

H-RAIN: AN ARCHITECTURE FOR FUTURE-PROOFING DIGITAL ARCHIVES

Andres Rodriguez

Dr. Jack Orenstein

Archivas, Inc.

Waltham, MA 02451

Tel: +1-781-890-8353

e-mail: arodriguez@archivas.com, jorenstein@archivas.com

1 Introduction

Traditionally, systems for large-scale data storage have been based on removable media such as tape and, more recently, optical disk (CD, DVD). While the need for increased storage capacity has never been greater, the inadequacies of traditional approaches have never been more apparent. This is especially true for fixed-content data: new government regulations and increasingly competitive market pressures have converged to underscore the importance of finding long-term storage solutions for fixed-content data that offer ready and secure access, easily scale, and are relatively inexpensive.

1.1 Shortcomings of removable media archives

Archives that rely exclusively on removable media share the following shortcomings:

- An archive system that commits physical data to removable media is also captive to the specific hardware system that enables read/write access.. As technology changes, these systems inevitably tend towards obsolescence. It is questionable whether the devices that are used today to read tape or disk will still be available and viable years hence—never mind the availability and viability of the vendor itself!
- As the archive grows, access becomes increasingly cumbersome and time-consuming. Data is not always readily available when it is wanted. Moreover, the administrative overhead that is required to provide timely access is unacceptably high, and—for many organizations—prohibitively expensive.
- Government regulations reflect a rising demand to maintain large amounts of data over long periods of time, and to guarantee their authenticity. Removable data is especially vulnerable to physical mishandling and corruption, both through physical deterioration and outside intervention, whether inadvertent or deliberate.

1.2 An alternative model

In general, when digital data is bound to tape or disk, it ceases to be a digital asset; instead, it becomes simply a physical widget that contains bits, with all the drawbacks previously cited. Long-term mass storage of fixed-content data requires a new type of storage model, where the data's physical location is completely separate from its logical representation. In order to achieve this objective, digital data must be stored in a digital archive that is scalable, reliable, and highly available.

Today's best hope of realizing this model rests in a network—or cluster—of inexpensive servers such as IA-compatible machines that can run a full Linux distribution. This model offers the following advantages:

- Various protection schemes (RAID-5, RAID N+K) safeguard files from multiple, simultaneous points of failure in the network, and guarantee that their data remains continuously available.
- Within the network, the archive system autonomously enforces policies that are associated with the stored files. These policies include retention period, file protection, and content authentication.
- Gateways for standard protocols (HTTP, NFS, SAMBA, CIFS) provide over-the-wire access to the archive.
- The archive is easily extended: as new nodes enter the cluster, the archive automatically invokes its own load-balancing and protection policies, and redistributes existing storage into the new space accordingly.
- A network-based archive can facilitate updates to files so they stay current with the latest applications—for example, format changes that are required by new end-user applications. Data migration of this type, on the scale required for large archives, is virtually impossible to achieve in tape-based systems.
- The data's actual location on the network is transparent to the user. During its lifetime in an archive, a stored file might be relocated across many network machines—or nodes—as the result of hardware upgrades, replacements, or load balancing. The reference to the file, however, remains constant, enabling users ready access to its contents without requiring knowledge of its physical location within the cluster.

1.3 Two architectures for online archives: RAIN and H-RAIN

In the last several years, various vendors have come forward with archive systems that implement the network approach just described. These all embody various implementations of RAIN (redundant array of independent nodes) architecture. RAIN archives are based on one or more clusters of networked server nodes. As nodes enter or leave the cluster, the cluster automatically adjusts by redistributing and, when necessary, replicating the stored files.

Currently, RAIN archives are typically delivered as proprietary hardware appliances, closed systems that are built from identical components. Evolution of these systems is carefully controlled by the vendor.

The architecture of H-RAIN—heterogeneous redundant arrays of independent nodes—differs from the RAIN architecture from which it evolved by making minimal assumptions about the archive's underlying hardware and software. In practice, this means that H-RAIN architecture can be implemented with commodity hardware. This relatively open architecture has two advantages over its RAIN progenitor:

- It adapts more readily to technological advances and site-specific contingencies. Administrators are free to replace components with superior hardware as it becomes available, thus improving storage capacity, performance, and reliability. Furthermore, they can choose among hardware options that best suit their requirements, such as CPU, memory, and disk capacity. For example, a cluster might be extended by adding new nodes with higher-performance CPUs, which can be used for CPU-intensive filtering operations. Incremental hardware additions and improvements might thereby measurably improve overall archive performance.
- Archive administrators can start small and scale up capacity incrementally simply by adding nodes as they are needed. Moreover, they are free to seek the best prices for storage cluster components. Given that component costs tend to decrease over time, cost-conscious administrators can reduce their average cost per gigabyte by spreading out purchases.

In general, an H-RAIN architecture enables users to upgrade their technical infrastructure while transparently migrating archive content to more up-to-date nodes. Improvements can be made incrementally, leaving the initial installation intact. As hardware prices fall, archive performance can be enhanced with better-performing nodes, and at lower cost.

2 Implementing H-RAIN architecture

Archivas' archive management system, Reference Information System (RIS), is based on the H-RAIN model. With RIS, organizations can create large-scale permanent storage for fixed content information such as satellite images, diagnostic images, check images, voice recording, video, and documents, with minimal administrative overhead.

In RIS' H-RAIN implementation, two features are salient:

- Distributed processing
- Autonomous management

2.1 Distributed processing

All nodes in a cluster are peers, each capable of running any or all of services that an archive requires. A cluster can be configured so archive services are distributed in a way that best serves the enterprise's storage requirements.

For example, a cluster can be configured symmetrically—that is to say, each node runs the same processes and daemons, including a portal server, metadata manager, policy manager, request manager, and storage manager. Each node bears equal responsibilities for processing requests, storing data, and sustaining the archive's overall health. No single node becomes a bottleneck: all nodes are equally capable of handling requests such as put and get operations. Furthermore, in the event of node failure, any other node can take over responsibility for the data that was managed by the failed node, so that user access to this data remains unaffected.

Alternatively, a cluster might be configured so that various services are distributed asymmetrically across different nodes. For example, if read requests are especially heavy

for a given archive, several nodes might be dedicated solely to request management and run multiple request managers and metadata managers, in order to maximize throughput to other nodes that store the physical data.

2.2 Autonomous management

Through policies that are associated with archived files individually, and the cluster collectively, the archive can manage itself without human intervention. Policies are set for the archive on initial configuration, and can (optionally) be set for individual files as they are archived. Taken together, these policies determine the archive's day-to-day operation. Through a policy manager that executes on each node, the archive monitors its own compliance with current policies, and when lapses occur, takes the appropriate corrective action.

For example, in the event of a failed disk or node, the system determines what data is missing and how best to restore it from data on the remaining healthy nodes, so that the protection policy for these files is fully enforced. Similarly, the system prohibits removal of an archived file before its retention period has elapsed.

Human intervention is rarely warranted, and usually only in response to system warnings that require outside action—for example, notification the cluster load has crossed a specified threshold, requiring the addition of new nodes.

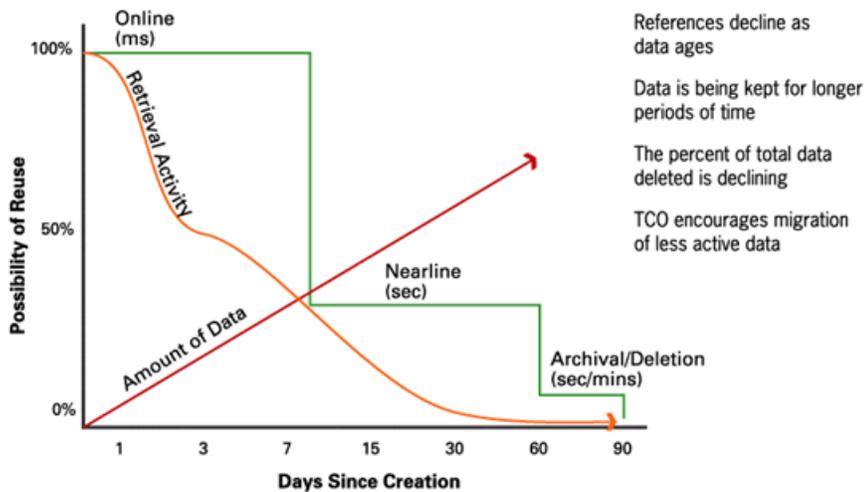
Four attributes characterize archive self-management:

- **Self-configuring:** Setting up large archive systems is error prone. An archive comprises networks, operating systems, storage management systems, and, in the case of RIS, databases and web servers; getting all these components to run together requires teams of experts with a myriad of skills. An autonomous system simplifies installation and integration by setting system configuration through high-level policies.
- **Self-protecting:** Policies that enforce document retention, content authentication, and file protection combine to protect an archive from loss of valuable digital assets.
- **Self-healing:** Serious problems with large-scale archives can sometimes take weeks to diagnose and fix manually. When the faulty device is finally identified, administrators must be able to remove and replace it without interrupting ongoing service. Autonomous systems can automatically detect software and hardware malfunctions in a node, and safely detach it from the archive. Further, because data is replicated across many nodes in the cluster, the failure of one or more nodes has no impact on data availability. Archivas' distributed metadata manager can find an alternative source for any data that resides on a failed node.
- **Self-optimizing:** Storage systems, databases, web servers and operating systems all have a wide range of tunable parameters that enable administrator to optimize performance. An autonomous system can automatically perform functions such as load balancing as it monitors its own operation.

3 Extending the H-RAIN model

With its H-RAIN architecture, RIS is capable of integrating with storage systems that use removable media such as tape or optical disk. In this scenario, the tape system is seen by RIS as simply another set of storage nodes; the physical location of data is managed by an RIS storage manager implementation that is specifically targeted to tape-based storage.

This capability is critical for a multi-tier storage and migration strategy, where data is stored in whatever medium best serves external access requirements. For example, a file that is frequently accessed should be archived in primary storage on a high-performance disk, while data that is rarely used can be stored on relatively low-performance media such as tape. Further, it is likely that access requirements for a given file will not remain constant, especially if the file is retained for a long period of time. In general, references to most types of data significantly decline as the data itself ages, as shown in the following figure:¹



With the advent of government regulations such as the Sarbanes-Oxley Act, enterprises are required to archive increasing amounts of reference data, and retain them for ever longer periods of time. In order to keep storage costs down, it is increasingly important that archive systems respond to changing access requirements by moving data easily from more expensive disk-based media to less-expensive removable media. By encompassing both disk-based and tape-based storage and providing a unified interface to both, RIS can provide a smooth migration path for aging data. Furthermore, RIS policy managers can automatically manage the migration, as determined by archive-wide or file-specific policies.

4 Conclusion

A digital archive system that is based on H-RAIN architecture offers the most economical, scalable, and effective solution for large-scale storage of reference data. Policy-based management minimizes administrative overhead while it provides the most reliable way to

achieve an archive's most important requirements: availability, reliability, and content authentication.

By extending the H-RAIN model to encompass any network node, including those that interface to tape-based systems, the potential exists to implement a multi-tier system where data is stored on the medium that best suits access requirements, and can easily be migrated to another medium as those requirements change.

¹ Fred Moore, "Information Lifecycle Management," Horison Information Strategies (<http://www.horison.com>)