N94-33807

VOLUME SERVER - A SCALABLE HIGH SPEED AND HIGH CAPACITY MAGNETIC TAPE ARCHIVE ARCHITECTURE WITH CONCURRENT MULTI-HOST ACCESS

Fred Rybczynski

Metrum,Inc.
10948A Beaver Dam Road
Hunt Valley, Maryland 21030
Voice: (410) 771-9207
FAX: (410) 771-9210
email: rybski@metrum.com

INTRODUCTION

A major challenge facing data processing centers today is data management. This includes the storage of large volumes of data and access to it. Current media storage for large data volumes is typically off line and frequently off site in warehouses. Access to data archived in this fashion can be subject to long delays, errors in media selection and retrieval, and even loss of data through misplacement or damage to the media.

Similarly, designers responsible for architecting systems capable of continuous high-speed recording of large volumes of digital data are faced with the challenge of identifying technologies and configurations that meet their requirements. Past approaches have tended to evaluate the combination of the fastest tape recorders with the highest capacity tape media and then to compromise technology selection as a consequence of cost.

This paper discusses an architecture that addresses both of these challenges and proposes a cost-effective solution based on robots, high-speed helical scan tape drives, and large-capacity media.

DATA CENTER PERSPECTIVE

Significant advances in magnetic tape drives, media capacities, and the integration of robotics now make it possible for most sites to maintain a significant portion, if not the entire set, of data in the computer center. Using these new technologies, the amount of floor space required for data storage is significantly reduced. (For example, the Metrum RSS-600b robot system contains 10.8 terabytes in less than 20 square feet of floor space. This is equivalent to approximately 60,000 reels of 9-track tape.) Media, housed within computer-controlled and robot-accessible carousels, is accurately identified by barcode readers integrated within the robotics. Following identification, media is rapidly retrieved and loaded by the robot. The high speed tape drives quickly locate the data and convey it to the computer host.

These new technologies enable the co-location of large volumes of data with computer hosts, thereby expediting data access and analysis. However, they are but mere tools, incapable of performing any data management function by themselves. Management of the data is performed by software that can manipulate the previously-described tools to achieve efficient data storage and retrieval. Various software data and storage management solutions are available. Some, such as UniTree™, perform a data migration function. They transfer files through a hierarchy of storage technologies measured by speed, capacity and cost (Figure 1) under the direction of a software-implemented algorithm responsible for managing the computer system's mass storage resources.

Others, such as AMASS™, perform a network-attached archival function (Figure 2). They present the entire archive storage capacity as if it were a huge disk-based file system. Data, referenced by a path-qualified file name, is transferred only in response to explicit commands received from a user or an application process. Although these archival systems are typically not delivered with software to perform behind-the-scenes file migrations, they can be used to accomplish a limited version of these functions.

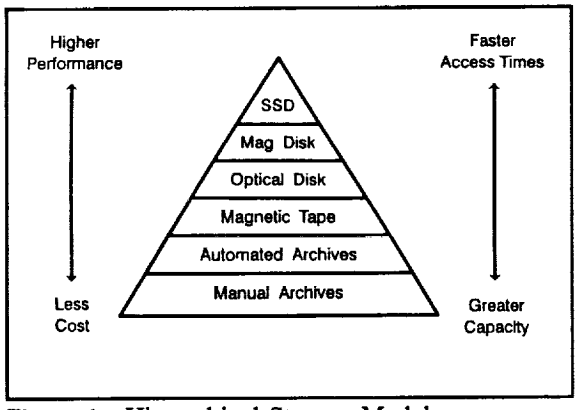

Figure 1. Hierarchical Storage Model

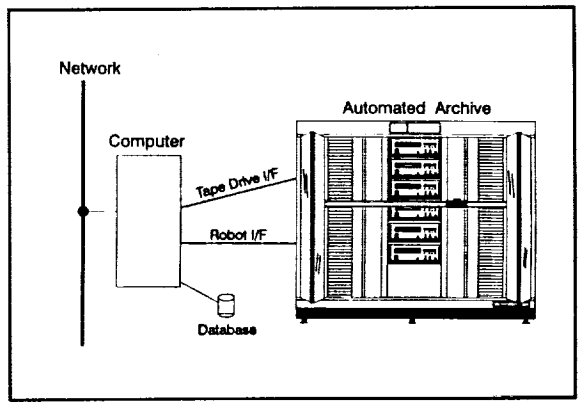

Figure 2. Network-Attached Archive Architecture

The predominant view of hierarchical and archival mass storage today is based on single-point concepts. First, a single database maps the storage system. Second, a single computer manages both the database and the storage system. Third, there is typically a single conduit for the transfer of data into and/or out of the storage system. While these singularities simplify the application of data management over the mass storage system, they have the undesirable side effects of slowing data access and reducing data transfer rates.

Instances may occasionally arise when it is very desirable to remove as many of these singularities as possible in order to expedite the storage and retrieval of data among cooperating computer systems. This "strawman" paper describes an architecture, termed Volume Server Architecture (VSA), for sharing large volumes of data and storage capacity among independent computer hosts.

DATA ACQUISITION PERSPECTIVE

Hierarchical and archival architectures are seldom acceptable for high speed data acquisition. The additional levels of data handling and detail tracking can introduce significant latencies to the data storage process, such that newly available data may be lost because its arrival rate is faster than the rate at which it can be stored. The data acquisition goal is to be able to store data at least as fast as the input rate; preferably faster. This takes shape with the selection of an acquisition technology that meets and/or exceeds the input rate.

The high-speed input data stream is often a multiplexed composite of independent sub-streams related by time. The decision to multiplex data is often determined by restrictions imposed by the delivery method. For example, perhaps only one satellite channel is available. An 8 megabytes per second (MB/S) data stream may actually be composed of 4 independent data streams, each arriving at 2 MB/S. While the data could be acquired at 8 MB/S, a cost effective alternative may be to acquire each of the 4 streams with a separate storage device at 2 MB/S.

It is frequently more efficient to store data in its de-multiplexed state if the data streams are primarily going to be analyzed individually and not as a composite. First, it is more efficient. Reading multiplexed data to extract 1 out of 4 bytes wastes 75% of bandwidth and increases the analysis period. Second, storing de-multiplexed data can extend the recording period because fewer media changes will need to be performed. Third, it can result in significant cost savings because a lower speed (and cost) recording technology can be used.

The volume server architecture discussed in this paper promotes automated data acquisition through the use of robots. It encourages data sharing because multiple hosts can have concurrent high speed access to the data storage reservoir through one or more locally-attached tape drives. Rapid access to recently-acquired data is possible even while more new data is being recorded.

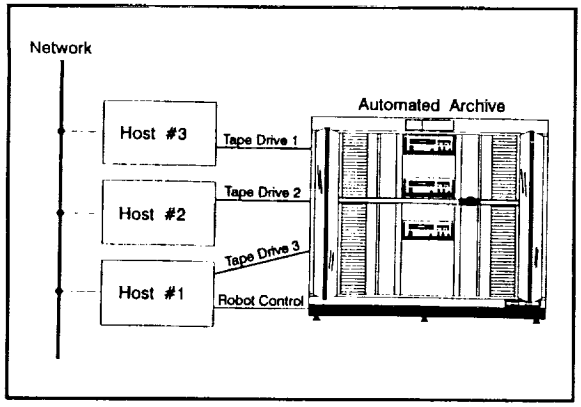

Figure 3. Simple VSA Configuration

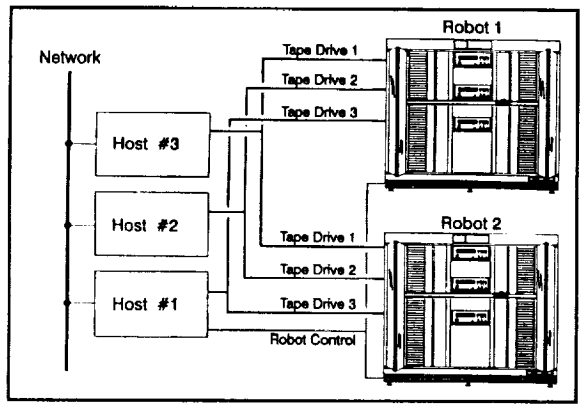

Figure 4. VSA with Multiple Robots

VOLUME SERVER ARCHITECTURE

The Volume Server Architecture (VSA) is only appropriate for select applications. It does not attempt to address the same needs as hierarchical or archival architectures and is not intended to compete with or replace them. Instead, its capabilities are best suited to applications:

- That acquire large volumes of data at high data rates.
- That have a need to share this data with other local systems on a continuing basis.
- Where the volume of data and/or the rate at which it must be received for processing exceed the available capacity of the local network.
- Where the use of operators to manually transfer media from one system to another presents unacceptable risk due to the possibility of the loading of incorrectly identified media and/or the possibility of loss of data due to misplacement or damage to media.

The purpose of the Volume Server Architecture is to:

- Maximize access to large amounts of robot-accessible storage capacity.
- Maximize access to large amounts of robot-accessible data.
- Maximize data transfer I/O rates by removing the restrictive actions of networks and providing direct access to data via a locally-connected tape drive.
- Provide a scalable architecture that can grow as the amount of on-line data grows.

The Volume Server Architecture is constructed using the same components as archival and hierarchical architectures:

- High speed tape drives.
- High capacity media for the storage of large amounts of data in a minimum "footprint".
- Cartridge-based media amenable to handling by robotics for "lights out" operations.
- A machine-readable media labeling system so that each cartridge is uniquely identifiable.
- A barcode reader to accurately establish media identity through automation.
- Robots to perform media transfer functions in response to computer requests.
- Robot-accessible storage bins for automated media storage and/or retrieval.
- A high speed interface linking the tape drive to the computer(s).

Figure 3 displays a simple VSA configuration showing concurrent data access for three computer hosts. Each of the hosts is directly connected to one tape drive in the robot, giving it direct access to all the data accessible to the robot. Host #1 controls the robot actions and processes all load/unload requests. The load/unload requests from Hosts #2 and Host #3 are received and responded to by Host #1 through the local area network link.

Table I. Database Information Fields.

Cartridge Inventory	Data Inventory	Free Space Inventory
Cartridge Barcode	Data Label	Cartridge Barcode
Storage Bin ID	Cartridge Barcode	Bytes Remaining
	Start Block Number	Robot ID
	Number of Bytes	Storage Bin ID
	Robot ID	
	Storage Bin ID	

The following discussions are based on the configurations depicted in Figure 3 and Figure 4.. These discussions not intended to be comprehensive, but only to provide sufficient detail to demonstrate the capability, feasibility, and extensibility of the volume server architecture.

STORAGE SYSTEM INDICES

Three databases are relevant to this discussion. The first database is called the cartridge inventory and links the unique barcode attached to each cartridge to its storage bin location within the robot. The second database is referred to as the data inventory and links a label (such as a file name) that is associated with the data to the data's storage location. The third database is referred to as the free space inventory and uses the cartridge barcode to track the location of available data storage space. Table I lists the information fields tracked by each database. The duplication of data fields between the database lists shown in Table I is for clarity. Specific database structure is left to the implementor.

The cartridge inventory is unique to each robot in the system. That is, each barcode and consequently each cartridge may only appear in one and only one cartridge inventory. This inventory is composed of only two fields: the cartridge barcode and the cartridge storage bin location within the robot.

The data inventory tracks information necessary to identify and locate any item of data available within all robots in the interconnected storage system (see Figure 4). This inventory is interrogated when the host knows the data label and wants to determine location information in order to access the data. It is updated whenever data sets are added to or deleted from cartridges in the storage system.

The free space inventory is queried whenever data is to be added to the storage system. It identifies those cartridges with space available to house the data. The cross reference to robot ID enables the host to select a specific robot system to store the data. This database is updated when new data is stored and when existing data, located as the last data set on a cartridge, is deleted and the previously-occupied space can be overwritten by new data.

The preceding database description is intentionally simplistic. For example, it does not address the possibility that data sets can span multiple cartridges, nor the possibility that data sets may be discontinuous across one or more cartridges. A more detailed development of the database structure and specifics of its implementation is beyond the scope of this "strawman" paper and is left to the reader.

OPERATIONAL DESCRIPTION

At system initialization time, Host #1 (the host controlling the robot) :

- Checks each tape drive position to be sure that there is no cartridge waiting to be stored. Cartridges are returned to their last known storage location. If that location is occupied, the cartridge is stored in the first open location. If no open storage location is found, Host #1 declares an error condition and solicits manual operator intervention.
- Uses the robot and barcode reader to establish the cartridge inventory database for that robot. Each cartridge storage location is investigated in order to determine the presence or absence of a cartridge. The barcode is read and stored whenever media is determined to be present.

At system initialization time, all hosts:

- Interrogate the tape drives directly connected to them to determine if a cartridge is already loaded. The cartridge can be ejected for storage or retained for further processing at the discretion of the local host.
- If a cartridge has been ejected, the local host must request that Host #1 either place the cartridge into a storage location or that Host #1 use the robot media handler to re-insert the cartridge into the tape drive (ie, "push" the cartridge back into the drive until the elevator mechanism grabs it and takes control).

During system operation, cartridge load and unload requests are issued to Host #1 for processing using the local area network. Table II lists the principal requests and corresponding responses possible for this dialog.

As shown in Figure 3, the tape drives are connected directly to the host requesting the service so that data is transferred directly from the host to and/or from the cartridge at maximal speeds. The data transfer bandwidth of the other tape drives in that robot is unaffected since each tape drive is connected to its respective host through its own dedicated interface bus.

Table II. Cartridge Manipulation Requests.

Command	Description	Possible Responses
INSERT *dd*	Use the robot to re-insert (push) the cartridge that is already loaded in the tape drive *dd* opening.	DONE
LOAD *xx dd*	Load cartridge with barcode *xx* into drive *dd*	DONE LOAD ERROR (drive problem) NOT AVAILABLE NOT FOUND PICK FAILED (cartridge problem)
QUERY *xx*	Determine whether cartridge *xx* is available or if it is in use by another tape drive.	AVAILABLE NOT AVAILABLE NOT FOUND
STORE *dd*	Return the cartridge in drive *dd* to its storage location.	DONE NO ROOM (robot full) PICK FAILED (cartridge problem)

It may occasionally be desirable to connect two or more tape drives, located within a single robot, to one host so that read and write operations can proceed independently, or in order to support multiple concurrent accesses. The additional tape drives could share the same interface bus or each could be connected using separate buses to maximize bandwidth. Decisions to utilize separate buses must consider the portion of time the tape drives might be involved in input and output versus the portion of time spent positioning to a new location on the tape. A tape drive places no load on the bus bandwidth during positioning functions.

EXTENSIBILITY OF THE ARCHITECTURE

Metrum tape drive features and robot configurations are identified and described in Tables III and IV in order to enhance the relevancy of further discussion. The information in these tables establish a real-world measure of capacity and transfer rates. All referenced components are available as commercial off-the-shelf (COTS) products. Their capacity and throughput numbers do not represent theoretical possibilities, but reflect actual system performance measures.

Extensibility is illustrated in Figures 3 and 4. Figure 3 shows the simple case of one robot shared by three hosts. Figure 4 shows how multiple hosts can access data residing in multiple robots.

Table III. Metrum Tape Drive Features.

RSP-2150 Tape Drive	2 MB/S Sustained
	4 MB/S Burst
	Track Addressable
	Record Addressable
	Robot-Compatible Media (S-VHS)
S-VHS Cartridge Media	DDC-258 (14.5 GB)
	DDC-343 (18 GB)
	$1.30 per Gigabyte

The American National Standards Institute Small Computer System Interface specification (SCSI, ANSI X3.131 1986) indicates that a host, through a single SCSI host bus adapters, could be connected to up to seven tape drives on a single SCSI bus. If each one of these tape drives were located in a different RSS-600b robot, the computer host would have immediate and unattended access to more than 75,000 gigabytes (ie, 7 tape drives * 10,800 GB/robot) of data storage capacity within a total floor space of less than 140 square feet. This is equivalent to approximately 420,000 reels of nine-track tape occupying more than 11,000 square feet of floor space. In essence then, each SCSI host bus adapter represents the potential for up to 75,000 gigabytes of robot-accessible data storage. At the same time, the seven RSP-2150 tape drives represent a total I/O bandwidth of 14 MB/S for that host.

Data archive capacity can easily be increased if the host is able to accommodate additional SCSI host bus adapters. Three host bus adapters represent the potential for 225,000 gigabytes of data storage capacity in 420 square feet of floor space. The equivalent volume of data on nine-track tape would require 1,260,000 reels of tape (at 6,250 bpi, 2,400 feet long, and 180 megabytes per reel) and more than 30,000 square feet of floor space. The entire 225,000 gigabytes of storage capacity can be shared with six additional hosts using the same principals as shown in Figure 3 and Figure 4.

[NOTE: The ANSI SCSI specification x3.131 1986 gives specific cable length limitations. These can be extended through the use of SCSI bus repeaters and/or fiber optic bus extenders.]

Table IV. Metrum Robot Configurations.

	RSS-48b	RSS-600b
Tape Drives	2	6
Cartridges	48	600
Capacity (18 GB/Cartridge)	864 GB	10,800 GB
Robot Cost Per Megabyte	$ 0.08	$ 0 02
Control Interface	RS-232	RS-232
Floor Space Required	6 Ft^2	19 Ft^2
Equivalent 9-trk Reels	4,800	60,000
Equivalent 9-trk Floor Space	125 Ft^2	1,600 Ft^2

The robot command language can be extended to enable a computer host to specify a robot identifier. In this way the host is able to access cartridges in automated storage systems configured with more than one robot.

- The host could QUERY if a data set is available in the cartridge inventory of a specific robot. The response could be "AVAILABLE", "NOT AVAILABLE", or "AVAILABLE IN rr", where "rr" identifies the robot containing the specified data set.
- The host can request that a cartridge be loaded into a specific drive in a specific robot. The robot identifier would be mandatory. Absence of a robot identifier would generate an "INCOMPLETE COMMAND" response.

The database and the robot command language can be extended to support controlled access to multiple copies of a data set. For example, the results of the database query may report that the primary tape cartridge copy with the requested data set is in use, but that another copy is available. If both the primary and secondary copies reside in the same robot, the robot-controlling computer simply loads the secondary copy. However, if the secondary copy resides in a different robot, the robot-controlling computer might respond to the LOAD request with "AVAILABLE IN rr". This efficiently conveys that the load request could not be satisfied and that robot "rr" has an available copy of the data set. The requesting host can then decide if it wants to issue a revised LOAD command, probably determined by the availability of its tape drive in robot "rr".

ADVANTAGES OF THE ARCHITECTURE

A significant advantage is the ability to scale system storage capacity in response to system needs:

- The number of tape drives within a single robot can be as few as one or as many as the robot can contain. (The Metrum RSS-600b robot can contain a maximum of six tape drives.)
- The number of robots is limited by the number of tape drive connections a host can support. It can range from as little as one robot and scale up to the maximum connections possible. Since each robot represents up to 10,800 gigabytes, a very wide range of data storage capacity is possible.

System bandwidth can be scaled in response to system needs:

- The implementation of multiple tape drives installed in one or more robots can dramatically increase the number of file transfer operations that can occur concurrently. For example, if seven tape drives were connected to a single host, up to seven I/O operations could occur concurrently in any combination of read, write and/or data search.
- The number of tape drives can be increased over time in response to rising system load in order to increase bandwidth. A host system can be configured with one Metrum RSP-2150 tape drive or up to the system-supported maximum. Table III lists the RSP-2150 tape drive features. Seven RSP-2150 tape drives can support a sustained transfer rate of 14 MB/S and burst rates of up to 28 MB/S for markedly less cost than a single 19mm tape drive. (This analogy is only valid if either the mandated data acquisition rate is in the range of the RSP-2150 or if the overall data stream can be demultiplexed and the resultant sub-streams of data have rates in the range of the RSP-2150.)
- A SCSI bus interface can transfer data faster than most network media-plus-protocol combinations. SCSI-connected tape drives with robot-assisted access to media in the storage system can be used to transfer data at sustained rates significantly faster than the network could support.

The ability to directly connect multiple hosts to the same data reservoir optimizes overall system throughput:

- 2 MB/S at each of 7 sites represents a significantly higher bandwidth than 14 MB/S at a single site if the data from the single site subsequently has to go through the restrictive bandwidth of a network in order to reach the other 6 remote sites. For example, 300 KB/S is typically the maximum sustainable transfer rate for a 10 megabit baseband Ethernet network.
- Data access times will be faster, since each tape drive can position directly to a data starting location on tape without having to first wait for data transfer processes on other tape drives to complete.

Improvements in operations:

- Automated media management means reductions in errors, loss of data, and associated recurring costs.
- Automated media management performs accurate and very fast media retrieve/store operations.
- The automated storage system can periodically analyze all media for wear without operator intervention. A computer host initiates the process and analyzes the final results. No data from the media needs to traverse the SCSI bus, therefore there is neither bus bandwidth nor host CPU impact.
- Excessive media wear can be determined before the data becomes unreadable. If a host detects media showing excessive wear, the data can be transferred to a new cartridge and the old cartridge identified for removal by an operator.

Costs are minimized:

- Less floor space means reductions in the amount of leased media storage space and related insurance and media transportation costs.
- Incrementaly augmenting the number of tape drives in response to rising system load affordably increases bandwidth. For example, seven RSP-2150 tape drives can support a sustained transfer rate of 14 MB/S and burst rates of up to 28 MB/S for markedly less cost than a single 19mm tape drive. Increasing bandwidth to 16 MB/S with the acquisition of one additional RSP-2150 is significantly less expensive than acquiring another 19 mm tape drive.
- The volume serve architecture is cost effective for a distributed computing environment because it allows sharing of the most expensive component (the robot) while still providing lights-out, operator-free support, minimizing recurring operating costs.

Limited risk:

- Configurations with multiple tape drives and robots distribute risk. Failure in a single component does not shut down the entire system.
- The Metrum components and storage media identified by way of example in this paper are based on standards.

- The Metrum RSP-2150 tape data format has been submitted to ANSI and is going through the standardization process.
- The Metrum RSP-2150 tape drive uses the standard SCSI-1 interface, supported by virtually all computers on the market through direct manufacturer support or through third-party offerings.
- Media wear and aging can by monitored dynamically with software by interrogating Metrum RSP-2150 tape drive registers.

DISADVANTAGES OF THE ARCHITECTURE

- The desired tape cartridge could be in use by another host/drive, resulting in a delay in access of indeterminate duration. It may be possible to minimize the number and frequency of these delays by making multiple copies of data cartridges for which access conflicts occur. The database structure would then have to be extended to support the concept of multiple copies of a single data set.
- The situation may arise when the cartridge containing the data set is available but the host's tape drive in that robot is already busy servicing a data transfer request. This situation may cause an unacceptable access delay. It may be possible to resolve this by any of the methods listed below. Some of these methods may require that additional database information fields be generated before they can be implemented.
    - Additional copies of the cartridge could be placed in other robots. If one tape drive is busy, perhaps at least one of the others may not be, thereby permitting immediate access to the data.
    - An additional tape drive could be added to the robot, space permitting.
    - Implement a data storage architecture capable of passing cartridges from one robot to another automatically.

SUMMARY

This paper has proposed hardware configurations that support the construction of large computer-accessible data archives. These configurations minimize storage costs and data access latencies while they maximize data transfer rates. Simple database constructs and a minimal robot control language have been presented. Commercial off-the-shelf hardware components were identified, by way of example, to demonstrate the feasibility and capability of this architecture.

Computer programs needed to implement this architecture, while not exceedingly complex, are not commercially available at this time. Non-commercial versions with limited functionality are currently under development.