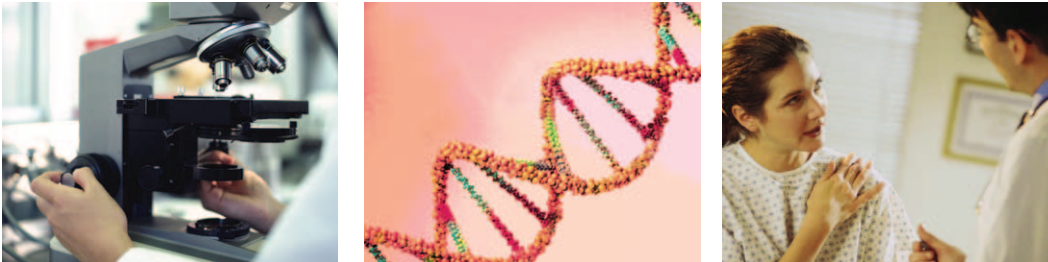




# Datalink Implements Customized, Multi-faceted Solution to Protect University's Valuable DNA Research Data



## THE CUSTOMER

Washington University\* is an independent research university that houses the Genome Sequencing Center, which focuses on the large scale generation and analysis of DNA sequence. The center is a leader in The Human Genome Project, a 13-year research effort to sequence and map all of the genes of the Homo sapiens species. The center also sequences the genomes of other species in a quest to better understand the human genome sequence and advance the study of biology.

In processing genomic data, researchers at the center load DNA samples into DNA sequencing machines, which in turn create files that are loaded into an Oracle RAC database. The database holds billions of files—about a terabyte of data that is growing at a phenomenal pace—that are used around the clock to conduct DNA analysis. If the database were to go offline, the entire lab facility would shut down. Beyond that, if the data were lost, the DNA sequencing research being conducted at the facility would be severely compromised.

## THE ISSUES

The Genome Sequencing Center was feeling increased pressure on its backup infrastructure. Failing backup streams were causing its tape-based backup methods to be unreliable and untrustworthy. The backup windows had grown longer and because of media write errors, it was increasingly difficult to obtain complete backups of the Oracle RAC database. With estimates that losing even a small fraction of the data could cost hundreds of thousands of dollars, ineffective backups were simply not an option. There was also concern among the IT staff that if they needed to restore the database, they would not be able to do it in a timely fashion.

## INDUSTRY

Research and Education

## LOCATION

St. Louis, Missouri

## KEY CHALLENGES

Increased pressure on backup infrastructure was affecting backup and recovery operation

## SOLUTION

Multi-faceted solution that uses disk-based backup, off-host backup, point-in-time copies, and clustered network-attached storage with replication technology for data protection and high availability.

## KEY SOLUTION COMPONENTS

Hitachi Data Systems  
ShadowImage software,  
Network Appliance fabric-attached storage systems,  
Network Appliance nearline storage system, VERITAS  
NetBackup software

## DATALINK PROFESSIONAL SERVICES

- Analysis
- Design
- Implementation
- Support

## BENEFITS

Mission critical data is highly available and protected in the event of a disaster or outage. Data can be quickly restored, assuring minimal impact on users.



Information Means The World.

## THE NEEDS

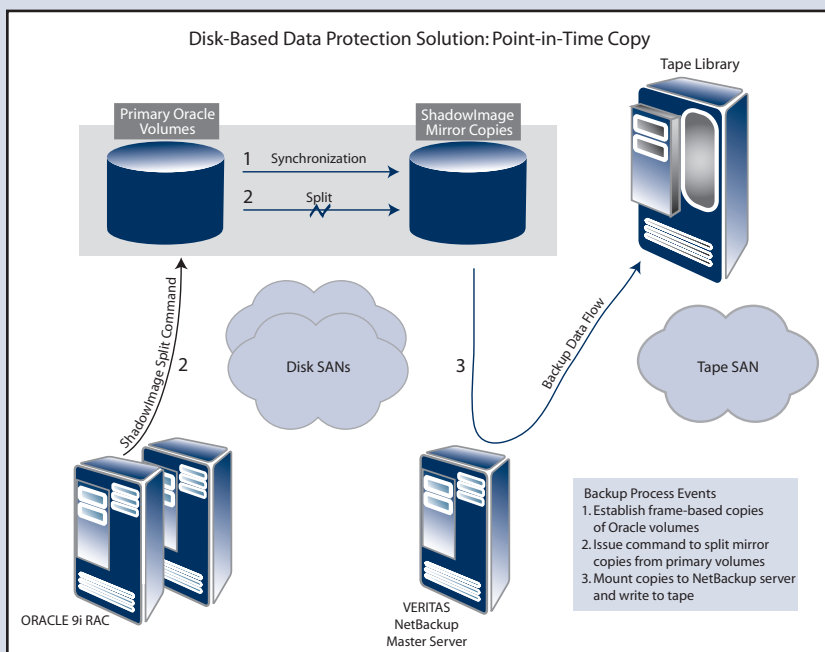
The Genome Sequencing Center turned to Datalink to help design a solution that would quell its concerns. Datalink consulted with the staff at the center and came up with the following objectives:

- Reduce the backup window
- Reduce the time to restore the database to under two hours if there was a problem
- Provide confidence that the valuable data was protected

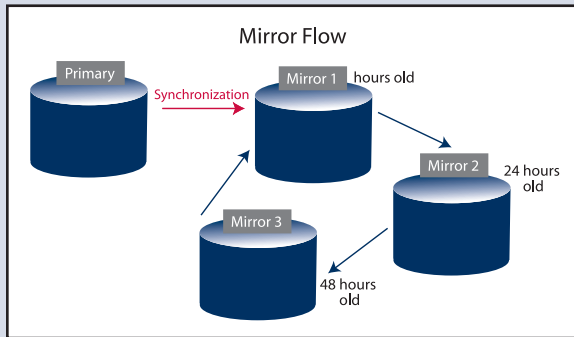
In helping Washington University achieve these objectives, Datalink also wanted to enable the organization to leverage its existing technology and expertise. Up until this point, the Oracle RAC data was stored on a Hitachi Data Systems (HDS) disk array with VERITAS NetBackup for backup software.

## THE SOLUTION

Using Hitachi Data Systems ShadowImage software, Datalink implemented a disk-based solution in which backup copies of the database are created in the form of mirror copies that reside on the HDS disk array. To meet the center's service level agreements, three mirror copies are made: one holds data that is a few hours old; one holds data that is 24 hours old; and one holds data that is 48 hours old. The mirror copies are split from the primary database and mounted on a different server where they are backed up to tape for archival purposes. The result is faster, more accurate backups that can be restored much more quickly.



*For its Oracle RAC database, the Genome Sequencing Center utilizes a disk-based solution in which multiple mirror copies of the data are created for backup purposes. The mirror copies are then split from the primary database and mounted on a different server where they are backed up to tape for archival.*



*To meet the Genome Sequencing Center's service level agreements, three mirror copies of the Oracle RAC data are made: one holds data that is a few hours old; one holds data that is 24 hours old; and one holds data that is 48 hours old.*

Here is how the solution works:

- The oldest of the three ShadowImage mirror copies is brought into sync with the database.
- A script begins the backup operation by suspending the Oracle RAC database for approximately five minutes. This usually occurs in the middle of the night.
- The ShadowImage mirror copy is split, creating a point-in-time copy of the Oracle RAC database.
- The Oracle RAC database is unsuspending.
- The ShadowImage copy is mounted on a different host and this point-in-time copy is backed up to tape.

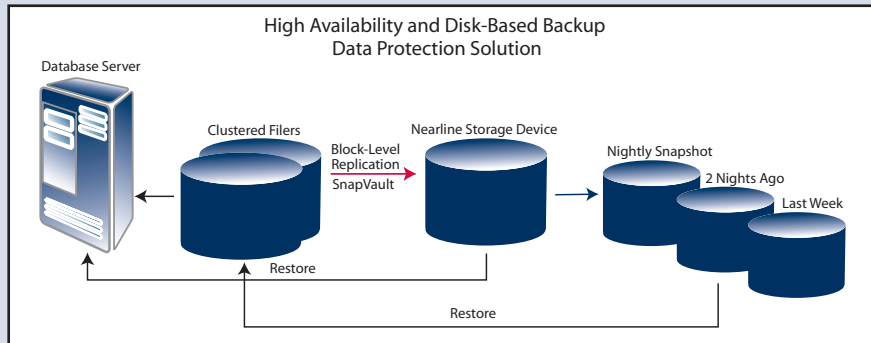
If a problem does occur and the center needs to restore data, it can access any one of the three ShadowImage copies that reside on the disk array.

## **MEETING OTHER NEEDS AS WELL**

As a result of the massive amounts of data generated through DNA processing, the Genome Sequencing Center was also facing issues with backing up the VERITAS NetBackup catalog. With a catalog that already consumed 400 gigabytes and was growing at an astronomical rate, the backups were taking too long and it was difficult to get them on to tape.

To alleviate this problem, Datalink implemented a Network Appliance disk-based nearline storage system to provide disk-based backups of the VERITAS NetBackup catalog. Nightly and weekly snapshots on the nearline device enable multiple copies of the catalog to be available if the need arises. By using a disk-based approach, the catalog can be saved and restored in a much shorter time frame.

To round out the storage infrastructure needs at the Genome Sequencing Center, Datalink also implemented a high availability and data recovery solution for an Oracle database



*The Genome Sequencing Center uses clustered network-attached storage with replication technology for data protection of a second Oracle database.*

that serves as a warehouse for gene sequencing data after it has been processed. Because the IO rate of this data is not extremely high, Datalink recommended network-attached storage in the form of Network Appliance's fabric-attached storage systems. The warehouse data—along with NFS data—is stored on two clustered systems. If one fails, the other one takes over—transparent to the end user. With downtime estimates of \$10,000 an hour (unrecoverable), a high availability solution of this magnitude was a must.

For recovery, NetApp SnapVault software takes a snapshot of this data and moves it to the nearline storage device where it can be quickly accessed for restore purposes. There are a couple of options for recovery. For quick restore, IT staff can make the volume on the nearline system read/write and mount it directly to the application server. Or, they can restore the volume back to the filer cluster and run the Oracle logs to get to the desired point in time. For additional protection, tape copies exist as well.

## **END-TO-END SOLUTIONS**

With cutting-edge research being conducted at the Genome Sequencing Center, it is only fitting that the center has the latest technology available for disaster recovery and high availability.



Information Means The World.

**CORPORATE HEADQUARTERS**  
8170 Upland Circle  
Chanhassen, MN 55317-8589  
(800) 448.6314  
www.datalink.com