TECHNICAL BRIEF

Lifecycle Fixed Content Manager 100 Series solution

Next-generation object storage solutions

DECEMBER 2004

1. Executive summary

In organizations of every size, IT managers are struggling with the explosive growth of information. E-mail, medical records, CAD/CAM drawings, backup data, clinical trial data and media files — these are just a few examples of industry information whose volume is growing at an astounding rate. Certain trends have contributed to this explosive growth and to the need to expand storage capacity. One trend is the digitization of information such as check images, medical images, photographs, video and voice data, which was previously kept on paper, tape or film. In addition, industry and government regulations now dictate how companies must keep corporate records, often requiring that records be retained for many years. Consider these forecasts from the Enterprise Storage Group:

- The volume of fixed content, or "digital assets retained for active reference and value," is growing at 92 percent every year through 2005 compared with 61 percent for traditional data.[1]
- The worldwide capacity of compliant records on all storage media will increase at a compound annual growth rate of 64 percent. On disk-based media, the growth rate is 172 percent. [2]

However, not all of this information is created equal. Some information, such as transactional database records, must be online and requires very fast read and write access. Other types of data, for example archived e-mail, need to be online but will be accessed occasionally and updated infrequently, if at all. A third category of data needs to be archived offsite to be safe in case of an onsite disaster, but will be rarely accessed.

Industry analysts recommend that enterprises create a multi-tier storage infrastructure that includes fast, primary disk storage for mission-critical, transactional data and low-cost, easy-to-manage secondary disk storage for fixed content (see page 4). Lifecycle Fixed Content Manager 100 Series solution is a software product that turns standard servers into scalable, secondary disk storage that meets enterprise needs to store increasing volumes of information online at a reasonable cost.

Lifecycle Fixed Content Manager 100 Series solution offers:

- **Easy integration** — Industry standard NFS and CIFS file access protocols allow easy deployment with existing applications.
- **Workgroup to enterprise scalability** — Add storage servers one at a time and buy only the capacity you need.
- **Increased storage efficiency** — Data coalescence (see page 6) reduces redundant data on disk up to 95 percent, lowering disk capacity needs.
- **Industry standard hardware** — Low-cost storage servers reduce acquisition costs.

Additional benefits of Lifecycle Fixed Content Manager 100 Series solution's architecture include:

- Highly secure fault tolerance in which all data is replicated on more than one server.
- Low system management costs via automated system management, easy scalability and self-healing.

[1] Reference information: the next wave, A "Snapshot" research study by Enterprise Storage Group, 2002

[2] Compliance: the effect of information management and the storage industry, ESG Impact Report, May 2003

2. Business concern

Extensive regulatory requirements for retaining information are new to modern business practices. This type of information can be divided into two broad categories:

- **Fixed content** — Examples include e-mail, office documents, media files, check and document images, medical images and patient records, and CAD/CAM drawings.
- **Regulatory-compliance records** — A subset of fixed content, compliance records, can include e-mail, financial documents and medical records that must be stored according to government and industry regulations.

These different types of data have similar characteristics. This information is:

- Accessed only occasionally, not as often as transactional data.
- Updated infrequently. Note however, that Lifecycle Fixed Content Manager 100 Series solution does not require data to be "fixed content."
- Storage-intensive, for example, X-rays and MRIs in a hospital environment.
- Potentially of high value to the enterprise; for example, customer information that can be reused to develop new service offerings for added revenue.
- Highly redundant; for example, different versions of the same file or files with large portions in common.

2.1 Fixed content

The Enterprise Storage Group (ESG) defines fixed content (reference information) as "digital assets retained for active reference and value." Fixed content includes many different types of data in diverse industries. Examples are e-mail and e-mail attachments, check and document images, medical images and patient records, and CAD/CAM designs. Examples of applications that generate fixed content include e-mail archiving, hierarchical storage management (HSM), document management and office applications. According to ESG, fixed content is growing at a rate of 92 percent CAGR, compared to 61 percent for traditional data such as transactional and database data.

Fixed content is often highly redundant. For example, an e-mail server saves multiple revisions of the same attached presentation. Another example is a project team that saves different versions of design files with large portions in common. Lifecycle Fixed Content Manager 100 Series solution is a low-cost storage solution ideally suited for safe and efficient archiving of fixed content. Lifecycle Fixed Content Manager 100 Series solution is less expensive than primary storage but offers fast data access compared to tape or optical media. Highly redundant data is coalesced for very efficient storage usage. Low price per gigabyte, scalability to tens of terabytes, seamless integration and simple system management satisfy customer requirements for a fixed-content repository.

2.2 Regulatory compliance

Many companies need to comply with government and industry regulations on the long-term retention of records such as e-mail, instant messages, financial documents, and medical and life sciences data. Regulations such as the following define rules for storing records to meet compliance requirements:

- SEC 17a-4 in financial services
- HIPAA in the healthcare industry
- Sarbanes-Oxley Act governing corporate accounting
- DOD 5015.2 in the Department of Defense

According to ESG, the worldwide volume of compliant records will increase by 64 percent compounded annual growth rate between 2003 and 2006. The various regulations stipulate requirements such as data retention periods, write once/read many (WORM) capability, and security and privacy of data. The SEC regulation 17a- 4 is just one of a large number of rules and regulations that govern the storage of archival records; however it is by far the most stringent. As such, it is emerging as an industry standard benchmark for storage vendors so that they can meet requirements for all regulations. In particular, SEC 17a-4 specifies:

- Records must be preserved in a non-rewritable, non-erasable format.
- The system must verify automatically the quality and accuracy of the storage media recording process.

3. Managing compliance data

Core features of the Lifecycle Fixed Content Manager 100 Series solution architecture make it highly suitable for storage of compliant records. Online disk storage is inherently more available and accessible than tape or optical solutions — an important feature when companies need to quickly search and retrieve compliant data. Data coalescence and scalability address the very large volume of data that will result from a company's compliance.

The Lifecycle Fixed Content Manager 100 Series solution specifically addresses SEC required retention periods and data verification. Core features plus Compliance Pack software make Lifecycle Fixed Content Manager 100 Series solution a magnetic, disk-based WORM device that enables enterprises to comply with the most stringent government regulations.

Lifecycle Fixed Content Manager 100 Series solution was designed as a flexible set of building blocks that can be combined in different ways to meet many customer requirements for storage and data protection. With the introduction of Lifecycle Fixed Content Manager 100 Series solution, StorageTek® offers a secure, permanent storage solution for fixed content and a solid foundation for future technology innovation.

The core product provides secure, permanent storage for large volumes of fixed content. Lifecycle Fixed Content Manager 100 Series solution is an object-based data repository with industry standard file interfaces for easy application integration, and data coalescence, which significantly reduces disk capacity requirements with advanced capabilities required for compliance with government regulations. Remote replication offers mirroring to a remote site so that the data can be restored in case of disaster.

4. A multi-tiered storage infrastructure

Industry analysts recommend that enterprises create a multi-tier storage infrastructure that includes high-performance primary disks, mid-tier disks and tape. For example, the Aberdeen Group defines "mid-line disk storage" as "high-capacity online ATA disks, which are in the middle tier between high-use FC/SCSI disk arrays and tape libraries."[3] IDC recommends a "tiered storage solution" that "incorporates the use of capacity drives." IDC also states that lower-cost disks "will stimulate the creation of new applications for disk storage arrays" such as "disk-to-disk-to-tape" (backup) and "online storage of fixed content."[4]

The tiers in a multi-tier storage infrastructure are:

- **Primary disk** — High-cost, enterprise FC- or SCSI-based disks for applications requiring fast read/write performance
- **Mid-tier disk** — Inexpensive, high capacity, somewhat lower performance, ATA disks for data that needs to be online, but is accessed and changed less frequently
- **Tape or optical** — offline, inexpensive storage for archiving large volumes of data

Lifecycle Fixed Content Manager 100 Series solution complements your existing IT infrastructure by off-loading infrequently accessed information from primary storage. It can either replace or complement tape devices and optical media, depending on the balance of access and scalability required for your business.

[3] Mid-line Disk Storage Emerging as Significant Cost-Saving Opportunity, An Executive White Paper by the Aberdeen Group, Inc. October 2003

[4] State of the External Storage Disk Market: 2003 and Beyond, Charlotte Rancourt, IDC's Storage Breakfast Briefing, Storage Network World, October 29, 2003

5. Lifecycle Fixed Content Manager 100 Series solution — key features and benefits

Feature	Benefit
Data coalescence	Lowers storage costs through reduced disk usage.
Fault tolerance	Meets business requirements for high availability and reliability.
Self-managing/self-healing	Requires minimal system administration and lowers system management costs. No interruption to services and continuous uptime when adding or removing servers.
NFSv3/CIFS file access protocols	Provides easy integration and investment protection for popular backup and recovery, content management and archiving solutions.
Encryption	Allows sharing of storage resources across an enterprise in a service-provider model without concern.
Security and privacy	Prevents unauthorized users or applications from accessing data.
Scalability	Enables scaling of capacity by adding storage servers and performance by adding portal servers — one at a time so you can purchase just what you need.
Snapshots	Protects against accidental or malicious deletion or corruption of data with read-only versions of the file system.
Data verification	Guarantees that each record is accurately written to Lifecycle Fixed Content Manager 100 Series solution.
Data integrity checking	Provides that data is not corrupted on disk, a highly unlikely occurrence.
Compression	Further reduces storage usage beyond the savings from coalescence.
Remote replication	Provides fast restoration of data in case of disaster.

Lifecycle Fixed Content Manager 100 Series solution is a secure, permanent storage solution for fixed content and a solid foundation for future technology innovation.

5.1 Object-based storage

Lifecycle Fixed Content Manager 100 Series solution uses an object-based storage model to store and manage data. Lifecycle Fixed Content Manager 100 Series solution access nodes give each block of data a unique name or content address, based on a cryptographic hash of its contents. Lifecycle Fixed Content Manager 100 Series solution uses this content address to identify data, to distribute data evenly across available storage servers, to test whether a block of data exists in the storage pool, and to efficiently retrieve the data. To distribute data evenly across all servers, each storage server is in charge only of certain ranges of block addresses. Applications do not need to understand and administer the physical location of a block of data on disk, nor do they need an application programming interface (API) to access data in Lifecycle Fixed Content Manager 100 Series solution. Content addresses also enable data coalescence, or the elimination of unwanted data redundancy, a key advantage over tape, optical and most secondary disk storage products.

Lifecycle Fixed Content Manager 100 Series solution uses the standard SHA-256 hash (256-bit) as a content address, which provides a greater guarantee of system integrity than shorter hashes used in competitive systems.

5.2 Data coalescence

Content addresses enable data coalescence, which provides that only one copy (plus a replica) of any block of data is stored. For all subsequent blocks with the same content, Lifecycle Fixed Content Manager 100 Series solution saves a reference, or pointer, to the block of data. By reducing the number of copies stored, Lifecycle Fixed Content Manager 100 Series solution reduces disk space usage, network bandwidth and the cost of storing information. Different terms are used in the industry for the concept of data coalescence, including un-duplication, commonality factoring and single-instance store. Lifecycle Fixed Content Manager 100 Series solution coalesces data at the block (or sub-file) level whereas competitive products coalesce data at the file level. Lifecycle Fixed Content Manager 100 Series solution's block-level coalescence results in more efficient disk usage because it looks at data at a more granular level. Today's enterprises store massive amounts of redundant data, especially backup data.

The traditional weekly full backup plus daily incremental backup schedule means that backup data sets typically include a large degree of duplication. Unchanged files are backed up at different times or from different servers.

Unchanged portions of different file versions are also stored redundantly. Lifecycle Fixed Content Manager 100 Series solution can reduce this data redundancy and the amount of storage used by up to 95 percent, depending on the data set.

5.3 Fault tolerance

Lifecycle Fixed Content Manager 100 Series solution is highly fault tolerant — there is no single point of failure. The system automatically replicates each block of data on more than one server. If a server fails, other servers take over and create additional replicas to replace data held by the missing server. This data is evenly distributed among the remaining servers and all servers participate equally in the data recovery. This shared recovery is faster than with mirroring, where the recovery burden falls on just two servers.

The idea of no single point of failure extends to hardware redundancy. Each server has two network connections, each connected to two different network switches. If a switch fails, no connectivity is lost.

5.4 Self-managing/self-healing

System management is the dominant cost in large-scale disk storage. Lifecycle Fixed Content Manager solution provides these automatic management features:

- **Self-healing** — If a server fails, its data is automatically replicated on other servers.
- **Data migration** — When a server is added, the system automatically migrates data to it.
- **Self-protecting** — All current data and historical snapshots are replicated; therefore no separate backup is required.
- **Low management overhead as a storage clique scales** — Adding capacity does not require any manual reconfiguration, data migration or downtime. Data remains accessible online.
- **Load balancing** — Data is evenly distributed across servers. CPU, network and disk resources of all servers are used in parallel, optimizing performance during read, write and failure-recovery operations.

5.5 Standard interfaces

Lifecycle Fixed Content Manager 100 Series solution offers industry-standard file access protocols for easy integration with standard backup, content management and archiving solutions. Users can add Lifecycle Fixed Content Manager 100 Series solution to their IT infrastructure without altering existing applications or incumbent backup strategies.

Lifecycle Fixed Content Manager 100 Series solution supports NFSv3 and SMB/CIFS via Samba. These interfaces map familiar file names and directories onto the underlying content addressed storage. Alternatively, to integrate Lifecycle Fixed Content Manager solution with a Microsoft Windows-based backup or archive application, you can run Microsoft Windows Services for UNIX 3.0 on the Windows server and then access a Lifecycle Fixed Content Manager solution volume directly via NFS.

5.6 Encryption (optional)

Storage resources can be shared within an enterprise and data can be replicated to an unsecured location with no danger of unauthorized access. If encryption is enabled, all blocks of data are encrypted using the AES encryption algorithm, and cannot be accessed by other Lifecycle Fixed Content Manager 100 Series solution users or applications.

5.7 Security and privacy

Lifecycle Fixed Content Manager 100 Series solution builds in strong security and privacy features at a fundamental level.

- **File access control** — Lifecycle Fixed Content Manager 100 Series solution access nodes preserve and enforce NFS file permissions. Only authorized users can access data.
- **NFS export control** — The Lifecycle Fixed Content Manager 100 Series solution installer can explicitly specify the client systems that can access data on Lifecycle Fixed Content Manager 100 Series solution storage.
- **Built-in privacy** — Although storage space is conserved dramatically through coalescence, even system operators cannot identify users that share the same content.

The system also protects against data deletion by malicious or accidental acts. The Lifecycle Fixed Content Manager 100 Series solution strictly enforces data retention policies, which is critical if a disk-based system is to provide the same level of protection as a system based on removable media such as tape or optical disk.

5.8 Scalability

Lifecycle Fixed Content Manager 100 Series solution scales from an affordable 2.3-terabyte configuration for an entry-level environment to a nearly 20-terabyte configuration for the most storage-intensive environments. A Lifecycle Fixed Content Manager 100 Series solution configuration consists of storage servers, providing disk capacity, and portal servers, providing application access through NFS or CIFS. To scale capacity, simply add storage servers; to scale throughput, add access servers. Storage servers can be added four at a time — allowing you to purchase just the capacity you need when you need it.

Lifecycle Fixed Content Manager 100 Series solution transparently adjusts to the addition or removal of storage servers — as a new server is added, data migrates to it automatically while the system continues to operate normally. There is no need to schedule downtime, manage volumes or build a new file system.

File systems that share a Lifecycle Fixed Content Manager 100 Series solution data store can each use any portion of the storage. The storage is treated as a single shared resource. Lifecycle Fixed Content Manager 100 Series solution allows file systems and volumes to grow as large as desired, limited only by the total storage space.

5.9 Snapshots

Snapshots are online read-only versions of the file system at a particular point in time, which protect against accidental or malicious deletion or corruption. Snapshots can be configured to run automatically at regular intervals; for example, daily. Snapshots can be used for daily backups and user initiated restores. If the snapshot directory is mounted for a user, the user can copy a file or file system from the directory to the user's directory.

Snapshots are simply lists of pointers to the addresses of data blocks stored in a clique at a certain point in time, from which a volume can be reconstructed. Thus, creating a snapshot requires only a small amount of additional disk space.

5.10 Data verification

When a block is deposited into the system, the access node storing it generates a content address and the server receiving it recalculates the content address. Data verification addresses the SEC 17a- 4 requirement, which states, "the system must verify automatically the quality and accuracy of the storage media recording process."

5.11 Data integrity checking

Lifecycle Fixed Content Manager 100 Series solution continuously checks the integrity of data with a background process, which identifies and rectifies bad blocks. This same process frees up space consumed by objects that have been deleted so that this space can be reused for additional data storage. As it traverses the data store, it defragments the disk and verifies the checksums of each block of data. In the event of a checksum mismatch, the system replaces missing blocks.

5.12 Compression (optional)

If this option is enabled, all blocks of data are compressed using the zlib compression algorithm. Compression will save storage space but slows down write performance. If both compression and encryption are enabled, Lifecycle Fixed Content Manager 100 Series solution compresses and then encrypts every block. The block is then deposited in the data store where it may coalesce with other blocks.

5.13 Remote replication for disaster recovery

This feature extends fault tolerance to offsite replication and disaster recovery. Two systems can be configured, in geographically separate locations, to replicate data over a WAN to each other in order to protect business-critical data from site disaster. In case of a catastrophic event at one site, data will be safe at the remote site and can be used to regenerate the original data store. Bandwidth usage for Lifecycle Fixed Content Manager 100 Series solution remote replication is highly efficient. Data coalescence prevents repeated transmission of identical blocks of data, decreasing network usage.

6. System components

A Lifecycle Fixed Content Manager 100 Series system consists of a storage nodes and access nodes. A storage node block is a set of four or more storage servers connected to each other via redundant switches. These servers are industry standard servers with internal disks that store and retrieve data.

Applications access the clique through one or more access nodes using the NFS or CIFS file access protocols. The Lifecycle Fixed Content Manager 100 Series solution Web console with a graphical user interface provides a single point of administration for the storage servers.

Each access node runs the following software:

- The Lifecycle Fixed Content Manager 100 Series solution NFS portal, which allows NFS clients to access the contents of the storage nodes.
- Optionally, SMB/CIFS via Samba, which allows Microsoft Windows clients to access the contents of the clique.
- The Lifecycle Fixed Content Manager 100 Series solution Web console used to manage the system.

Each Lifecycle Fixed Content Manager 100 Series solution storage node runs the Lifecycle Fixed Content Manager 100 Series solution storage server software, which provides self-healing, load balancing and instant scalability.

The access nodes subdivide data into blocks, each with a unique content address. Servers route, store and replicate the object blocks. The system saves one copy of this block, plus a replica. If an original block is corrupted or lost, the replica takes over. To spread management of blocks over all the servers as equally as possible, each storage server is in charge of only certain ranges of block addresses.

6.1 Lifecycle Fixed Content Manager 100 Series solution configuration

The basic configuration of the Lifecycle Fixed Content Manager 100 Series solution consists of two access nodes and four storage nodes, providing 2.3 terabytes of usable storage. Storage can be added in blocks of four storage nodes up to a total of 18.6 terabytes usable storage. The base configuration also includes a battery backup (UPS) and 2 Gig-E 24-port switches.

7. Deployment architecture

A Lifecycle Fixed Content Manager 100 Series solution deployment consists of Lifecycle Fixed Content Manager 100 Series software on a cluster of networked access nodes, storage nodes and a Web-based management console. Storage servers provide disk capacity and portal servers provide application access through NFS or CIFS. Because multiple access nodes can write and retrieve data to the underlying storage servers in parallel, adding access nodes increases performance.

Lifecycle Fixed Content Manager 100 Series solution storage servers and access nodes do not need to be co-located. Locating access nodes near application servers optimizes network performance. Because the access node coalesces the data, the volume of data sent across the network is a fraction of the original data set. Data sent from access nodes to storage servers can be encrypted so that there is no danger of unauthorized access.

Multiple applications can securely and privately share the underlying Lifecycle Fixed Content Manager 100 Series solution storage. Applications can each access the system through a dedicated access node or applications can share access nodes. In this picture, an e-mail server, various application servers and a backup server all write to Lifecycle Fixed Content Manager 100 Series solution storage. The e-mail server and the backup server each access Lifecycle Fixed Content Manager 100 Series solution through a dedicated access node. The UNIX and Linux servers share an access node.

8. Benefits of Lifecycle Fixed Content Manager solution

The Lifecycle Fixed Content Manager 100 Series solution software product turns standard, server hardware into scalable, content addressed, secondary disk storage that meets enterprise needs to store high volumes of information online at a reasonable cost. Lifecycle Fixed Content Manager 100 Series solution was designed to store fixed content and compliance records. Object-based storage traditionally provides these benefits:

- Data coalescence, which minimizes the storage of redundant copies of data and thereby reduces storage capacity needs
- Highly secure fault tolerance, in which all data is replicated on more than one server
- Low system management costs via self-healing and easy scalability

Lifecycle Fixed Content Manager 100 Series solution's architecture offers additional benefits over earlier generations of object-based storage:

- Easy integration with industry standard NFS and CIFS file access protocols
- Workgroup-to-enterprise scalability allowing you to buy just the capacity you need
- Increased storage efficiency with block level (versus file level) data coalescence
- Industry standard hardware, which reduces total cost of ownership

Lifecycle Fixed Content Manager 100 Series solution is an affordable solution ideally suited for safe and efficient storage of fixed content such as e-mail archives and office documents. Lifecycle Fixed Content Manager 100 Series solution will scale over time as enterprises add additional applications, such as regulatory compliance, which can all share Lifecycle Fixed Content Manager 100 Series solution's secure storage resources.

ABOUT STORAGETEK®

Storage Technology Corporation (NYSE: STK) is a $2 billion global company that enables businesses, through its information lifecycle management strategy, to align the cost of storage with the value of information. The company's innovative storage solutions manage the complexity and growth of information, lower costs, improve efficiency and protect investments. For more information, visit www.storagetek.com, or call 1.800.275.4785 or 01.303.673.2800.

WORLD HEADQUARTERS

Storage Technology Corporation
One StorageTek Drive
Louisville, Colorado 80028 USA
1.800.877.9220 or 01.303.673.5151

FT 0004 A 12/04