Starfish Scan Accelerator for VAST

Summary

Clemson University, a Starfish client for many years, had performance challenges backing up their large VAST storage volumes due to large numbers of small files needing to be scanned and copied during each backup operation. After deploying Starfish's Scan Accelerator for VAST in March 2026, the Clemson Storage team can now regularly complete backups within their target window of 24 hours. This includes two volumes that together contain more than 2.4 billion files and over 160 million directories.



Customer

Clemson University

Challenge

Unable to back up all seven research volumes in a 24-hour window due to small files numbering in the billions.

Solution

Starfish Scan Accelerator for VAST was deployed against the two largest volumes.

Result

Backup for all seven volumes, totaling over 2.4 billion files and 3.5 PB, was easily completed in 24 hours.

© 2025 Starfish Storage Corporation. All Rights Reserved.

Challenge & Context

Clemson's Research Infrastructure storage team, led by Storage Architect Becky Ligon, has been a Starfish customer for many years and uses the system to address a range of use cases, including as a backup solution. The team is responsible for backing up seven of the university's busiest research volumes, all housed on their VAST storage system. Prior to installing Starfish's beta-release Scan Accelerator for VAST, only five volumes could be backed up within 24 hours. Backing up the remaining two volumes was so time-consuming and labor-intensive that they regularly disrupted the backup schedule for the entire dataset, a reality which Becky and her team found unacceptable.

At the root of the problem was the so-called **"small file tax."** Two volumes carried over 1 billion small files each, plus millions of directories. Most filesystems lose efficiency when this many small files require metadata to be queried (lstats) for subsequent operations. The file markers on the disk, known as inodes, have to be read one at a time, increasing system load and causing slowdowns and lags throughout the system. Scanning these inodes for changes to determine if the file needed to be backed up put an unsustainable drag on system resources. Backups frequently slowed to a crawl or stopped altogether. This was in spite of the fact that the amount of data being backed up was not excessively large for a 24-hour window.

Large Research Volumes	Capacity	Metadata Scanned Per Run
Project volume	3 PB	1.3 billion lstats; 42 million listdirs
Home volume	185 TB	1.1 billion lstats; 125 million listdirs

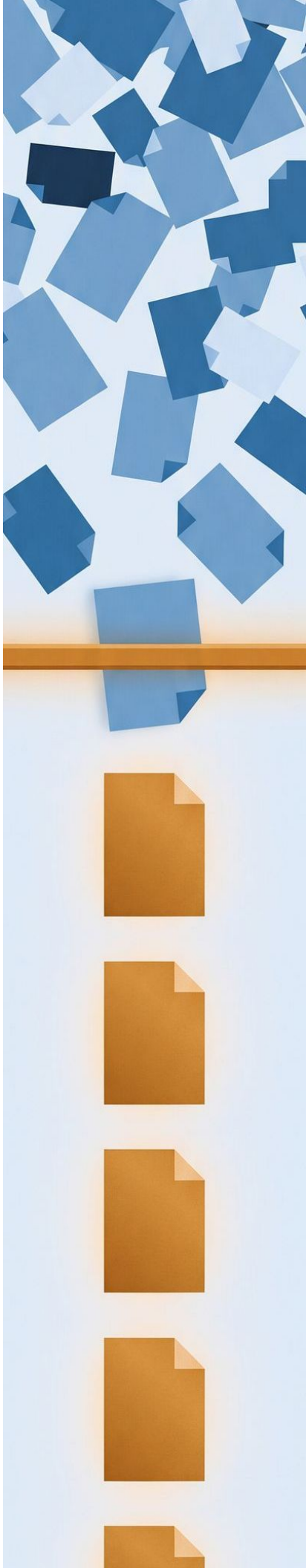
Capacity and metadata volume for Clemson's two largest research volumes.

Given the large number of inodes, scanning these volumes took more than seven hours each. Scanning was only the first step. The subsequent copy job for the Projects volume took more than 16 hours, and the copy job for the Home volume took more than 6 hours (for the same reason as before).

Before Solution	**Scan Time**	**Copy Time**
Project volume	7+ hours	16+ hours (170 destination media sites)
Home volume	7+ hours	6+ hours

Scan and copy times before deploying the Starfish Scan Accelerator for VAST.

The only dependable option was to scan and copy the volumes independently, which consumed more than 4 days each week. Even with this strategy, Becky had to monitor for failures throughout the process. This adaptation left little room for the other five volumes, which slipped to once-weekly backups.



The Solution

Becky learned that Starfish was developing a scan accelerator for VAST while attending Supercomputing 2025 and promptly agreed to be a beta testing site for it. The Starfish team worked with Becky to install the accelerator on Clemson's existing infrastructure. Clemson needed no new hardware.

Today, Becky runs two VAST accelerator instances on a single Starfish agent, one for each "problem" volume. The Accelerator's ability to efficiently query the VAST catalog for updates eliminates the need to scan the entire file system for changes. (Clemson's VAST system is configured to create new system updates once an hour.) Once the Accelerators detect an update to the overall file system, they scan the affected files for changes in metadata. Instead of copying the full filesystem for every backup, Starfish only copies files that have changed since the last backup.

The Results

In a representative run on June 30, 2026, querying the Projects volume data in the VAST catalog for changes took only 38 seconds. Querying the Home volume data in the VAST catalog for changes took only 23 seconds. Subsequent to that, the Starfish Catalog scanned only the changed files. By reducing the file system scan to only changed files, scanning took minutes not hours. Now, on average, scanning changed files takes appx 15 minutes or less vs. a previous total scanning time of more than 7 hours. Thanks to frequent, smaller updates based on only changed data, copy time for the Home volume dropped from 6+ hours to 15 minutes (Copy time for the Projects volume remained at 16+ hours since these copies were sent to over 170 different destinations).

With Scan Accelerator	Catalog Query Time	Catalog Detail	Time to Scan Changed Files	Backup Time to Backup Sites
Project volume	20–40 seconds	326,776 lstats and 28,301 listdirs	~15 minutes	16+ hours to 170 sites
Home volume	20–40 seconds	358,126 lstats and 19,535 listdirs	~15 minutes	15 minutes

Table: Scan and copy times after deploying the Starfish Scan Accelerator for VAST.

The overall reduction in scanning times and copy times for backing up these two volumes meant that backing up all seven of the university's volumes could now be completed in 24 hours. Backup times now meet the standard that Becky and her team set for the institution, and the team's time is free to focus on more strategic storage problems.

> "The Starfish Scan Accelerator for VAST was definitely a game changer for us. It used to take seven or eight hours just to scan our project space. Now, with the accelerator, we're talking about a scan that runs in a matter of seconds, leaving us much more time to do the actual backup."
>
> **— Becky Ligon, Storage Architect, Clemson University Research Infrastructure**

© 2025 Starfish Storage Corporation. All Rights Reserved.

About Clemson University

Clemson University is a public research university in Clemson, South Carolina. Its Research Infrastructure team supports high-performance computing and data storage for research computing across campus, managing large-scale volumes on Starfish alongside VAST Data storage systems.

About Starfish

Starfish provides the world's most scalable and versatile file management platform, for organizations running the largest and most complex storage environments: pharma and biotech, semiconductor, aerospace, manufacturing, energy, finance, and academic and government research. The platform features an Unstructured Data Catalog for indexing and extracting metadata for every file, combined with Starfish's fully programmable Automation Engine for automating data flows. Starfish Remote Agents deliver high performance, parallelized file system scanning and data movement at scale, including native Scan Accelerators for a variety of storage platforms. Starfish Zones enable data owners to see and act on their own data, and storage admins to see and account for who owns which files. The Starfish platform gives organizations control of their files in large-scale, mixed-vendor environments that span thousands of shares, tens of billions of files, and hundreds of petabytes so they can cut storage costs, simplify operations, and enforce lifecycle, security and compliance policies. Starfish customers manage 2.5+ exabytes today. Starfish was founded in 2013 and is headquartered in Waltham, MA. Learn more at starfishstorage.com.

© 2025 Starfish Storage Corporation. All Rights Reserved.