

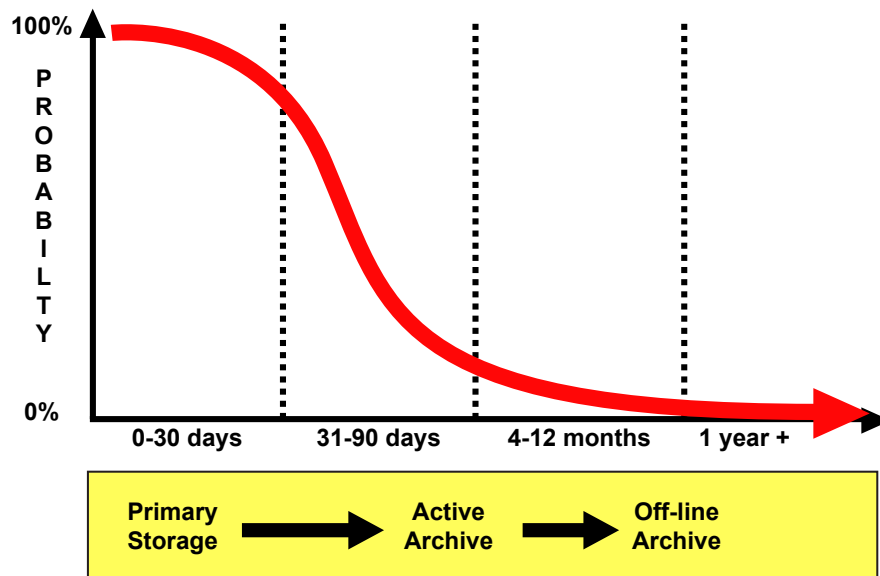
## Performance Considerations for Archiving Data

By: Randy Kerns — Chief Technology Officer, ProStor Systems

Archiving data is a practice utilized in Information Technology for managing data that makes decisions based upon the effect of time on the probability of access. The change in the probability of access can be exploited for cost avoidance in both capital and operational expenses as well as for near term gains. While the benefits of archiving data are realized in economics, the deployment must consider the performance requirements and the basic function of an archiving system.

Performance considerations for archiving involve the characteristics of the archive system as well as the overall nature of what constitutes archiving of data. The differing implementations in archiving systems and their impact on performance only have meaning when the complete process for archiving is taken into consideration.

Archiving of data is not just a one way movement of data even though the basic reason for archiving is that the likelihood of access is very remote. The general probability of access of information curve is represented in *Figure 1*. While this does not characterize all information, it is a very general case and fits a majority of information. Another way that this is represented is that the value of data changes over time where it is less valuable and therefore less likely to be needed.

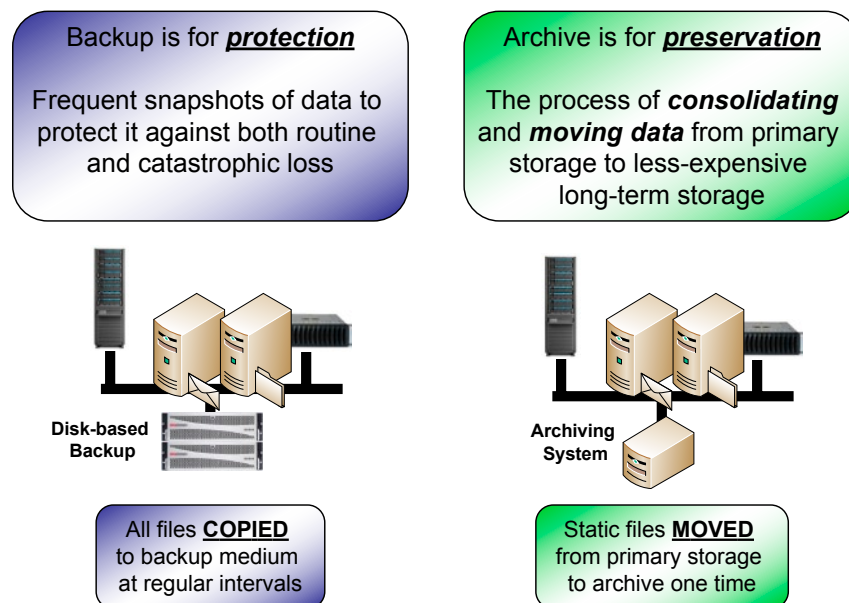


**Figure 1:** *Probability of Access*

## Understanding Archiving

While archiving has been utilized in Information Technology seemingly forever, the definition of what constitutes archiving has been diffused somewhat. To understand the performance considerations of archiving data, the very basis of what archiving is needs to be understood. Quite simply put, archiving is moving data off one storage medium (typically primary disk storage) to an archiving system when the probability of subsequent access has reached a point that the need for access on primary storage does not justify the economic costs.

By archiving the data off of primary storage, it no longer consumes space there and is no longer included in the backup process. By the definition of archiving, this is not a defined “backup process” and the data in the archive is not necessarily in an encapsulated form typically produced by backup software. *Figure 2* illustrates the difference between archive and backup.



**Figure 2:** *Difference Between Archiving and Backup*

Another common misconception or misuse of archiving is for storing fixed content data that may be frequently accessed. In this case, information that does not change or changes infrequently is being stored on a storage system other than primary storage because the frequency of data protection can be reduced due to the unchanging

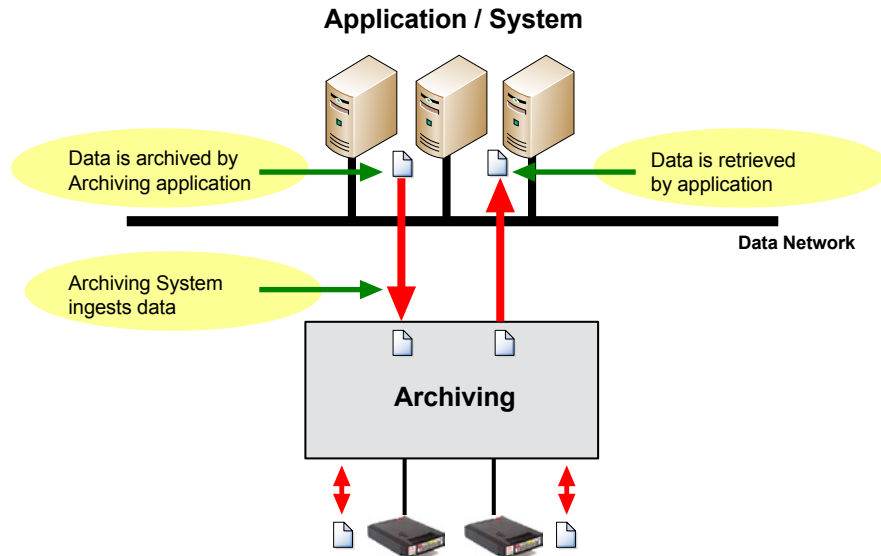
nature of the data. Archiving with the low probability of access and the controls around compliance and retention is not the same as storage of fixed content data and the two different types of information storage may have significantly different performance requirements. Because of the differences in the way information is treated and the different performance requirements, an archiving system may not be the optimal storage system for fixed content data.

## **Performance for Archiving**

Archiving is about moving data which is in the form of files or objects. The fact that archiving is a physical move and not real-time processing of information focuses the performance aspect on throughput. Throughput is measured in the amount of time it takes to move a given amount of data and is usually quoted in megabytes per second or MB/s. This measure is about the capability to move data – how much over time and not the number of input-output operations per second (IOPs) that can be performed. If the archiving system has archive data that is commingled with regularly accessed information for real time processing such as some fixed content data, the performance required for archiving can be unclear and make it difficult to develop a meaningful archiving plan.

Fundamentally, archiving is a write-dominated operation. Data is archived with the premise that it is unlikely to be retrieved again and certainly not on any regular basis. A retrieval is considered an exceptional event. Consequently the write throughput is the most critical performance measure. The scale of the number of files or objects archived is highly dependent on the archiving application and the size or scale of the Information Technology operation. Compared to the normal processing requirements for I/O, archiving is a generally a low volume activity after an initial surge that occurs when an archiving system is first installed.

From the archiving system perspective, data being written to the archive (whether file or object) is called an ingestion of data. The archiving system will manage the data and place it on a storage medium based on a set of rules or parameters. Retrievals of information may require recalls of media for the archiving system. *Figure 3* illustrates an overview of archiving.

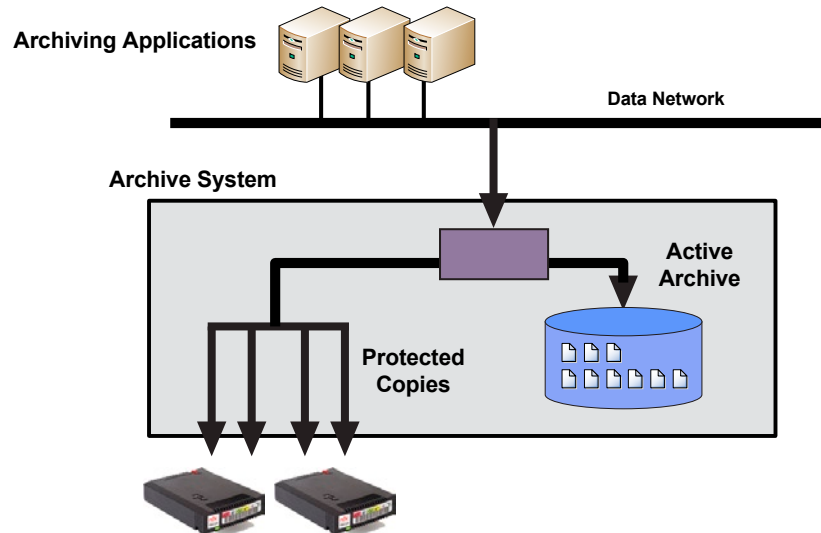


**Figure 3:** *Archiving – Moving Data*

The storage devices used in the archiving system can be a significant factor in the performance as seen by the archiving application software. The degree at which the individual device performance is visible to the archiving application is a system design issue up to a point. Ultimately the archive data must be written to a storage device but the characteristics of archiving with a very high write to read ratio and modest to low ingestion rates can be exploited to favor one technology over another.

An archiving system may utilize different storage device technologies that have different characteristics. In addition, combinations of technologies may be integrated to meet some performance demands. One method of implementing an archiving system is to use an active archive which is a storage area where files or objects may be retrieved directly by applications or archiving software and also may create copies of data on removable media to meet protection needs. *Figure 4* shows a diagram of an active archive along with the removable media copies. The active archive provides another layer of opportunity in the archiving system to meet needs of applications that have archived data: the data in the active archive can be accessed without any internal movement or promotion of data from a removable device and the data is “visible” with security controls to the individual applications. Managing the residency of data in the active archive is usually an automated function based on the capacity and performance requirement settings. In addition to providing access and protection for archive data,

the regulatory compliance factors that include immutability, serialization, encryption, and secure deletion can be implemented in the archive system.



**Figure 4:** *Archiving System with Active Archive*

The devices used in archiving include magnetic tape, optical disks, fixed disks, and removable disks. Considerations of the devices used in archiving regarding performance include:

- *Bandwidth* – the sustained performance in MB/s for moving data with primarily write operations
- *Latency* – the initial access time to store or retrieve individual files or objects
- *Random access* – retrievals typically are for specific files or objects and occur in a seemingly random fashion.

Other characteristics that may or may not have some bearing on performance but are more important for other considerations include: hardware-enforced WORM mode for data immutability, removability of the storage media, encryption and digital fingerprint (content address) of the data element, reliability of the storage media on retrieval, longevity of the storage media without a requirement for forward migration with new versions of device technology, and potential technology obsolescence.

The archiving application and the server where it is running may have impact on the performance for archiving. The archiving system is receiving (ingesting) data that is being moved to it by some data mover software – usually an archiving application. If that application is inefficient or the server that it is executing on has some limitations, the overall archiving process may be impacted. Both of these elements will need to be part of the performance considerations.

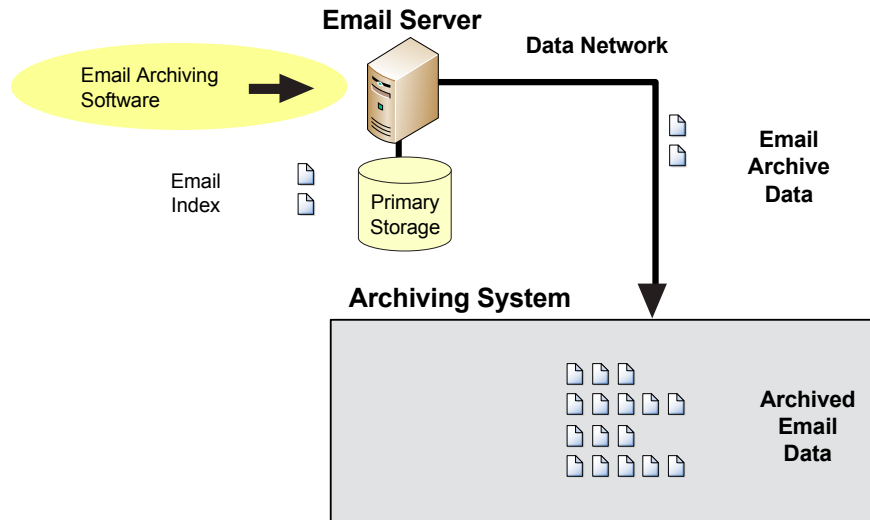
The data transferred to the archiving system is usually sent over a data network. The network bandwidth and its overall efficiency may be a gating item regarding the ability to archive data. There may be a great deal of variability in the data rate of the network given the dynamic nature of traffic that may share the network. Sizing and monitoring the data network is as important a performance consideration as any of the other elements.

## **Archiving Applications**

Applications that do archiving of data fall into a very narrow classification of functionality. The archiving applications are typically sold as independent software and come from a variety of vendors. The performance of each of the different archiving applications can vary widely with the archiving system only being a part of the equation. The processing done by the archiving application, the database that may be used for maintaining metadata, the collection and moving processes and the potential content indexing that is done are all highly variable factors. The most likely archiving applications and their performance implications are:

**Email Archiving** – Email archiving works with existing email systems such as Microsoft Exchange and applies a set of rules around collecting, indexing, and retaining email and attachments. The email archiving software usually does indexing and maintains a database with metadata about the emails. Archiving of email data varies based on the software vendor but usually includes archiving of a group or segment of emails based on time or owner, attachments that may be reduced to a single instance to avoid repetition and retention based on regulations or business practices. Large environments that centralize email in some manner may drive a significant number of emails into the email archiving software. The performance needs for the email archiving software are related to the number of emails that are handled and the policies regarding what information (and when) gets pushed to the archiving system. Some email systems may convert emails to another format for later searching or may create a database that contains content indexes. These are typically compute intensive operations such that the server system running the email archiving software is the first consideration for performance regarding email archiving.

Figure 5 illustrates an email archiving system.



**Figure 5: Email Archiving Operation**

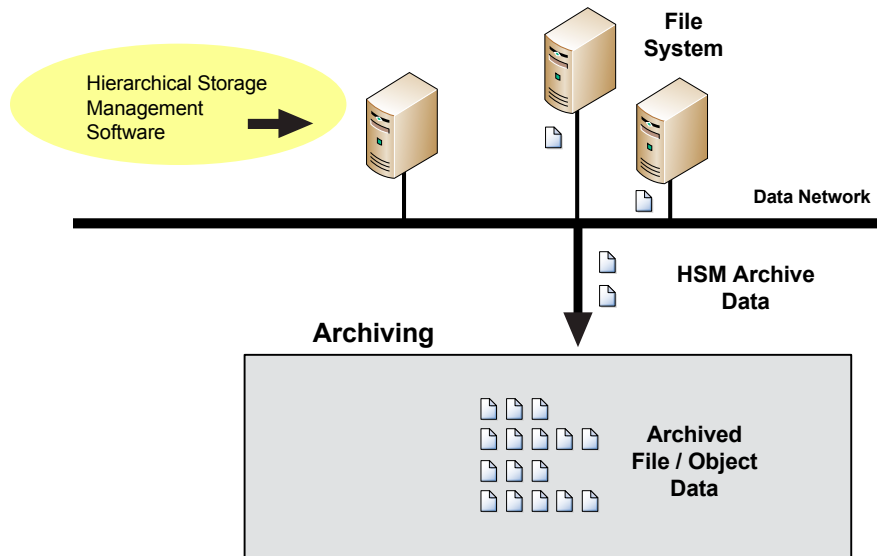
In one email archiving system that was evaluated, a centralized email archiving software application was handling emails for a very large company. In this case more than 250,000 emails were being driven into the archiving application per hour. The emails were processed and containerized (grouped together for storage as a larger block of information) and sent to the archiving system. For this environment the following performance was seen:

- 250,000 email messages per hour
- 100K bytes – average archive element (containerized email and attachments resulting in just under 200,000 elements)
- 5.6 MB/s throughput for ingestion of emails by archiving storage system
- Retrieval rates were insignificant

**Hierarchical Storage Management (HSM)** – HSM software uses rules to move data from one storage level to another and is the primary software tool used for archiving of information. The rules are the key to HSM effectively meeting the business needs for Information Technology. Usually the probably of access profile from *Figure 1* is used

in setting up rules such that the time of last access becomes the trigger for HSM decisions regarding moving of the data from one level of storage to another. Other very specific to business rules can be chosen as well. The most straight-forward of HSM usages is to move files based on the rules. Some of the differing implementations add features to index the data and containerize it. The net effect of containerization from a performance standpoint is to provide a more consistent size range for transfer which yields a more consistent throughput requirement. Direct movement of individual files can be highly variable regarding the transfer length as compared to the overhead for processing which yields a variable demand on the archiving system.

Figure 6 illustrates HSM archiving operation.



**Figure 6:** *HSM Archiving Operation*

An analysis of an HSM archiving software product in a very large system environment where a large amount of data was being archived based on a rule set that was aggressive with the archiving of data (meaning that it was still relatively high on the probability of access curve) was captured and is shown in the following statistics. Because of the aggressive rules for archiving, the retrieval rates were very high as well which is somewhat atypical of most archiving systems.

- 2,000 files per minute (on average) were migrated to the archiving storage system
- 100K bytes for average container size (actually ranged from 70K to 130K)
- 1,000 file retrievals per minute
- 5 MB/s throughput required on the archiving system

**Enterprise Content Management (ECM)** – ECM is similar in concept to HSM but is more complex in that usually unstructured data is being indexed and containerized to put into some management structure. ECM software will manage when and what data is to be moved to an archiving system based upon the management rules. Structured data that is extracted from a database is managed in a similar fashion regarding movement to an archiving system. The ECM software systems are very complex regarding managing information according to business rules and requirements but from an archiving of data standpoint, the software moves data similar to the other archiving software applications. Performance from the standpoint of the archiving system is similar to that of HSM operation.

**Other Archiving Software** – There are other software archiving applications that are in use as well as the three highlighted. One is the simplistic data movement using system tools and scripts or execs for controls. Included in this type are the drag and drop type of graphical interfaces to invoke the system movement.

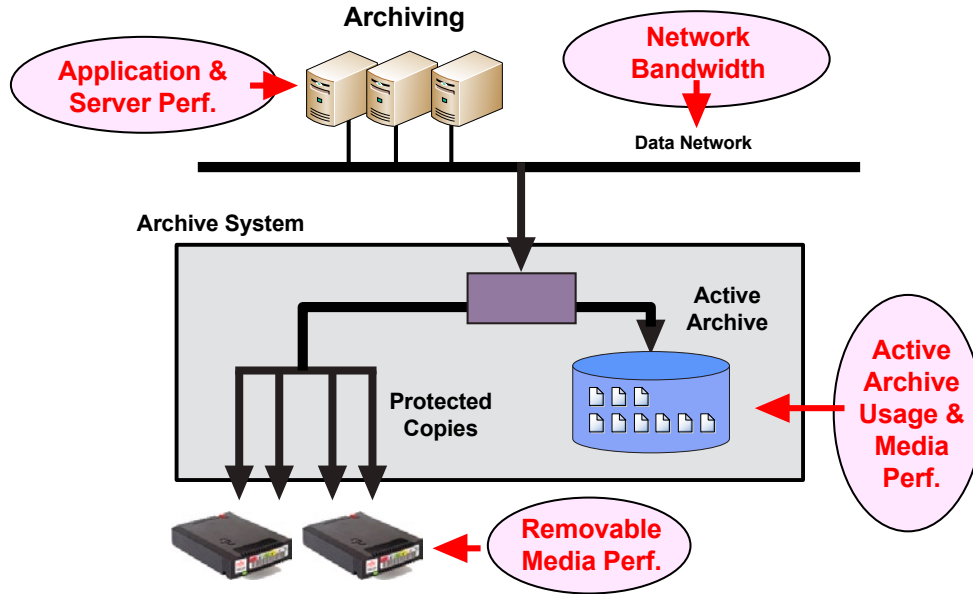
Another type of archiving software seen is the custom implementation that many IT operations have developed to meet their specific needs. The custom archiving software may have different performance impacts and demands on the archiving system so each would need to be understood to effectively plan usage of an archiving system.

Some applications including databases and industry specific software implementation such PACS may have an archive function included that understands the internal structure of data and has some automation regarding archiving controls that is dependent on the application itself. Similar to the custom implementations in IT, each of these integrated applications will have to be understood to be able to characterize the performance needs.

## Overall Look at Performance Considerations

Archiving represents significant opportunity in the economics of storing data. One of the necessary elements in realizing the benefit is sizing the requirements for the archiving system which includes meeting the performance needs both initially and for the long term. As a review of what needs to be considered for performance of an archiving system, the following list encapsulates the majority of items:

- *Needs and usage of archiving* – Generally termed “archiving requirements,” this is really an understanding of the need for archiving and how archiving will be used in a particular environment. Within any company there are many different demands that roll up to a general archiving category but each will have to be considered independently.
- *Amount of data* – While this seems simple, it is one that will be surprising because the successes of an archiving implementation typically result in more data being “made available” for archive than was initially exposed in the planning. With that anomaly accounted for, the capacity growth of data eligible for archive can be estimated and the performance demands taken into account with the capacity growth.
- *Types of data* – Different types of data have characteristics that affect performance requirements. Simple attributes to consider are the size and number of files or objects that will be archived. Extremely large numbers of small files may be equal in capacity to a small number of large files but the overhead in handling or indexing may have very different performance implications.
- *Regulatory compliance* – By the very nature of archiving data, there tends to be a set of regulatory compliance or business governance rules from a myriad of sources that may be applied to the data. Handling compliance may introduce some additional overhead elements or limit the type of archiving system and media storage that must be factored into the performance considerations.
- *Archiving system elements* – once requirements are understood, the characteristics of the physical environment and the archiving system must be analyzed to determine if performance capabilities will meet the requirements. There are many different aspects to consider for the complete archiving environment. Figure 7 highlights the elements of concern for performance in an archiving environment. Those particular areas are pointed out in red with a balloon.



**Figure 7: Performance Areas of Concern for Archiving Environment**

## Summary

Putting together the performance considerations for archiving requires planning on a larger scale. Not only do the performance considerations need to be researched but how to monitor and report on archiving performance will need to be addressed. Performance monitoring is rarely an isolated activity so the other elements about archiving such as economic benefits and regulatory compliance adherence will usually be included in any reporting activity.

It is important to understand that archiving is another one of many Information Technology disciplines and must be incorporated into the overall planning strategy. The effects on capacity planning, capital and operational expense reductions, and legal risk control are major parts of archiving contributions to an overall strategy. Performance planning is a part of that much larger picture.