Design of a WORM Filesystem

Terry Stokes
Dell EMC/Isilon
Who am I?

- 15+ years in network protocols.
- Previously worked at Blue Coat Systems, RadioFrame and Microsoft Anti-trust (MCPP).
- Co-founded a network capture device startup in 2003.
- Currently work in the AIMA group at Isilon.
Disclaimers

- Not Isilon Smartlock design.
- Dated design, developed in 2003.
- I’m not a filesystems expert.

Product called the “Compliance Appliance”.

Capture and archive electronic communications.

Incorporated a WORM filesystem.

10+ “sweat equity” employees.

Two WORM filesystem patents.

Failed to get funding.
Compliance Appliance
Inspection, Analysis, Discovery and Reporting
Store & Archive
Traffic Capture (email and IM)
Other Traffic
Enterprise Data Center
Traffic...
Policy-Based Capture, Retention, Access Control

Application Dashboard
- Reports
- Discovery
- Workflow

SEARCH
Messages And Message Relationships

Index Store

Indexer: Message Relationships

Archiver: Structured Objects

External Storage

LDAP Directory

IT

Litigation & Compliance

SOAP Query & Analysis

Email
IM
Web
Existing Archives

IT Network

...
WORM Storage 2003

- Optical storage (CD-R, optical tape, etc).
- EMC Centera CE
  - Content-Addressable Storage
  - Content hashed for storage location.
- NetApp Snaplock
  - Plugin for Data OnTap
  - Software prevents file deletion.
Customer Research

- Talked to 10-15 potential customers.
- Banks, Brokers, Healthcare, Manufacturers.
- 10-15 investment firms.
- Several had home grown software solutions.
- A few simply backed up optical storage.
- Issue was greater than simply providing a better WORM storage solution.
# Regulations

<table>
<thead>
<tr>
<th>SOX Section 404, 802</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIPAA Protected Health info</td>
</tr>
<tr>
<td>SEC 17a-4</td>
</tr>
<tr>
<td>NASD 3010/3110</td>
</tr>
<tr>
<td>NYSE 440/472</td>
</tr>
<tr>
<td>Basel-II Supervisory Control</td>
</tr>
<tr>
<td>CA SB 1386 Privacy and Protection</td>
</tr>
<tr>
<td>US Patriot Act Cust Identification Pgm</td>
</tr>
<tr>
<td>ISO 17799 Retention &amp; Safeguarding</td>
</tr>
<tr>
<td>DOD 5015.2-STD ERecords Management</td>
</tr>
</tbody>
</table>
Multiple Protocols

- SEC rules are ambiguous:
  - All broker-dealer communications must be held for 3 years on non-rewriteable and non-erasable storage.
  - Originally SEC only enforced email storage.
  - Added instant messaging requirement.
  - What about broker web pages? VoIP?
- Each protocol requires a separate solution.
Slow Message Retrieval

- Message discovery requests (from legal and SEC) have a time limit.
- Many SEC requests involved several brokers, with each broker search taking 2-3 days.
- By the time search was complete, only a day or two left to review before handover.
- Didn’t know what was being handed over.
- Legal Discovery industry born from this.
Deletion of Expired Messages

- Emails can hurt you!
  - Harmful emails have been used in several notable trials, Microsoft for one.
  - One unexpired email can indirectly hold hundreds of expired emails on optical storage media.
- It is important for a company to delete emails as soon as regulations allow.
Survey of Filesystems

- Reviewed top ten filesystems for ideas.

- **ReiserFS:**
  - Uses B+ trees for file objects (inodes, directories and file data).
  - Employs tail packing.

- **Sun XFS:**
  - Use B+ trees for tracking free extents and free inodes.
  - Uses extents instead of blocks.
Extents
Distributed Storage
## Distributed Storage

<table>
<thead>
<tr>
<th>Start Date</th>
<th>ID Start</th>
<th>ID Stop</th>
<th>Location</th>
<th>Storage Partition</th>
<th>State</th>
<th>Free MB</th>
<th>Access ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3/06</td>
<td>0000</td>
<td>1234</td>
<td>London1</td>
<td>\groupL1a</td>
<td>ready</td>
<td>52369</td>
<td>132</td>
</tr>
<tr>
<td>10/23/05</td>
<td>1235</td>
<td>2254</td>
<td>NYC1</td>
<td>\groupN1a</td>
<td>ready</td>
<td>43221</td>
<td>23</td>
</tr>
<tr>
<td>10/23/05</td>
<td>2255</td>
<td>3378</td>
<td>NYC2</td>
<td>\groupN2a</td>
<td>ready</td>
<td>96676</td>
<td>34</td>
</tr>
<tr>
<td>10/23/05</td>
<td>3378</td>
<td>4865</td>
<td>NYC2</td>
<td>\groupN2b</td>
<td>ready</td>
<td>45312</td>
<td>35</td>
</tr>
<tr>
<td>9/18/06</td>
<td>4866</td>
<td>7697</td>
<td>Boston1</td>
<td>\groupB1a</td>
<td>ready</td>
<td>12314</td>
<td>80</td>
</tr>
<tr>
<td>6/16/06</td>
<td>7698</td>
<td>8745</td>
<td>NYC2</td>
<td>\groupN2c</td>
<td>ready</td>
<td>23890</td>
<td>34</td>
</tr>
<tr>
<td>9/18/06</td>
<td>8746</td>
<td>9999</td>
<td>Boston1</td>
<td>\groupB1b</td>
<td>ready</td>
<td>67114</td>
<td>85</td>
</tr>
<tr>
<td>3/27/04</td>
<td>2687</td>
<td>3956</td>
<td>NYC1</td>
<td>\groupN1a</td>
<td>read only</td>
<td>43221</td>
<td>23</td>
</tr>
<tr>
<td>7/14/03</td>
<td>5586</td>
<td>6132</td>
<td>NYC1</td>
<td>\groupN1b</td>
<td>read only</td>
<td>324</td>
<td>23</td>
</tr>
</tbody>
</table>
Unsupported Operations

**Supported Operations**
- Filesystem Mount/Unmount
- Filesystem Statistics
- Volume Statistics
- File Creation
- File Reading
- File Deletion (if past retention period)
- File Annotation (auditing functions)

**Unsupported Operations**
- File Deletion (prior to retention period)
- File Content Modification or Appending
- File Attribute Modification
- Directory Support (creation, listing, deletion, etc)
- Symbolic Links
Specialized Storage

- Formatted for electronic communications.
  - Email
  - Instant Messaging
  - HTTP – Web Pages
  - VoiceOverIP
- File content and meta data encrypted.
Specialized Storage

- Restricted File Writes
  - Write out entire file or not at all.
  - No appends or modifications.
  - NVRAM backed journal for crashes.

- Restricted File Reads.
  - Only two types:
    - Read inode (for searches)
    - Read entire file (for viewing/gathering)
Layered Filesystem
Data Encryption/Compression Layer

Virtual File System (VFS)

Block Accumulator

Text Data

Compression

Decompression

Encryption

Decryption

File Reader/Writer

Off-Box NFS/SAN Storage

On-Box WORM FS Layer
WORM Layer
Volume Layout

- Volume Boot Block
- Public Key Certificate
- Filesystem Super Block
  - AG Header
  - Inode and Data Blocks
  - AG Header
  - Inode and Data Blocks
Structure Data

INDEX SPACES

Message Relationships

Headers: Relational DB

Attachment Associations

Content + Att: Reverse Keyword

Structured Objects

Indexed Spaces

Msg Protocol | Msg Headers | Msg Content | Msg Attachments | Msg Protocol
Superblock Format
AG Header Format

- AG Metadata
- Free Extents List
- Free Indoes List
- Lockbox
Inode Format
File Format

- Meta Data
  - Item Headers
  - List of Related Hashes
  - Message Headers
  - Entire Message or Message Beginning

- Version
- Type
- Flags
- Create Date/Time
- Retention Period
- Original Size
- Num of Related
- ...
Disk File Format
Automatic File Deletion

- Hourly process to delete expired files.
- Deletion holds table:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Type = Any and Groups = Execs</td>
</tr>
<tr>
<td>2.</td>
<td>Type = Email and Users = John Smith, George Martin, Pete Nobody</td>
</tr>
<tr>
<td>3.</td>
<td>Type = Spreadsheets, Email, Reports and Groups = Execs, Finance</td>
</tr>
<tr>
<td>4.</td>
<td>Type = VoIP and Groups = Customer Service</td>
</tr>
<tr>
<td>5.</td>
<td>Type = Any and Users = All</td>
</tr>
</tbody>
</table>
WORM Characteristics

- Disk structures and data encrypted.
- Non-standard disk format.
- Unsupported file and directory operations.
- Access to file contents restricted via software.
Performance Characteristics

- Use of extents.
- File contents optimized on disk.
- Tail packing in inode.
- Read operations optimized for archival.
- Structured file format.
- B+ trees for free inodes and extents.
If I did it again…

- Investigate SED drives.
- Drop support for off-box storage.
- Include disk defragmentation scheme.
- Add rudimentary directory structure.
- Remove distributed storage by location.
Questions?

terry.stokes@emc.com