



iArchives Accelerates Document Digitization  
Using DataDirect Networks S2A Storage Array

www.datadirectnet.com

## Challenge

Finding a storage solution that could consolidate 25 different file systems on a single array, scale to massive capacity for image archiving, and deliver high performance to simultaneously support multi-core, multi-threaded image processing and the ingest of new images from multiple scanners

## Application

iArchives converts microfilm and original print content into searchable, digitized online databases using high-speed scanners paired with proprietary image manipulation and optical character recognition (OCR) techniques

## Solution

- S2A storage array provides a cost-effective, high-capacity consolidated storage pool
- S2A performance speeds image capturing and processing operations.
- DataDirect S2A system allows use of cost-effective SATA drives with high reliability



## iArchives Accelerates Document Digitization using DataDirect Networks S2A Storage Array

iArchives' vision is to be the world leader in transforming microfilm and other print content into searchable, digitized, online databases. To achieve that vision, iArchives provides technology and processes that substantially reduce the cost and time it takes to archive documents while enhancing the user's experience in exploring those documents. The company's state-of-the-art software converts print or microfilm-based content into a customized database searchable over the Internet or an intranet.

“I've never seen an array as fast as the DataDirect Networks array!”

— Daniel Leaberry,  
Senior Systems Administrator  
at iArchives

Customers such as the Dallas Morning News, Brigham Young University, the Library of Congress, the National Archives and Records Administration, the Church of Jesus Christ of Latter Day Saints, and the University of Utah, rely on iArchives to transform massive amounts of historical documents, records, and other content into digital archives that appear in their original context with the added benefit of being searchable. iArchives also hosts archived content through Footnote.com, the consumer website of iArchives.

### Efficient Digitization Requires High-Performance Consolidated Storage

The company's challenge was to consolidate 25 different file systems and associated storage into a single, common storage pool that could rapidly ingest scanned content while simultaneously providing high-speed access to technicians using custom software that implements proprietary OCR (Optical Character Recognition) techniques and unique algorithms to scan, de-skew (or auto-straighten), crop, clean up, enhance, and index images. Because the image capturing and processing operations occur concurrently, there are heavy

demands on the storage system.

For ingesting content, iArchives uses five microfilm scanners and one microfiche scanner that run around the clock, seven days a week, streaming images at 100Mbit/s each. In addition, they regularly receive drives ranging in size from 750GB to one terabyte from partner companies who do their own scanning but want to have iArchives perform indexing on the content.

iArchives operates a cluster computer consisting of more than 250 multi-threaded CPU cores, resulting in 400 or more threads hitting the storage array at any one time for processing. Each large original TIFF image will produce five derivative images, plus associated XML metadata files, after processing. With 25 separate file systems, scanned data would be stored on the system with the most available space, depositing data among the different file systems and making it very difficult to locate, manage, and process the images. In addition, the cluster was often underutilized because the storage couldn't support simultaneous processing by all the nodes.

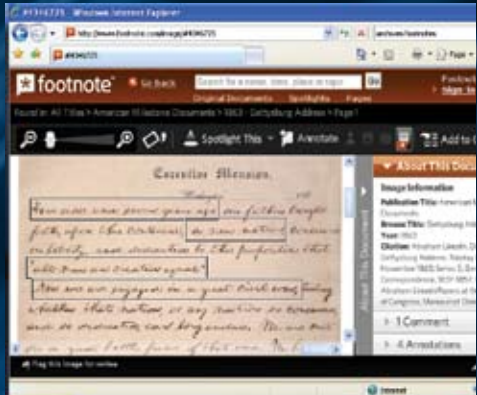
“I worry about how fast my processing is, and how fast the finished images can be served up to the customers,” explained Daniel Leaberry, iArchives' Senior Systems Administrator. “Our primary issue was not bandwidth for the scanners, but rather providing a storage system that could be responsive when every processing node hits the storage array and expects an immediate response.”

He continued, “We were originally going to use generic ‘white box’ storage servers, but realized that they would be hard to support. We evaluated a number of storage solutions, but we were concerned about the complexity of implementing a failover solution with one product we considered. The Lustre file system community widely uses DataDirect Networks' storage, so based on that and other positive research, we decided to invest in a DataDirect Networks solution, which was still competitive on price. I'm very happy with our decision.”

www.datadirectnet.com



Investigation and Trial Papers Relating to the Assassination of President Lincoln



Abraham Lincoln, Draft of the Gettysburg Address 1863, Page 1



Project Blue Book Report Document

## DataDirect Networks' S2A Solution

iArchives storage system is anchored by DataDirect Networks' S2A (Silicon Storage Appliance) solution, with an S2A controller fronting 86TB of cost-effective SATA drives and using InfiniBand interconnects paired with the Lustre parallel file system.

The S2A storage controller incorporates DataDirect Networks' SATAssure intelligent SATA drive management system, which makes large pools of SATA drives reliable, increases their uptime, and ensures their data integrity. SATAssure was important in the iArchives centralized storage configuration because it allowed the use of lower cost, larger capacity drives without sacrificing reliability or data availability.

The system's performance has impressed Leaberry and iArchive's technicians by delivering sustained high-throughput reads and writes as multiple systems simultaneously access content for processing while ingesting new content at the same time. Additionally, the system's fault-tolerant architecture with inherent zero-time failover appealed to the team for hassle-free reliability.

"DataDirect has such a fantastic array. We're not yet using more than a quarter of its performance capabilities. My big file benchmark test yielded over one gigabyte per second throughput through the file system, and that's on just a singlet," Leaberry enthused. "The

stability is marvelous. We've never had a controller crash or had any hiccups or failures at all. In the course of three months since we installed it, we've pushed at least 86TB through it and never had an issue."

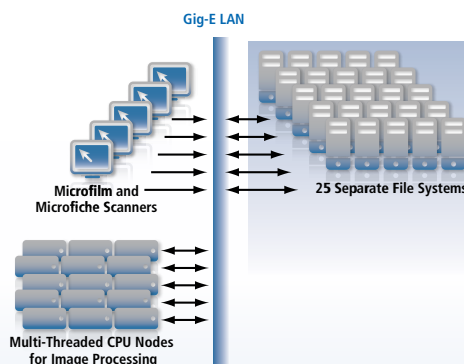
Leaberry and his team liked DataDirect Networks' unique design approach, which integrates multiple zero-latency, full-access host ports, each able to access the entire pool of storage at full speed at the same time, while delivering solid reliability.

He continued, "We're now adding new content at the rate of about 600GB per day, or nearly 20TB per month. The remarkable performance of the DataDirect Networks system has allowed us to dramatically increase the rate at which we are able to add new content, letting us complete projects for our customers much faster than before the DataDirect Networks implementation."

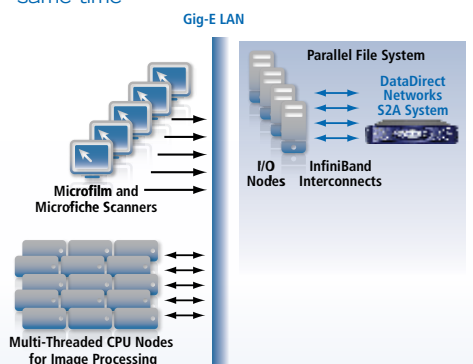
To keep pace with this growing need, and because the S2A-based system has reduced his costs by half, Leaberry and his team are planning to add an additional DataDirect Networks S2A array in the near future.

Leaberry summed it up, "I expected the DataDirect Networks solution to be much more expensive. I was pleasantly surprised at the price/performance ratio. In fact, I've been telling everyone I know that if they need a pool of storage in the neighborhood of 90TB or more, they should buy DataDirect Networks."

Before S2A, data was scanned to one of 25 separate file systems, and trying to process images using all the CPU nodes would cause crashes



With S2A, simultaneous scanning and processing to a common storage pool is possible, and all CPU nodes can reliably operate on the file system at the same time



DataDirect Networks is the leading provider of scalable storage systems for performance and capacity drive applications. DataDirect's S2A (Silicon Storage Appliance) architecture enables modern applications such as video streaming, content delivery, modeling and simulation, backup and archiving, cluster and supercomputing, and real-time collaborative workflows that are driving the explosive demand for storage performance and capacity. DataDirect's S2A technology and solutions solve today's most challenging storage requirements, including providing shared, high-speed access to a common pool of data, minimizing data center footprints and storage costs for massive archives, reducing simulation computational times, and capturing and serving massive amounts of digital content.

Major corporations, supercomputing centers and rich media organizations, including AOL, Ascent Media, Autodesk, Boeing, CNN, Disney, Federal Reserve Board, FedEx, Ford, Hess, Kodak Gallery, Lawrence Livermore National Laboratories, NASA Ames, RIOT, Sandia National Laboratories, Sony, Technicolor, Time Warner, Thomson, Universal, and Veritas DGC, utilize DataDirect high performance, high capacity solutions.

**DataDirect**  
N E T W O R K S  
Performance. Capacity. Innovation.

9351 Deering Avenue . Chatsworth . California 91311  
phone +1.800.TERABYTE (837.2298) . fax +1.818.700.7601  
sales@datadirectnet.com  
www.datadirectnet.com