SOLUTION BRIEF

# XENDATA STORAGE FOR GENOMIC DATA

XenData

### Active Archives

Active archives based on data tape and disk provide secure long-term storage for genomic data. For organizations that have up to a few hundred terabytes of data, cloud object storage is cost-effective, minimizing the need for capital investment. But for larger storage requirements, the lowest cost approach overall is to invest in a scalable tape-disk system. After the initial purchase, the storage capacity can be grown very economically.

## OVERVIEW

The high data rates achieved with Next-Generation Sequencing (NGS) methodologies have resulted in high growth in the volume of stored genomic data. Cloud object storage from providers such as Amazon Web Services, Google and Microsoft is a convenient approach to meeting this growing demand for data storage. However, the cost of cloud storage grows approximately linearly with the volume of stored data, making cloud storage uncompetitive above a few hundred terabytes. XenData storage systems based on robotic libraries with high-capacity data tape and disk cache deliver solutions that scale very cost-effectively.

## CLOUD OBJECT STORAGE – COSTLY AT A SCALE

Cloud-based storage for NGS data is easy to implement, requires minimal fixed costs and, from a technical perspective, it scales well. Unfortunately, the costs also scale linearly and many cloud providers charge fees in addition to the monthly storage cost. These fees include egress and rehydration costs which are explained below. They are difficult to calculate accurately[1] and can make it expensive to leave a cloud provider.

Egress fees, sometimes called bandwidth charges, are the costs incurred every time you move data out of the cloud. Most cloud providers do not charge for ingress of data to their cloud. Egress fees are typically many tens of dollars per terabyte of downloaded data. For organizations with more than a petabyte, it is very costly to migrate from the cloud provider. For every petabyte, the egress fees will be many tens of thousands of dollars.

Rehydration fees are specific to cloud providers that offer lower cost storage tiers. Examples of lower cost storage tiers are AWS Glacier tiers and the Microsoft Azure Archive tier. Moving your data to these tiers lowers the monthly storage cost but it means that the data is no longer immediately available. To become accessible, it must be migrated to an 'online tier' which takes from minutes to many hours. This migration process is termed 'rehydration' and there is a cost associated with it. For example, the cost to rehydrate every petabyte of data from the Azure archive tier is $20,000 or more. This cost is in addition to the egress fees.

Even though it is often difficult to estimate the costs, using cloud storage for up to a few hundred terabytes of genomic data is usually a good choice. However, for higher data volumes, a XenData tape-disk storage system provides lower cost and, above one petabyte, the cost savings are substantial.

## TAPE-DISK SYSTEMS SCALE COST-EFFECTIVELY

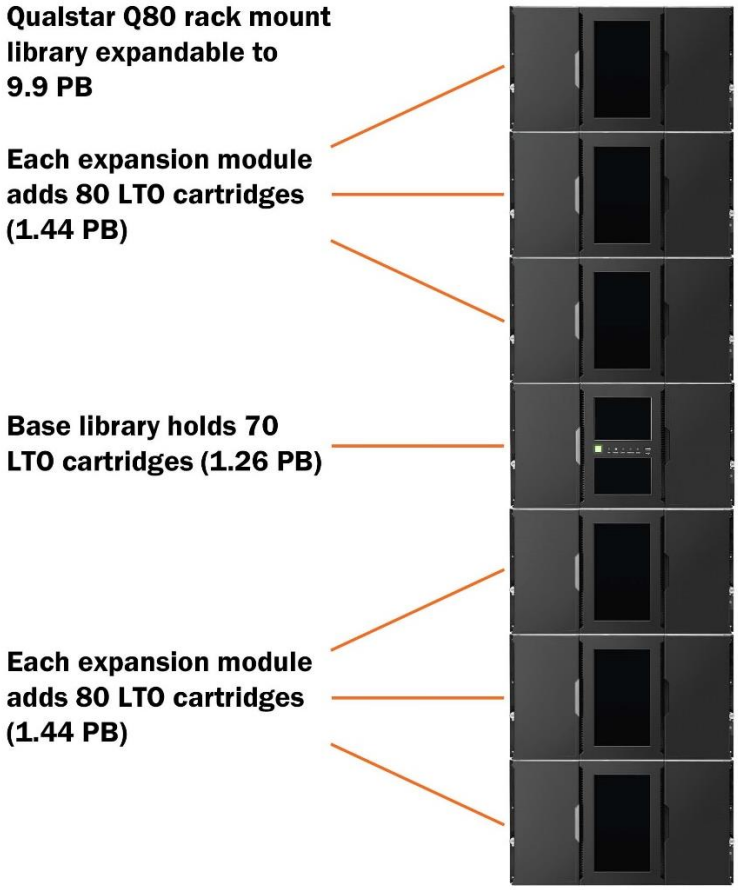
For organizations that have large volumes of genomic data, an on-premises data storage system that uses a robotic tape library and disk cache is attractive because it scales cost-effectively to many petabytes. The tape library and disk cache are managed by intelligent software running on a server to easily integrate with the existing on-premises IT infrastructure.

The dominant data tape format is LTO (Linear Tape Open). It is currently in its ninth generation with a twenty-year history. Over that period, new generations have been introduced regularly with increasing capacities[2]. Each LTO-9 data cartridge holds 18 TB natively, or more with compression, and LTO-9 tape drives provide high transfer rates up to 3.2 Gbps. The data cartridges are extremely durable with a minimum data life of 30 years.

Robotic LTO libraries are available from many manufacturers including Dell, HPE, IBM, Oracle, Qualstar, Quantum and Spectra Logic. Most library manufacturers offer models which scale by adding physical expansion modules, thereby increasing the capacity of the library. LTO drives may also be added to increase the overall data transfer rate.

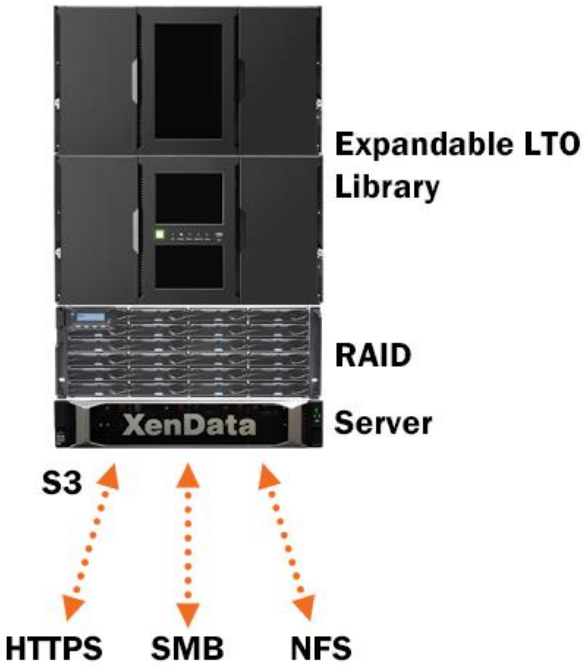
After the initial purchase, a scalable LTO storage system can easily be expanded as illustrated in the image below which can initially be implemented with a 1.44 PB capacity and then expanded progressively to 9.9 PB.

**Example Expandable LTO Library**




## XENDATA TAPE-DISK SYSTEMS

XenData offers a range of storage servers that manage robotic LTO libraries and disk cache. The X100 is suitable for organizations that need storage solutions which scale easily to many petabytes. It is available as a single server or cluster[3] which manages one or more LTO libraries and includes a disk cache up to 280 TB. Its intelligent software creates a scalable file server that integrates via NFS or SMB network protocols. Additionally, it can be configured as private cloud storage with an S3 interface and accessed remotely via secure HTTPS.




When accessed via an NFS or SMB network share, writing and reading is like writing to and reading from a disk-based file server. Files written to the X100 are initially stored on the disk cache and then the intelligent XenData software writes the files to the designated LTO cartridges in background. When reading a file that is not cached, there is typically a delay of about 2 minutes while a tape cartridge that contains the file is loaded into an available LTO drive.

When accessed via the S3 interface, writing and reading is similar to writing to and reading from public cloud storage. Writing starts immediately but with reading non-cached data there is a delay of about 2 minutes before the file starts to be restored.

The X100 has a lot of advanced functionality, including:

- Replication of LTO cartridges, automatically creating a data protection copy
- Disk cache with a capacity up to 280 TB and highly configurable disk retention policies
- Option to synchronize the file-folders on any accessible network share to LTO

The X100 is available as a single server or as a server cluster for high availability and more information is available in reference 3.

## Contact Us

**XenData**

Address: 20005 State Highway 88
Suite D, Pine Grove, CA 95665
Phone: +1 925 465 4300
Email: xendata@xendata.com
Website: www.xendata.com

**XenData Europe**

Address: Sheraton House,
Castle Park, Cambridge CB3 0AX, UK
Phone: +44 1223 370114

# REFERENCES

1. N.Krumm, N. Hoffman, 'Practical estimation of cloud storage costs for clinical genomic data', Pract Lab Med, PMID 32529017, (2020)
2. https://www.lto.org/roadmap/
3. https://xendata.com/Assets_Products/X100_Product_Brief.pdf

***About XenData***

*XenData is a global provider of cutting-edge active archive systems based on LTO data tape and hybrid cloud. The company has over 1500 large LTO storage systems installed in over 90 countries which it supports from its facilities in Walnut Creek, California, USA and in Cambridge, UK.*