Conquering the Scalability and TCO Challenges of Managing Email

Turning Data Into Information

Executive summary

This white paper focuses on the business case for the use of the HP StorageWorks Reference Information Storage System as a solution for dramatically reducing the total cost of management of email systems.

Reference Information Storage System is an appliance designed to provide large-scale storage, access, and content-based retrieval for reference information. Reference information is defined as records that are not expected to change. Valuable information exists within this pool of reference data—whether as a record of an event, a piece of collateral that is potentially re-usable, or documenting the details of a project or product.

Email, and its associated attachments, are prime examples, and clearly ubiquitous across all businesses. Reference Information Storage System will automatically harvest emails from primary storage systems and place them on Smart Storage Cells. This storage pool is designed as a grid-architecture that combines processing power, indexing, and storage capacity—turning data into information.

Finally, Reference Information Storage System can aid businesses to meet the growing number of data retention regulations by providing long-term storage of specific business records in a secure and tamper-proof manner.

Target audience

CEOs/CIOs and business managers who are asking the following questions:

- How can we reduce the cost of our email systems?
- How can we keep pace with explosive email growth, without exceeding our IT budget?
- How can we improve the performance of our email system?
- How can we reduce the time employees spend on administrative tasks related to their email, such as inbox quota management or searching for a specific message or attachment?
- How do we reduce the waste in our email system? We know there are several duplicated messages as well as stale items, but we still store and back up all of them.
- How can we address the growing number of data retention regulations and how are we going to manage to retain records for long periods while sticking to our IT budget?

Introduction

More than 50% of data stored on enterprise storage systems is reference information—records that are not expected to change. Examples of reference information are email, digital audio and video files, sales collateral, product/project documentation, and also vertical industry-specific items such as check images for finance and patient medical records for healthcare.

As email is so ubiquitous in today's business world, this white paper focuses on how Reference Information Storage System dramatically improves the total cost of ownership (TCO) of email—although the same arguments can be made for other reference data types, and Reference Information Storage System can support many other file types.

The email problem

The core problem presented by email data is that it is poorly managed. Email represents such a large volume of data that poor management has a significant impact on the performance and cost of email systems. Furthermore, email is highly duplicative in its nature, so a lot of email data that is stored is actually waste. The emergence of data retention regulations, which specifically list email, cause further complexity.

Performance and cost

Initially conceived as a straightforward routing function—transferring messages from one user to another—email systems have morphed into a vast storage and performance headache. The volumes of data transferred have grown exponentially—through increased users, more messages per user, and increased size of messages and attachments. Coinciding with this growth, users have adopted the mailbox as their default filing system. It is common for messages, appointments, and working documents—reports, slide-sets, and spreadsheets—all to be stored and neatly filed, in many cases, in a user's mailbox.

This is the core problem: Email systems are getting choked and cannot cope with such large volumes of data. More servers and storage are added at ever-increasing frequencies, and IT budgets simply cannot cope.

Scale

A large enterprise could easily have 50 TB of email data, but the critical fact is 50 TB can represent over one billion individual messages. It is the scale that causes the problem—how to effectively manage such a large pool of records and understand which must be kept and which should be removed.

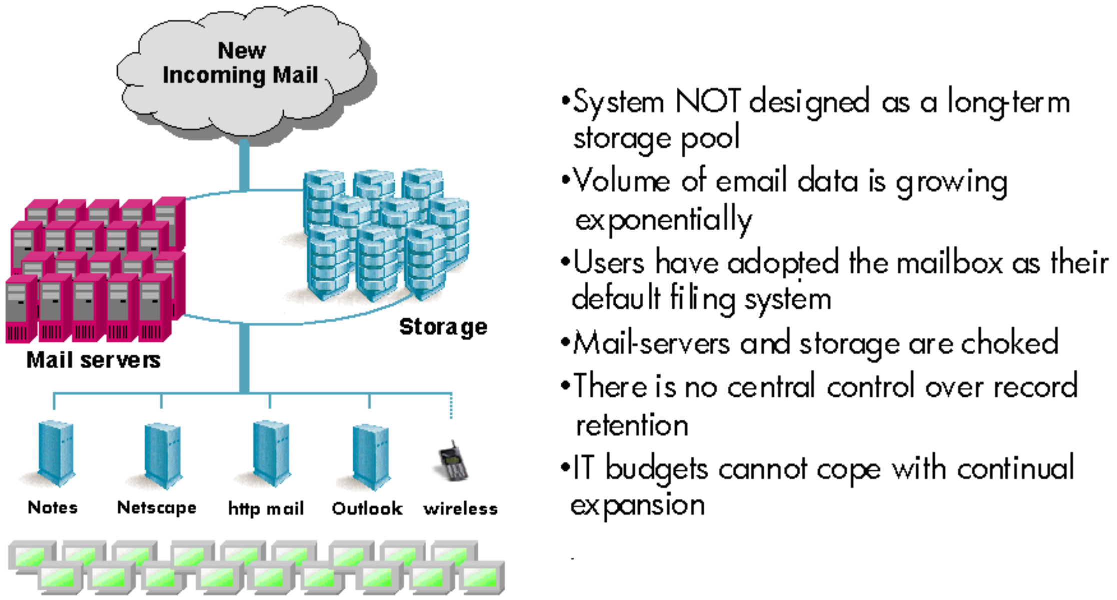
IT departments often try to shift this management burden onto individual users by imposing mailbox quotas, forcing users to either delete messages or archive to personal folders. But by making this shift, IT actually surrenders any control it had, and the problem will escalate. This is because users will manage their mailboxes in an overall random fashion—some will delete items they actually need to keep, others will keep too much. Archiving to personal folders, after deleting the original message from the email server, only pushes the storage/performance problem to another location—and actually can escalate the problem if multiple users archive the same attachments. And in reality, IT will still be called in when there is a problem but may try to recover a message that no longer exists on the backup tape of the email server.

Data retention requirements

Finally the emergence of data retention requirements is now causing an additional email storage problem. Acknowledging that IT budgets are already strained by current levels of email use, the challenge is that businesses may be forced to retain certain messages for increasingly longer periods of time.

As email forms a record of a past events, good business practice, as well as an increasing number of external regulations, require that this record is retained. Audit trails and the ability to defend a dispute are good internal reasons for retaining email communications. External regulations such as Sarbanes Oxley and SEC 17a-4 make specific requirements for email retention for businesses in specific regions and industries.

Figure 1. The email problem: mail systems are not designed for volume storage



- System NOT designed as a long-term storage pool
- Volume of email data is growing exponentially
- Users have adopted the mailbox as their default filing system
- Mail-servers and storage are choked
- There is no central control over record retention
- IT budgets cannot cope with continual expansion

Solving the email problem

The key things that must be understood to address the problem of more effectively managing email data or any other reference data are:

- It is not simply a question of managing storage—it is all about managing billions of unique records.
- Archiving is not about putting data away—it is about retrieving records when they are needed.
- IT cannot rely on users to retain the "right" records—enterprise-wide policies around electronic record retention are needed.

Managing billions of records

Providing increasingly lower cost storage for reference data will not significantly reduce the overall TCO. The fact is that storage alone only counts for one part of the overall cost. As outlined in the previous section, 50 TB of email data actually comprise more than one billion individual email messages. The challenge is not to manage a 50 TB block of storage, where using low-cost disk arrays could reduce the cost; the challenge is actually to cope with more than one billion individual records and transform the 50 TB of data into one billion pieces of manageable information. This requires integration with the original application to effectively capture the records and an indexing capability to be able to understand each record.

Archiving is all about retrieval

Harvesting emails from primary systems to a lower-cost archive pool is only realistic if users are guaranteed they can retrieve their messages just as quickly as with their current system. Again, it is a scale issue—users will be looking for one specific message or attachment from a pool of billions of records—like looking for a needle in a haystack. Therefore, comprehensive search and retrieval tools are needed, along with the processing power to rapidly return a search.

Enterprise-wide archiving policies are needed

If the organization is going to effectively address data retention regulations, and also unlock the intellectual capital that resides in emails and their attachments, then there must be some centrally enforced policies around which emails are retained, for how long, and when are they eventually deleted.

The solution components

The key functionality needed to effectively archive email and solve the core email cost and management problems are:

- Capture
- Index
- Store
- Search
- Retrieve

For many years it has been possible to buy a component-level solution and integrate the preceding capabilities, but with the Reference Information Storage System, HP is delivering an all-in-one solution.

Reference Information Storage System

HP StorageWorks Reference Information Storage System (RISS) is an all-in-one appliance designed to provide large-scale storage, access, and content-based retrieval for email and other types of reference data. All the necessary storage hardware, archival software, and operating infrastructure are included.

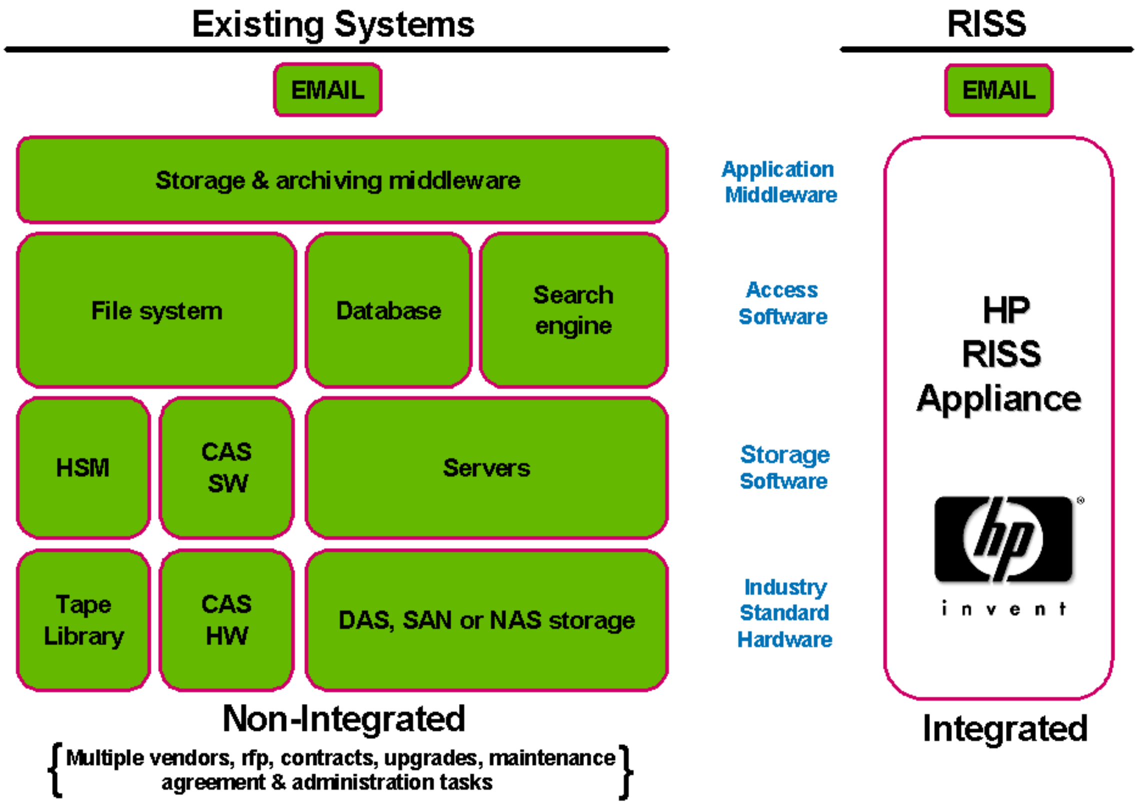
RISS delivers on the key requirements identified in the previous section by providing the following three core values:

- Automatically harvest data from primary systems—**manage** system costs
- Full content indexing and grid computing architecture for rapid search and retrieval—**exploit** your information assets
- **Mitigate** risk with time stamping and authentication to help settle disputes, maintain audit trails, and address data retention regulations

As opposed to other solutions that take weeks to install due to complicated integration tasks, Reference Information Storage System's solution takes generally only a few hours. It seamlessly integrates with many leading messaging applications, and allows many other file types to be archived. For details on which file types are supported, refer to the latest support matrix on the Web at www.hp.com/go/ILM.

Figure 2 shows the complexity involved in a "do-it-yourself" solution compared to RISS.

Figure 2. Cost, integration, and compliance



Reference Information Storage System features and benefits

The key features of RISS are as follows:

Manage costs

- Selective archiving to reduce overburden on primary system
- Automatic harvesting of messages and attachments from users' mailboxes based on central policies
- Archiving policies can be set company wide or at the department or even individual level
- Migration of legacy files to archive
- Automatic removal of duplicates from the archive
- Seamless integration, no user-training needed

Exploit reference data

- Application-aware archiving
- Powerful application-integrated search tools
- Full content search—search entire message and attachment

Mitigate risks

- Automatic archiving of **all** records for compliance purposes
- Data is digitally signed, time stamped, and protected from erasure
- Retention policy management and shredding after expiration

When installed, end users see an additional folder named "Active Archive" in their normal email client view; for end users, this is the only visible change to their email experience. There are two methods for archiving emails and attachments—automatic archiving or selective archiving.

- Automatic archiving enables **all** messages to be archived to the Reference Information Storage System (this helps meet data retention requirements).
- Selective archiving—Policies to harvest out of inboxes and onto the Reference Information Storage System. This delivers a *mailbox extension* that gives users a mailbox that is scalable to a large level.

By using HP StorageWorks Reference Information Storage System, the IT administrator can set policies—either globally on a departmental or even individual basis for automatically archiving emails and their attachments. This removes the randomness of user-initiated archiving.

The Reference Information Storage System can be used to store additional reference data as well such as office documents, voice recordings, video clips, instant messages, and so on, freeing up expensive primary storage. Desktop documents are dragged and dropped into the Reference Information Storage System "My Active Archive" bin, and are then transferred to the Reference Information Storage System where they are content-indexed, compressed, and stored.

Archiving costs are reduced by duplicate-detection. Reference Information Storage System will only store one copy of any given email or attachment, even if multiple users have asked to archive it. For example, if a company-wide email with an organizational slide-set is sent out—the message and slide-set will only be stored once on Reference Information Storage System, but all users will have a pointer to that one record.

After emails and documents are stored in the Reference Information Storage System, they can be queried and retrieved rapidly with advanced searches that pinpoint the exact email or document in a matter of seconds. Merging of the search and folder paradigms improves user productivity and empowers users with information at their fingertips.

In addition, the Reference Information Storage System search feature is highly advanced in that users can search *content* in addition to header information—meaning all information can be queried, including the content of email attachments and documents. Most importantly, search results are returned rapidly, despite such comprehensive search capabilities.

The underlying architecture to RISS is central to the way it delivers its value. HP has developed Smart Storage Cells to address the scale problem of managing billions of unique records.

Smart Storage Cells

The key benefit of the Reference Information Storage System architecture is that it takes a "divide and conquer" approach to the large storage problem. So instead of attempting to manage billions of emails in one block, the Reference Information Storage System is deployed as a massively parallel architecture of Smart Storage Cells.

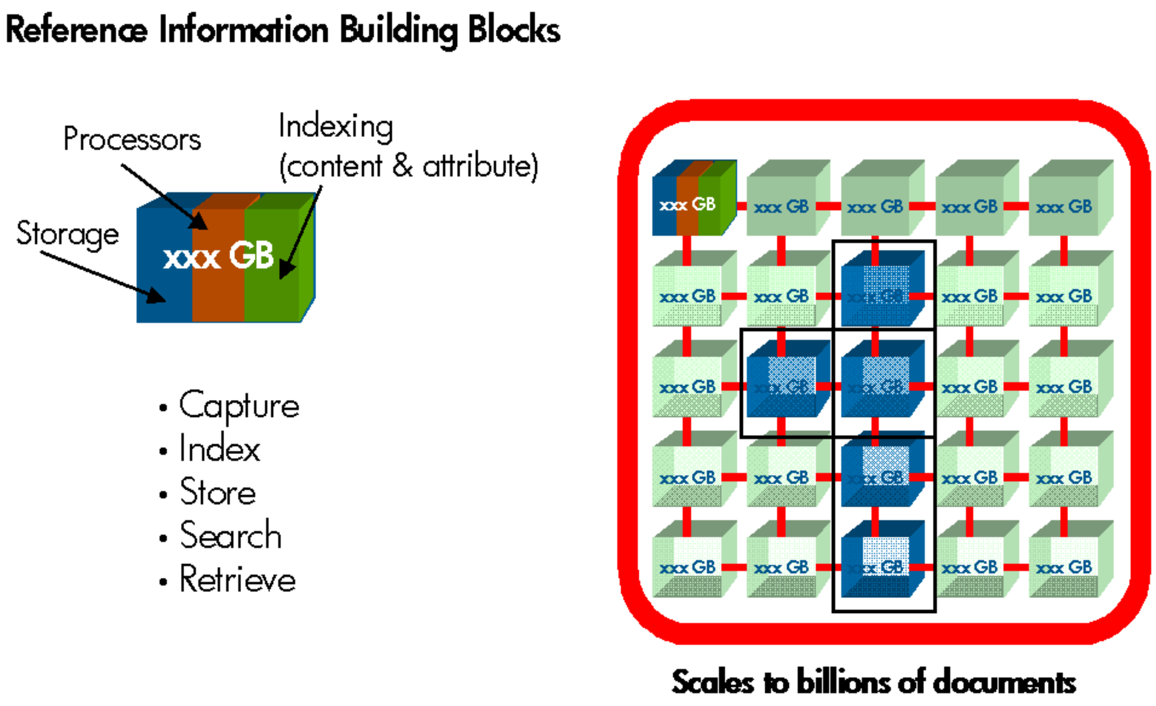
Each Smart Storage Cell contains the core elements needed in managing reference information:

- Indexing capability
- Storage capacity
- Processing power

This is what enables rapid retrieval times. The system has a higher proportion of processing power to storage capacity, and is searching through smaller chunks of data for the user's record. And as stated earlier in this white paper—archive is all about retrieval.

Furthermore, the architecture of the Smart Storage Cells enables scalability to a very large level—as processing power is added each time more storage capacity is added, there is no degradation in performance.

Figure 3. Solving the scalability problem



Alternative solutions

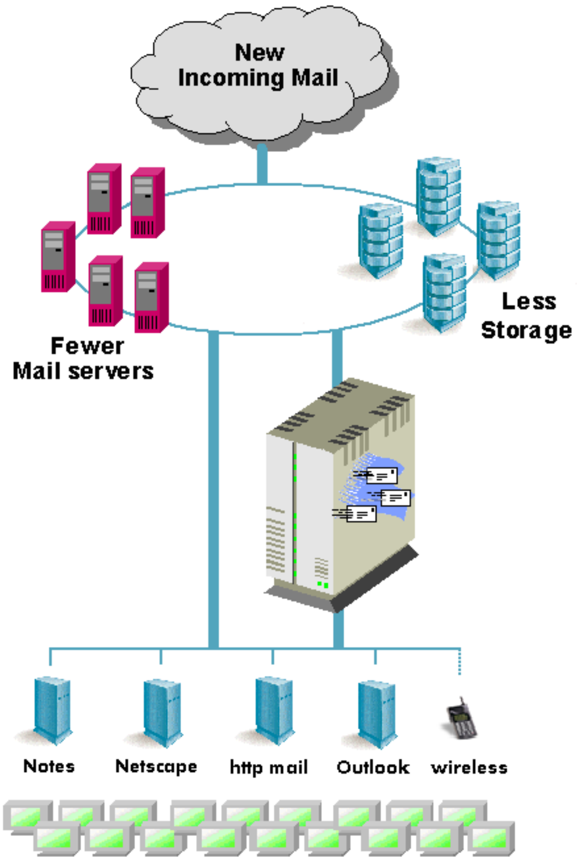
As outlined throughout this white paper, HP StorageWorks Reference Information Storage System delivers a complete solution to the problem of managing the cost of email systems as well as effectively providing access to reference data and assisting companies in their steps to meet data retention regulations.

Adding more servers and storage to existing email systems does nothing to cut the cost of email management. Harvesting messages to RISS can reduce the total cost of email management by placing data in a lower cost archive pool and by automatically deleting duplicates.

Implementing mailbox policies and making users archive or delete messages does not effectively address the cost issue—inevitably there will be multiple copies of the same documents. Furthermore, there is no central control over retaining important records and making valuable information widely available. RISS allows for central archiving policies to retain important items and delete unimportant records, and allows users rapid retrieval.

It is possible to build an email archiving solution from components readily available. However, archiving requires far more than just a low-cost storage pool—and integrating all necessary components is a complex, time-consuming, and inevitably costly process. RISS provides an all-in-one solution.

Figure 4. The RISS solution reduces the TCO of email



- •Reduce Email expense with active archiving
- •An increase in users/volumes does **NOT** demand an increase in servers and storage
- •Mailbox extension
- •Removes duplicate messages and attachments
- •Policy-based archiving
- •Retain all business-critical records
- •Transparent to end-users
- •Rapid retrieval of files
- •Infinite scalability

Conclusions and recommendations

While email has become an invaluable application in the workplace, its growth and storage requirements are growing out of control. There are a host of potential solutions being offered to address this problem; unfortunately, the majority of these are expensive, difficult to integrate and manage, or disruptive to end-user productivity. In general, they are symptomatic remedies, as opposed to permanent solutions for the source of the problem.

The correct solution to the email storage problem is one that provides customers with abundant storage, is simple to install and manage, does not affect end-user productivity, and ultimately lowers TCO. The HP StorageWorks Reference Information Storage System represents a "true solution" to the email growth problem:

- **Lower TCO**—Since the majority of email storage is placed on the Reference Information Storage System, the customer can greatly reduce the number of servers needed. This results in significant cost savings in terms of administration costs, maintenance costs, training costs, and downtime. A 10,000-user company could expect substantial cost savings over a three-year period.
- **Improved user productivity**—End users will receive a supplemental "Reference Information Storage System Inbox" in their email client, which requires virtually no training. More importantly, obstructive mail quotas and other restrictions for end users can be avoided.
- **Abundant storage**—The Reference Information Storage System offers 1 TB to 25 TB of reference data storage, per device, in 100-GB mirrored increments.
- **Simple to install and manage**—Because of the system's plug-and-play architecture, installation takes generally only a matter of hours. The product also offers self-diagnosis and self-healing features to reduce the administration requirements.

Competitive products focus on storage (retention), but Reference Information Storage System has been designed from its conception to focus on retrieval. Reference Information Storage System delivers a truly valuable solution. Enterprises that deploy the Reference Information Storage System will enjoy rapid relief to the email storage crisis, along with a more efficient email system operating at a significantly lower TCO.

Frequently asked questions

What technical capabilities of the ENSAextended architecture does this product enhance?
This product provides a proof point for the HP Information Lifecycle Management vision by providing intelligent, automated archiving of selected emails and Microsoft® Office documents according to the value of the data and without user or operator intervention. Uniquely, the product's powerful search engine and fast retrieval methods have full Application Integration to enable easy access to stored data making this an "active archiving" solution that turns data into information.

My company must demonstrate compliance with specific regulations relating to data retention, data accessibility, and data authenticity. Does this product provide compliance?
This product supports a compliant data storage environment. Individual regulations such as SEC rule 17a or the Health Insurance Portability and Accountability Act have specific clauses that define how data must be stored, for how long, and with what levels of accessibility. The HP solution provides a number of features that enable compliance with these regulations, including writing data in write once read many (WORM) format to disk, electronically fingerprinting the data by way of digital signature methods to ensure that any changes to the data can be easily identified, preventing deletion of files until the end of the retention period, limiting access to archived data to authorized users, and providing compliance reports and audit trails. Using these compliance building blocks, the system can be tailored to address any data retention and management regulations that apply.

Email growth is out of control in my company. Can this product help?
Reference Information Storage System automatically migrates data from expensive enterprise disks to more affordable disk-based Reference Information storage, thus reducing cost while maintaining full data accessibility. In addition the product includes a .PST importer tool that allows consolidation of locally stored and unmanaged personal archives into a centrally located and professionally managed Reference Information store. End user–initiated retrievals by way of intuitive search tools further reduce the cost of managing long-term information repositories.

How do I find the specific email or Microsoft Office document I need from a multi-terabyte archive?
Reference Information Storage System uniquely supports free text searching of stored data initiated from any web browser or directly from the Microsoft Outlook GUI. This means that any stored data object can be quickly located and retrieved. Other metadata such as date of dispatch/receipt or recipient for messages or file creation date for Microsoft Office documents can be used as additional retrieval key words.

For more information

With HP StorageWorks Reference Information Storage System, storage and retrieval of reference data has never been easier. This self-contained, self-managing, and fault-tolerant system transforms unstructured reference data into exploitable information. Now you can manage storage costs and comply with data retention regulations with one solution. Now you can deploy a single, highly scalable cost-effective solution to manage excessive reference information.

www.hp.com/go/ILM

5982-4298EN, 04/2004