Storage Options

There are multiple options for storing research data available at UAMS. Here are a few:

Storage Suitable for Identified Data

The UAMS IT department does offer storage options for research data that might have identifiable data, such as patient information regulated by HIPAA, or student information regulated by FERPA. Because of the sensitive nature of such data, this storage is managed by UAMS IT to comply with various UAMS policies and procedures. Identifiable data must be de-identified before it can be processed on Grace, since Grace does not comply with all UAMS policies for HIPAA and FERPA. Naturally, data that does not fall under HIPAA or FERPA rules may also be stored in these systems.

Research NAS

The UAMS IT department maintains a high-performance campus Network Attached Storage (NAS) system. One section of that NAS is dedicated to research storage. The research area is logically separated from the clinical areas, to minimize interference between the two. The NAS system replicates data in 3 locations to guard against data losses. One is in the Primary Data Center. Another is in the Secondary Data Center on the opposite side of the UAMS Campus. The third is at the NW campus in Fayetteville. UAMS IT makes periodic snapshots in lieu of full backups. The snapshots generally are taken nightly, and held for 30 days. The Research NAS is considered HIPAA compliant, hence can be used to hold identified data.

The Research NAS is divided into isolated spaces. Having access to one space does not give access to other spaces. Each space can be mounted separately, if permissions allow. Research NAS spaces can be mounted from any systems that support NFS inside the UAMS Campus firewall. The Research NAS cannot be mounted outside of the firewall.

To reserve space on the Research NAS, or to mount a space that you have access rights to, please contact UAMS IT directly. UAMS IT currently charges $1.50 per GB (subject to change) to reserve space on the Research NAS. The reservation lasts 5 years. After that. to continue holding that space a researcher would have to purchase another 5 year increment at the then going rate. The charge is commensurate with the actual cost of purchasing the storage hardware needed to maintain the 3 replicas of the space, including the redundant disks for error coding that helps protect the data from hardware failures.

Box

UAMS IT has also negotiated with box.com to provide cloud storage for UAMS faculty and staff. The Box storage is considered HIPAA and FERPA compliant, and can be accessed anywhere on the internet, assuming the user trying to access it has been given permission. Being pure cloud storage, access to data on Box is not has speedy as locally stored data, or data stored on the Research NAS. Because of the lower performance, storing data on Box may not be appropriate for projects managing large amounts of data.

Box does provide means for sharing data, both inside and outside of UAMS, via cloud interfaces (e.g. through a browser). Of course, a Box user should be careful that they do not share HIPAA or FERPA data inappropriately.

Box provides the Box Drive app for Mac and for Window that allows a user to access cloud based directory trees. Box Drive optionally sychronizes locally stored directory trees with Box, giving a relatively simple and automatic method for backing up those directories, with the ability to use them offline. In addition to Box Drive, there are dozens of apps that can be used to organize and deal with data stored on Box on a variety of devices.

Please contact UAMS IT for information about setting up a Box space and about Box storage quotas.

Storage Not Suitable for Identified Data

There are other options available on campus for storing data, which have not been vetted by UAMS IT and IT Security Offices as being safe for sensitive data, such as data that must comply with HIPAA or FERPA. As such, these systems should never be used to store data that contains student or patient data, and care should be exercised when storing data that needs extra protection. The researcher using these storage systems is responsible for assuring that data stored in these systems is not being stored inappropriately.

Local Storage

User do have the option to store data on their own systems, for example the local disk on their laptop or workstation, or on a local storage system (e.g. a departmental NAS). In most cases, these storage systems are not being managed by IT in a fashion compliant with UAMS policies for HIPAA and FERPA, so care should be taken by users to not store identifiable or sensitive information inappropriately. In general, sensitive data should use one of the UAMS-supplied storage systems, and any copies that are synchronized or copied to local storage should be on protected, encrypted drives according to UAMS policies. In addition, users are responsible for dealing with their own backup and disaster recovery needs.

We strongly suggest that local storage should not be the primary location for research data. For safekeeping, the primary copy should be on one of the other storage systems, depending on the sensitivity of the data and research needs, and appropriately backed up. The local copy should be considered a scratch copy, only used for local processing and analysis of data.

Cloud Storage

There are multiple cloud-based solutions for storing data, such as iCloud, OneDrive, Google Cloud Platform, Amazon Web Services, and IBM Cloud Object Storage (formerly Cleversafe). Use of these systems is at the researcher's discretion, and the researcher is responsible for complying with UAMS IT and UAMS policies. Often the researcher is also responsible for arranging the financials for using these storage systems. Keep in mind that even with cloud-based storage, the researcher should be concerned with backup and disaster recovery requirements. Not all cloud storage services guarantee the safety of data stored in their services.

Research Object Storage System (ROSS)

The UAMS HPC team, with the assistance UAMS IT on the hardware, manages a 4.2 PB Research Object Storage System (ROSS). It is based on the Dell/EMC Elastic Cloud Storage (ECS) product, and offers multi-tenant object store. In philosophy it is somewhat similar to other cloud-based storage options, except that it is locally managed. A coming hardware refresh will allow a portion of ROSS storage to be replicated in NW Arkansas courtesy of our partners at the University of Arkansas Center for High Performance Computing. ROSS is now part of the Arkansas Research Computing Cooperative, which supports the Arkansas Research Platform project.

ROSS is visible inside the UAMS firewall, but is not currently directly visible outside of the UAMS firewall. Essentially any system on the UAMS network has access to ROSS. ROSS supports multiple protocols for storing and retrieving data, including S3, OpenStack Swift, NFS, and HDFS. There are dozens of tools available that can move data to and from ROSS. One of our favorites is rclone, which is universally available across multiple operating systems.

The UAMS HPC team also maintains a data transfer node (DTN) that uses ecs-sync to move data between ROSS and Grace's cluster storage. The ecs-sync program can move data faster while using less compute overhead than rclone. But some might find ecs-sync's command line utility a bit more difficult to use. Using the DTN is a good option for moving data to the Grace Cluster Storage prior to running a job that analyzes that data, and then again after the job completes to move the results and changed files back to ROSS.

The DTN also includes an Aspera High Speed Transfer Server (HSTS) that is capable of moving data in and out of ROSS at near wire speed, including to locations outside of UAMS. In addition to the HSTS, we have an Aspera Faspex package manager that facilitates data sharing at HSTS speeds to outside collaborators with just their e-mail addresses.

While storage on ROSS is highly reliable, it is still a good idea to maintain a backup copy of data on alternate technology.

Before joining the ARCC, the College of Medicine at UAMS, who originally purchased the system, was suggesting that reserving storage on ROSS would cost $70 per TB for a 5 year reservation, which is significantly less expensive than commercial Cloud Storage, especially when taking into account the networking and data egress fees that many cloud storage providers charge. Note that this $70 charge was less that 30% of the actual cost of the storage hardware. In other words, use of the storage was heavily subsidized by UAMS IT. This charge, among other policies governing the use of ROSS, is being re-evaluated by the ARCC steering committee. So when requesting a storage space reservation on ROSS, keep in mind that there may be a charge to use it, depending on what the ARCC Steering Committee decides. But until the charge question is settled, storage may be reserved by simply sending a request to the HPC administration team (mailto:hpcadmin@uams.edu).

GRACE Cluster Storage

Most HPC users are already familiar with Grace's 2PB internal cluster storage. This is a highly parallel, high performance, distributed, shared file system based on IBM's Spectrum Scale software running on DDN's Gridscaler hardware. Grace's cluster storage is more expensive than other types of storage maintained by the HPC team.

Grace's cluster storage, both local and shared, is considered scratch storage, hence is not backed up. Each user on Grace is responsible for their own backups. Users should employ some other storage option to maintain the primary copy of their data.

Grace's storage system is not intended for long term storage of data. Since Grace's internal cluster storage is limited in size, users should move any data that is not in active use to one of the other storage options for safe keeping. Currently we rely on self-policing to keep Grace's storage free of inactive data, but we may be forced to change that policy and begin automatic deletions of data, which is another reason why users should start now to get into the habit of keeping primary copies of data somewhere else.

OURRstore (Oklahoma University and Region Research Storage)

OURRstore is an NSF funded project (circa $1 million) that set up an Exabyte scale cold archive for research data. We are eligible to use this storage resource for certain classes of research data. It is a robotic LTO tape handler front ended by a disk-based cache system. Access to data in OURRstore is through the cache system and a set of Data Transfer Nodes (DTNs). LTO tape has an expected archive life of 15 years, with an error rate orders of magnitude lower that spinning hard drives. Being conservative, the OURRstore team suggests planning on an 8 year lifetime. The OURRstore system will always make at least 2 copies of data, one that stays in the robot, and a second that is sent back to UAMS as an offline, offsite backup of the data. An optional third copy can also be made that is stored offline in Oklahoma. This is the second LTO-based archive that the University of Oklahoma team has set up. The first one (Petastore) has been in operation for almost a decade, and has proven extremely reliable and safe.

A cold archive is for data that is not in active use, but that should be kept long term. A common mantra with OURRstore is "Write Once, Read Seldom". It is not a problem getting data back from OURRstore. It is readily available using various network protocols (sftp, Globus Connect, etc.), and can be retrieved at any time. Data just recently stored will start streaming back nearly immediately, since it likely is still in the OURRstore disk cache. Older data might require a few minutes to fetch the data from the LTO cartridge before it starts streaming back. The underlying technology streams data at about 300 Mbps for the current generation of media (LTO-7M).

Since it is regionally accessible, OURRstore is not appropriate for identified or sensitive data.

In addition, data destined for OURRstore must meet strict size requirements. OURRstore will accept file in size from 1 GB (minimum) to 1 TB (maximum). The 'sweet spot' are files between 10 GB and 100 GB for best storage efficiencies and performance. Typically we suggest consolidating smaller files into larger archive files using compressed, encrypted tar or zip prior to sending to OURRstore.

In addition to long, stable storage life, a major advantage to OURRstore is cost. Since NSF funded the hardware and setup costs, and OU covers the administration and ongoing management costs, users of OURRstore only need to purchase media and pre-paid return shipping labels to put data in OURRstore. At recently quoted prices, which continue to edge downward, that works out to under $20 per uncompressed TB for two replica copies, or $30 per uncompressed TB for triple replica copies kept at different locations, with an expected life of at least 8 years, possibly longer. Depending on the data, compression of 2x or more is typical, cutting those costs in half. This is a tiny fraction of what cold archive in the cloud, such Amazon deep Glacial, costs. This is the least costly option for stable, long-term storage of data.

If you are interested please contact the HPC admin team (mailto:hpcadmin@uams.edu). The system is available now.