Data Mobility Group

www.datamobilitygroup.com

HP Reference Information Storage System

Architecture Overview

White Paper

May 2004

 76 Northeastern Blvd. Suite 29A Nashua, NH 03062 Phone: 603.881.9031 Fax: 603.881.9035

E-mail is a mission-critical application for all types of businesses—both large and small. But as e-mail volumes grow, it becomes increasingly difficult to keep these applications up-and-running 24x7—a necessity in today's business climate.

Trying to control the e-mail situation, administrators often place limits on the size of user mailboxes. However, doing so often turns out to be more counterproductive than not. Users can spend hours each week filtering through e-mails trying to weed out those that do not need to be saved but still end up exceeding posted limits. And when they do, e-mail administrators are again called into action—often away from higher-priority tasks—to address the problem.

The e-mail burden on the IT department is further complicated by an increasing number of regulatory compliance requirements centered on corporate records retention, many of which dictate how e-mails should be saved and archived for potential future access.

Thanks to its acquisition of Persist Technologies last November, Hewlett-Packard now has a way of effectively dealing with the growing e-mail problem. In fact, its new product—Reference Information Storage System (RISS)—is not limited to the storage of e-mails but can also be used to store other content, including presentations, Word documents, medical images, and other unstructured data types.

The Problem with E-mail

As all users of e-mail can attest, maintaining e-mail in-boxes is no simple task. It requires fastidious "purging" of in-box messages, back-and-forth copying of critical e-mails from desktop to laptop (and vice versa) in the hope that critical information is backed up, and the help of e-mail administrators to lift size limitations, find files, etc.

As for e-mail administrators, they fare no better. They are charged with the task of managing e-mail systems that appear to be growing without bounds. What starts off as a single copy of a PowerPoint presentation can quickly become 20 copies or more. A user e-mails the presentation to 10 co-workers, who three days later distribute it to another 10 people. Not only is the original e-mail message saved multiple times but so are any attachments.

As e-mail servers and storage get larger and larger, the challenge before administrators intensifies. Backups can take hours, restores even longer. Locating a specific e-mail message that has been backed up or archived can take days, if it is located at all.

What is required is a solution that reduces the management complexity, minimizes storage requirements, and makes it quick and easy to retrieve any e-mail.

Enter RISS

RISS was designed to take the large problem of trying to manage growing amounts of e-mails and break it into smaller, more manageable pieces.

Based on a grid-computing architecture, RISS is composed of many storage "smart cells." Each cell is self-contained; has its own processor, storage, and content indexing; and is mirrored to another storage cell for redundancy.

Storage smart cells are federated to form a storage smart cell fabric (see Figure 1), which is highly scalable. Adding a storage smart cell not only increases the grid's storage capacity but its processing power and content indexing capabilities. This allows the RISS to scale exponentially without sacrificing processing power.

Storage smart cells are responsible for storing, indexing, and protecting the data. When additional storage is needed, storage smart cells can be added non-disruptively to the RISS infrastructure. The cells are automatically discovered and added to the available storage pool. Operational Smart Cells manage the storage system and enforce policies.

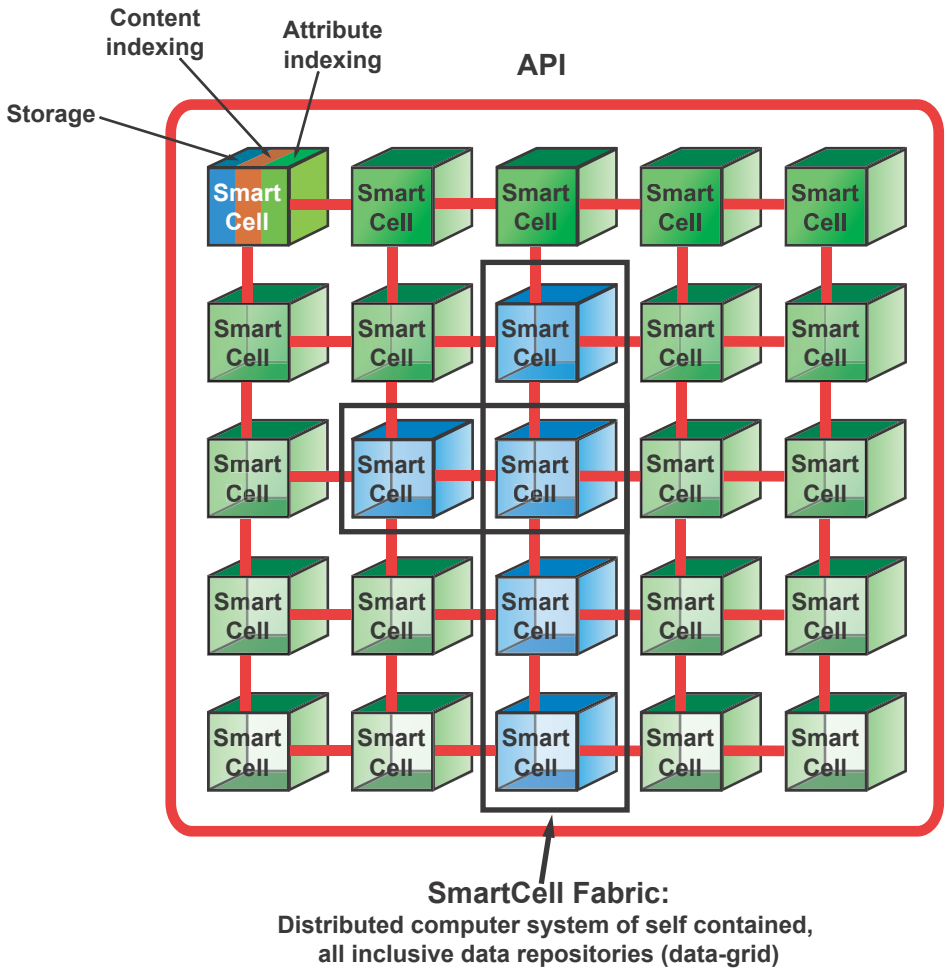

Figure 1

Domains and Repositories

RISS can be broken into domains or several smaller entities for management purposes (see Figure 2). For example, an administrator may decide to create separates domains—each with its own policies—for divisions within a large organization. In particular, domains may be useful in service bureau environments or to segregate regulated and non-regulated documents and e-mails.

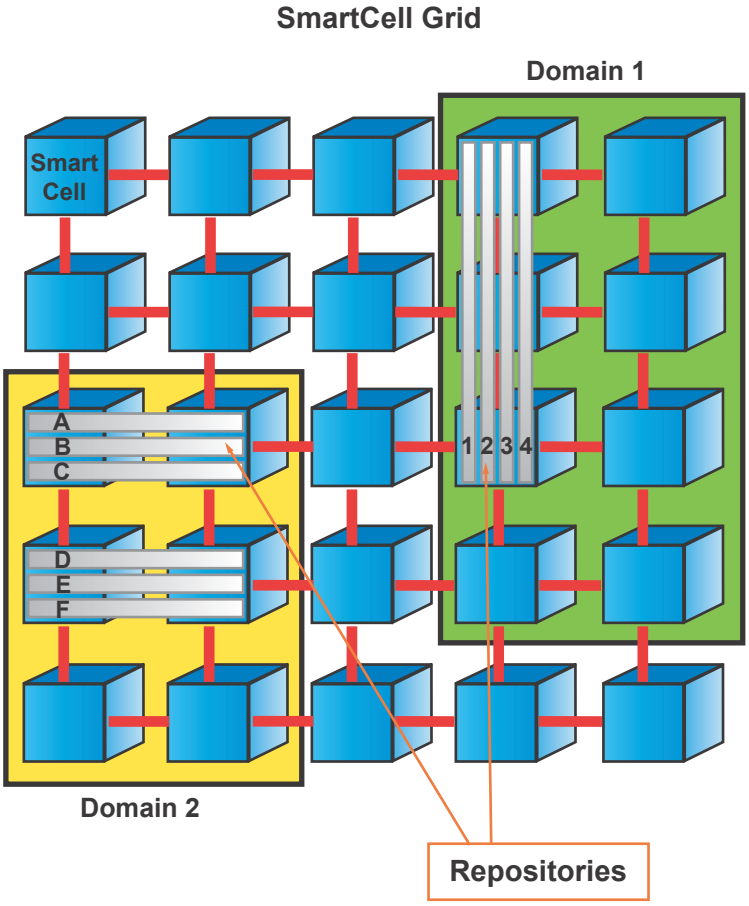

Figure 2

Each domain contains at least two storage smart cells, which are physically partitioned from the remaining cells. Each domain has its own backup policies and authorization settings, and can be managed by its own administrator.

Within each domain are one or more repositories. Each repository is a logical entity that spans more than one storage smart cell. A repository can be thought of as one user's mailbox; the domain would contain multiple user mailboxes. Access Control Lists determine who has access to a particular domain or a repository.

RISS and Microsoft Exchange

RISS is well suited to store both regulated and unregulated messages for several years. In a regulated environment, e-mail messages are immediately directed to the storage system; in an unregulated environment, the system periodically "pulls" or "harvests" messages from the Exchange environment.

Unregulated Environment

In unregulated environments, messages are delivered to the Exchange server in a traditional fashion. At user specified intervals, the RISS pulls or harvests the data from user mailboxes. Pre-set policies determine what data to harvest and when to harvest it.

For example, a typical policy might tell the system to harvest all e-mails after seven days. In this selective archiving mode, RISS becomes an extension of the Exchange server and offloads all e-mails that are eight days old or more, thereby reducing the size of the Exchange server. E-mails can also be selectively harvested based on other criteria (e.g., sender, recipient, subject, etc.).

All harvested messages (and their attachments) are stored on the HP RISS and are deleted from the Exchange server. Users see few changes: The message header appears in the user's mailbox; a small tombstone icon tells the user that the data has been successfully saved to the RISS. To retrieve messages, users simply double-click on the icon. The messages are retrieved from RISS and sent to the Outlook client, not to the Exchange server.

Users do not have to actively manage the size of their inboxes. Messages are automatically routed to RISS after the mining operation is complete. This option simplifies the management of e-mails and reduces the storage requirements of the Exchange server.

In regulated environments, additional support is required. New compliance regulations, such as SEC 17a-4, Sarbanes-Oxley and HIPAA, require documents to be retained in an unmodified form for several years. For example, e-mails detailing financial trades must be reviewed by supervisors, and they must be retained for extended periods and made quickly accessible in audits or discovery processes. Failure to do so can result in hefty fines, negative publicity, etc.

RISS supports this regulated environment through flexible storage rules and fast searches. Regulated e-mail messages are automatically delivered to RISS before they appear in the user mailbox, and each message is signed with a digital signature to verify its integrity. Retention policies, which dictate the period of time that an e-mail must be saved, can be lengthened but not shortened. Supervisors can review e-mails via a Web-based GUI.

Storage rules determine which domain and repository particular e-mails are sent to. For example, if Bill is a regulated user (and Bill's supervisor, John, is required to review his e-mails), the storage rules would specify that Bill's e-mails be sent to both Bill's repository (i.e., Domain 1, Repository 1) and John's repository (i.e., Domain 1, Repository 2).

In this example, John can view of all of Bill's e-mail while Bill can only view his own. However, only one copy of the e-mail is actually stored in the RISS. The company adheres to storage policies (and regulatory rules) without having to allocate additional storage capacity.

Recognizing that not all e-mails need to be regulated, HP has designed RISS to support both regulated and unregulated content within the same storage system. In fact, the system is designed to support changes in regulatory status on the fly—a feature that is particularly important as workers' responsibilities change. Today, John is an unregulated user, but tomorrow he could be promoted and become a regulated user.

Elimination of Duplicate Messages

Storing duplicate messages can be costly. Saving multiple copies of a single 3 MB PowerPoint presentation, for example, can unnecessarily chew up many MBs or more of valuable storage capacity. Cost-effective e-mail archiving solutions must eliminate or reduce redundancy.

RISS detects duplicate messages through two different processes: single instancing and duplicate filtering. Single instancing recognizes when e-mails have been sent to multiple recipients by searching through e-mail identifiers, such as sender, recipient, etc.

The original message is routed to all individuals on the distribution list but is only stored once. Though these individuals appear to have their own copies of the original e-mail (and any attachments) in their own repositories, what they really have is a "view" to the original stored message.

Duplicate filtering, meanwhile, detects and filters e-mails that are sent to/from separate Exchange servers; it works with Exchange server journals. While Exchange does a good job of reducing duplicate messages when they reside in a single Exchange server, when the recipients span several Exchange environments, it falls short. This is where RISS comes in. The RISS detects these duplicates by creating a message hash for each message during the harvesting process. Incoming messages are compared to the hash and duplicate messages are filtered out.

Message Retrieval

While storing messages is important, being able to quickly retrieve them is paramount. Messages can be retrieved by many search criteria, including the message date, recipient, sender, subject, or a text string embedded within the message.

RISS does not have to query all storage smart cells to locate all matching e-mail records. The secret to the storage systems' high-speed searching capability is that it knows which subset of storage smart cells is likely to contain a particular message.

For example, the device knows that all messages stored within the first six months of 2003 are contained in the first x storage cells; therefore, it searches only these cells—and it does so in parallel. Because the system does not have to query all storage smart cells to retrieve messages, the search time for queries is reduced significantly

Once RISS finds the lists of messages that match the specific criteria, the message are presented to the user, who decides which messages need to be retrieved and which ones do not. These messages are then shipped to the requestor inbox or PST.

Data Protection

RISS provides several levels of data protection. First, each storage smart cell is mirrored to another smart cell within the same repository. RISS writes to a storage smart cell and its mirror synchronously and verifies both writes through CRC checks before the write operation is completed.

Mirroring within RISS protects against the failure of a storage smart cell, but it does not protect data from a failure at the primary data center. For disaster-recovery purposes, a primary RISS can replicate remotely to a secondary RISS system over an IP, T3, OC3/12, or DC3 connection. Replication occurs at the domain level and is bi-directional.

For example, Domain 1 in primary RISS 1 can be replicated to remote RISS 2, while Domain 2 in RISS 2 can be replicated to RISS 1 (see Figure 3). If RISS 1 fails, or there is an outage at the local data center, the Exchange application fails over to RISS 2, which now performs the stores and retrievals.

When the local data center comes back on-line, fail-back operations begin. The contents of the original domain—and any new e-mails stored during the outage—are replicated back to RISS 1. RISS 1 then resumes its original role of storing and retrieving e-mails.

For users who do not have a second data center, HP has designed RISS so that they can back up data and digital signatures to Write-Once Read-Many (WORM) media (e.g., WORM tape or HP Magneto-Optical disk) The target device can be local or connected through an IP WAN network.

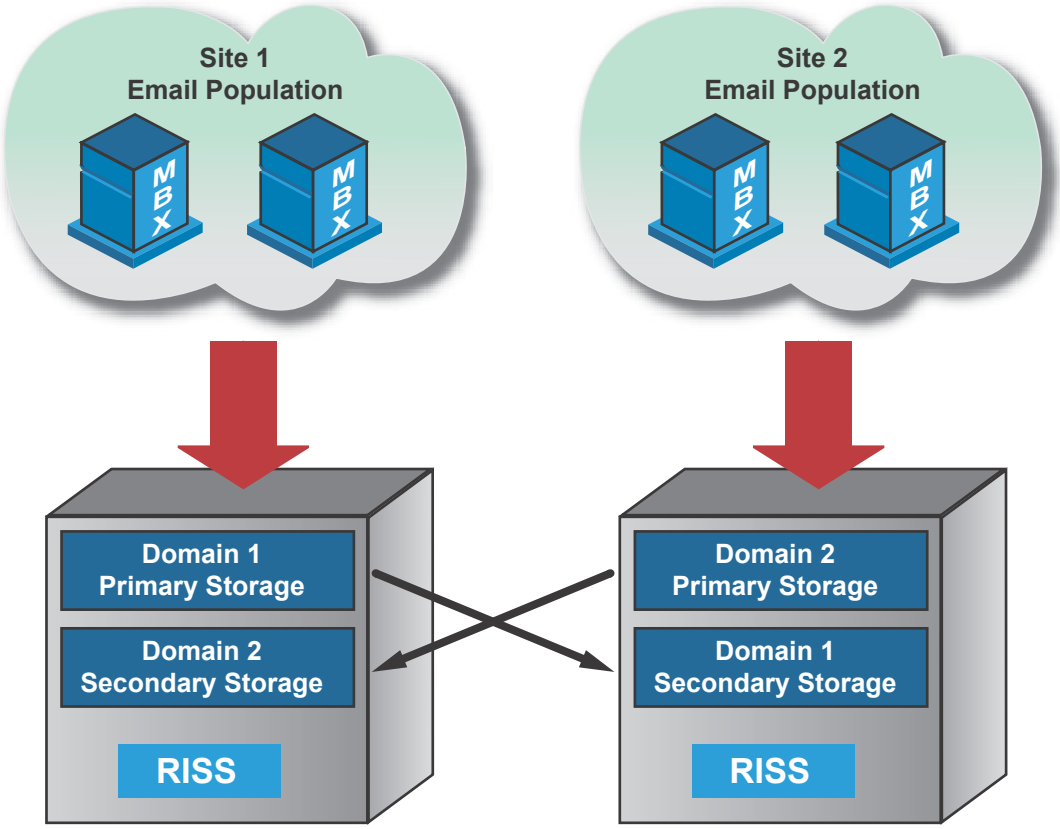

Figure 3

Data Integrity

Regulated e-mail must be stored in its original unmodified format. RISS maintains data integrity in several ways. First, all data written to a storage smart cell is also synchronously mirrored to a second storage smart cell within the same domain. A CRC check is performed on the data before the operation is considered complete.

RISS also creates a digital signature for each stored message. This digital signature consists of two parts—the hash of the original message and a time stamp that tells when the message arrived at the system. The hash and time stamp are then encrypted using 128-bit DES encryption. A private key is used to create the signature and then discarded; a public key is retained. The digital signature and message is then passed to the storage smart cell. The data is indexed and compressed and the data, index, and signature are stored on both the primary storage smart cell and its mirror.

When a request to retrieve an e-mail is received, only storage smart cells that are thought to contain the e-mail are queried. The e-mail and its digital signature is returned, the digital signature is recalculated from the e-mail and compared to the original digital signature for validation purposes, and, if the signatures match, the e-mails are returned to the person making the request.

Security

Storage devices that are used to archive regulated data need to be secure. RISS has many built-in physical and logical security mechanisms. For example, the RISS can be broken into domains to physically isolate divisions within a company or one company from another. In these cases, divisions or individual companies are assigned their own domains; access to and from these domains is governed by Access Control Lists.

Communications within RISS take place within two redundant private IP subnets, which are not exposed to the outside environment. To prevent unauthorized access, RISS uses Network Address Translation (NAT) technology to translate IP network addresses so that the receiver of a transmission does not know what the actual IP address is.

Access is also restricted by a built-in firewall; each RISS system has at least two firewall/NAT blades for redundancy. All external HTTP access is protected by Secure Socket Layer (SSL).

RISS—An Integrated Solution

Selecting and implementing the right storage systems to support the retention of documents and e-mail for long-term compliance is difficult. The right solution must provide:

- physical and logical security
- data integrity through technology such as digital signatures
- several different layers of data protection to prevent data loss during disaster recovery
- fast store and search capabilities
- a scalable architecture for future data growth

RISS provides all those features in one integrated system. Unlike other solutions, which may require users to purchase separate components (e.g., e-mail extractor software, disk storage, and firewalls), RISS integrates all of those features under one cover, eliminating interoperability and bottleneck problems.

visit us on the web
www.datamobilitygroup.com

5982-5979EN, 5/2004