storage. Up to four 5480 units can be interconnected as one archive with pass-through ports.

- **RSS 48** and **RSS 600** store 48 and 600 T120 1/2 inch helical scan tapes each holding up to 14.7 Gigabytes of data for a total of 0.7 and 8.8 Terabytes respectively. The RSS 48 uses one or two drives, the RSS 600 uses one to five drives.

- **DataShelf**, a manual archive, stores 3480 cartridges, T120 cassettes, all three sizes of 19 millimeter cassettes, and up to 16 user-defined media types. A single DataShelf archive can support multiple media types and sizes. Storage is organized into rows, columns, shelves, and bins. Total storage capacity is limited only by facilities. This archive uses any storage drive type compatible with any supported media type.

- **Stage**, a manual archive, has the same capabilities as the DataShelf except the storage is free form. The stage archive can be used as an area to receive media for import or export or as a processing station for media needing cleaning, degaussing, certification, or other client desired processes.

The VolServ software design provides for addition or removal of individual archives from an established volume server system. Reconfiguration of an archive does not interfere with the operations of other archives. New archives and drives can easily be added to the volume server family with a minimum of effort. The volume server is modular so archive-dependent changes are constrained to the archive manager software. Clients can use multiple archives in a variety of ways: archives can be shared to provide client interusage, archives can be operated independently to provide data privacy and control, or archives can be structured in a hierarchy so media can be migrated between any combination of compatible archives. Media movement between two archives is directed by software, but performed manually. An operator can put an archive in an unattended mode. When a movement request involves an unattended archive, the VolServ software can cancel the request or wait until the archive is again attended.

5. Client Interfaces and Relationships

The VolServ software provides a control/status interface to the client software over a network. The VolServ software is connected to client program computers through an Ethernet or Fiber Distributed Data Interface (FDDI) connection using standard Remote Procedure Call (RPC) protocols. This connection allows multiple clients to share a volume server system. Figure 2 shows a stylized volume server environment with several potential file serving clients and the currently supported EMASS volume server archives. The client provides three hardware/software capabilities to use the EMASS volume server, namely the file server software application, the data and control interface to each drive, and the VolServ connect interface. These capabilities reside on each client machine connected via network to the volume server. EMASS imposes no limit to the number of clients that could be connected on the network path to the VolServ control processor.

A client system is a hardware/software package performing data management services for the client's own use or as an intermediary to other client programs. The VolServ software provides a high level interface relieving the client system of the need to know the storage architecture. The volume server offers transparency by locating and moving media based on its internal database. VolServ software provides a programmatic command set that allows a client to integrate any file management program with Client Interface Software (CIS). Through this CIS, the VolServ software may communicate with current and future file systems. In addition, EMASS has an Application Program Interface (API) and a command line interface (CLI) which simplify the client's interface design by residing on each client's computer and providing the RPC network interface.

The client interface can be implemented as an application program or as a modification to an operating system. Commands sent by an application program pass the required information to the volume server. For example, the VolServ software mounts the appropriate medium to allow the application to perform its read and write activities. One volume server can simultaneously perform similar services for several clients. Operating systems may include file management functions. If the operating system is providing file management, the VolServ CIS would be included as part of the operating system. Application programs would issue commands through the operating system which access the volume server transparently and the user would not be required to learn a new application.

The client provides file server software that determines what data is recorded on each medium and tracks data location with a cross reference to specific volume name(s). The client migrates data files onto media, identifies which medium contains which file, and requests the medium to restore data. The client provides the drive interface (data and control path) to each drive and knows which files are placed on which media and at what security or privacy level. Finally, the client provides the VolServ connect interface. This interface emphasizes standards so the client machine can use the RPC interface for all command and status data passed to/from the volume server. This interface represents only one load on the client's Ethernet or FDDI network. All traffic internal to the Volume Server is conducted over a separate Ethernet path. The client needs no direct interface with the archives for robot control.

6. Operations and Administration

System administrators and operators work directly with the VolServ software through the GUI for configuration, reconfiguration, archive management, media management, resource allocation, and daily maintenance operations for the volume server system. The System Administrator initializes and configures the volume server system and defines the associations between volume server components. The System Administrator login ID provides access to all VolServ software functions, while the operator login ID allows access to a subset. System Administrator/Operator interface GUI runs under any Motif window based manager. The VolServ software accesses the GUI directly from the VolServ control processor's console or remotely via a network. The control processor uses the OSF/Motif™ windowing system. The VolServ software must be installed on the host and INGRES® and the X-window manager must be running. The UNIX® shell used to initiate the GUI must be configured with environmental settings established through the software installation script.

The GUI operations are grouped and accessed through three types of consoles: the System Management Console, the Archive Console, and the System Log Console:

- **System management console** - provides access to logical operations and administrative functions. This console is generated by a System Administrator/System Operator.

Access is controlled through the use of passwords. Multiple consoles can be used simultaneously by System Administrators/Operators. Console GUIs are grouped into four functional categories: Media Operations, Administration, Configuration, and Queries/Reports.

- **Archive console** - used to execute media movement commands and manage the media stored within an individual archive. A volume server system has a separate archive console for each configured archive.

- **System logging console** - displays system messages. Logging consoles are generated automatically by the VolServ software. Message levels are defined during configuration by the System Administrator. System messages, generated during system operations, provide information about events occurring within the volume server system. The logs can be displayed on one or multiple consoles and/or directed to one or multiple files. Operator options for managing the display provide for clearing the display, printing the text in the display buffer, saving the information into a file, and setting auto-scroll on and off. The upper limit of displayable information is configurable. Once the limit is reached, the oldest messages are removed as new messages enter the display buffer.

7. Summary

The EMASS Volume Server provides all the media management capabilities to build a mass storage system *now* using a design and current hardware that can be used well into the future. These capabilities include:

1. **Transparent media and drive management**. The client provides the volume server with a media name or class and a drive or drive pool. The volume server advises the client when the medium is mounted (or moved).

2. **Multiple archive control**. The volume server is configured to manage several types of archives. Storage flexibility is enhanced by the capability to manage multiple copies of each archive.

3. **Multiple media (and drive) types**. The volume server supports standard media types and 16 user-specified types. The volume server supports robotic mounting of drives in robotic archives and operator mounting of drives in manual archives.

4. **Easily expandable**. The same VolServ software supports one archive or a variety of archives. Archives can be added with minimal impact to existing operations. Based on designed-in modularity, when new archive types are added to the volume server design, these archives could also be added to an on-going site.

5. **Advanced mounting algorithm using media classes and drive pools**. The mounting algorithm considers media, drives, drive-load balancing, robots, and operator support (for cross-archive mounts) in choosing the best media-drive pairing for each mount.

6. **Media migration between archives**. The volume server uses media type and class capability to support automated or operator-directed media migration from any type of archive (manual or robotic) to any other type.

7. **Full-featured manual archives**. Manual archive support includes a simple table-top used for media entry, up to a multi-row shelf archive with thousands of volumes, multiple media types, and a variety of drive types.

8. **Software portability (now runs on Sun, IBM and Convex platforms)**. The software emphasizes basic UNIX concepts. It can be readily ported to other platforms running versions of the Unix operating system.