CASE STUDY

Hormel Institute UMN preserves cancer research data and breakthrough findings with Spectra Logic solution

> " We have to store a lot of data for long periods of time. A single microscope can capture 10TB of data in as little as a few hours, and a published breakthrough on cancer research can take anywhere from a few years to a couple of decades. The BlackPearl and tape solution from Spectra provides a scalable and cost-effective way to store multiple data workflows with varying retention requirements. "

Jeffrey McDonald, director of IT, Hormel Institute at University of Minnesota

Spectra BlackPearl and T950 Tape Library

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA   MAYO CLINIC

About the Hormel Institute

The Hormel Institute, University of Minnesota (UMN) is a leading cancer research department of UMN and part of the Masonic Cancer Center, an NCI Designated Comprehensive Cancer Center. Collaborative research partners include Mayo Clinic, MD Anderson Cancer Center, Columbia University, University of Arizona and more renowned centers worldwide. The Hormel Institute tripled in size in 2008 and again doubled in size in 2016. Currently the faculty, researchers and staff are comprised of 140 leading cancer research scientists and 20 cancer research sections. Over the next few years, The Hormel Institute UMN will add another 130 new faculty, research and staff jobs as part of its expansion as it continues to perform world-class research in the quest to prevent and control cancer.

The Challenge

Started in 1942, the Hormel Institute has a seven-decade history of making significant scientific discoveries aimed at improving the health of the world. Its highly successful research scientists are focused on accelerating answers to cancer by determining the basic molecular mechanisms of cancer development and finding better ways to prevent, detect and treat the disease. Their research efforts generate over half a petabyte of data per year via two primary avenues: cryoelectron microscopy and bioinformatics research.

Cryogenic electron microscopy (cryo-EM) is a technique applied on samples cooled to cryogenic temperatures to determine biomolecular structures at near-atomic resolution. The process uses electron microscopes to create high-resolution images of molecules, which scientists render three-dimensional to identify ways to make cancer drugs more effective. Hormel is home to one of the few cryo-EM microscopes that are active in the United States (and the only one in Minnesota). The Hormel Institute must store these processed images for the life of the researcher or according to the appropriate grant terms. Given the volume of data produced by their cryo-EM microscope, Hormel can generate up to 10TB of data per day. They wrestle with further compressing this massive image data as well as dealing with the large quantity of small files generated in each microscope run.

In addition, Hormel also conducts bioinformatics research with data provided by National Institutes of Health (NIH) and other research organizations. In this

The Hormel Institute is located in Austin, Minnesota.

research, Hormel scientists analyze large data sets from DNA sequencers to find patterns and commonalities that may help identify a cause or facilitate treatment for cancer. Specific rules around data retention and deletion vary depending on user agreements.

Finally, their primary disk storage was undersized and consistently running above 85% utilization, meaning they were constantly running out of disk space. The institute lacked an appropriate archive environment that would allow them to store data indefinitely and recall it seamlessly.

The Solution

The Hormel Institute chose to implement a two-tier workflow with storage technology from Spectra Logic, deploying a Spectra® BlackPearl® Hybrid Storage Platform with 10 8TB HDDs and a Spectra® T950® Tape Library with four LTO-8 tape drives. The object-based BlackPearl provides a single portal to multiple storage targets in a Perpetual Tier of storage. The Perpetual Tier complements the organization's existing Primary Tier of storage and provides lower-cost storage options, including public/private cloud, NAS, object storage disk and tape.

The implementation allows Hormel to effectively balance access, cost and long-term preservation objectives. After users have finished operating on the data on an HPC system and the data needs to be archived, GNU Tar multivolume archives are used to aggregate it into 50GB chunks before the data is moved to BlackPearl via a 40 GigE network. Globus research data management software connects Hormel's primary NFS-mounted storage system to BlackPearl and releases batch transfers of data. File metadata and tags are retained for continuous management of the data across its lifecycles. BlackPearl writes two copies of data to the T950 tape library. The T950 seamlessly compresses the data by 20%, providing Hormel with a reduction in size at no additional cost or time. Data is written using the open Linear Tape File System (LTFS) file format, a common tape format that presents a standard file system view of the data to deliver a future-proof architecture without vendor lock-in.

> "The economics just make sense. Our investment in the Spectra solution gives us a total of 3.1PB of storage capacity at a cost of under $10/TB per/year. And this number takes into account depreciation cycles, annual support costs and a 5-year and 15-year product lifetime for BlackPearl and the T950 respectively."
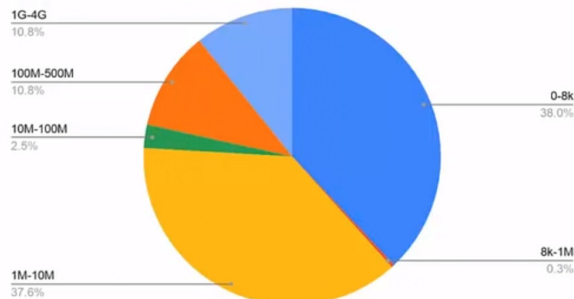
Jeffrey McDonald, director of IT, Hormel Institute UMN

Typical Directory structure from Run

Data Stats
- 7.2 TB
- 103 directories
- 35,860 files
- Average File size is 220MB
- Standard Deviation is 623MB

Files Sizes in Representative Data Set

- 1G-4G: 10.8%
- 100M-500M: 10.8%
- 10M-100M: 2.5%
- 1M-10M: 37.6%
- 0-8k: 38.0%
- 8k-1M: 0.3%

This graph illustrates a real data set from a cryo-EM microscope run at the Hormel Institute.

The combination of BlackPearl and tape delivers a highly-scalable, cost-effective architecture that supports the Hormel Institute's high-impact research. Their system has 3.1PB of storage capacity at a cost of under $10/TB per year, support cost included. Their system is fast enough that it could transfer an entire petabyte of data in about 100 hours, with an average transfer rate well over 300MB/sec.

Environment Snapshot

- Spectra BlackPearl Multi-Purpose Hybrid Storage Platform
- Spectra T950 Tape Library
- Four LTO-8 tape drives
- SpectraVision camera
- BlueScale® software with standard encryption
- Globus data management software
- Java CLI

Why Spectra?

- Scalable, reliable and adaptable to growth
- Open standard technology to prevent vendor lock-in
- Cost-effective technology
- Seamless workflow integrations

Solution Recap

Spectra BlackPearl – BlackPearl provides an object-based, hybrid storage platform that easily and cost effectively scales up as object and file volumes grow; adapts as operational requirements change; and enables easy synchronization of data between on-prem and cloud storage.

Spectra T950 Tape Library – Designed for data integrity, data security and reliability, the T950 tape library affordably scales in capacity and throughput, and supports multiple generations of current and future tape formats.