

**DataDirect Networks**  
**Optimized Storage Solutions for**  
**Life Sciences Data**

## **Table of Contents**

---

<b>Introduction</b>	<b>1</b>
<b>Developments in “Omics” Data Discovery</b>	<b>1</b>
<b>The BioInformation Data Deluge</b>	<b>2</b>
Bandwidth Delivery	2
Transactional Capability	2
BioData Volume Capacity	3
<b>DataDirect Networks: Worldwide Extreme Storage Provider</b>	<b>3</b>
<b>Introducing: The GridScaler Networked File Storage System</b>	<b>5</b>
<b>Summary</b>	<b>6</b>

---

## Introduction

Over the past 10 years, advances in life science technology enabled a revolutionary understanding of biological phenomenon & biochemical systems. These advances in technology have ushered in the dawn of advanced drug discovery, technologies and treatments that promise to accelerate cure disease and improve human life quality and life expectancy. Over time, continued advancements will make the benefits of these breakthroughs accessible to more and more of the world's population.

Medical and scientific researchers now have at their disposal extremely powerful tools for profiling and analyzing molecular, proteomic and genomic interactions at the atomic level. Often called "omics" research, these bioinformatics disciplines are rapidly advancing the quality and precision of "omics" tools and the resultant data discovery and analysis techniques. Tied closely to the IT industry, these progressions traditionally track with advances in CPU processing capabilities and parallel data processing methodologies. In 2009, these bio-scientific tools have undergone another step-change advancement in capabilities and research facilities are now finding themselves challenged with the amount and rapidity of data being created by bio-scientific tools.

## Developments in "Omics" Data Discovery

Fueled by investments from the pharmaceutical and scientific endowment communities, researchers pushed the capabilities of bio-analysis equipment to drive a deeper understanding of biological phenomena and have ushered in the dawn of massively parallel analyses. This drive has resulted in the introduction of revolutionary bio-analysis equipment including:

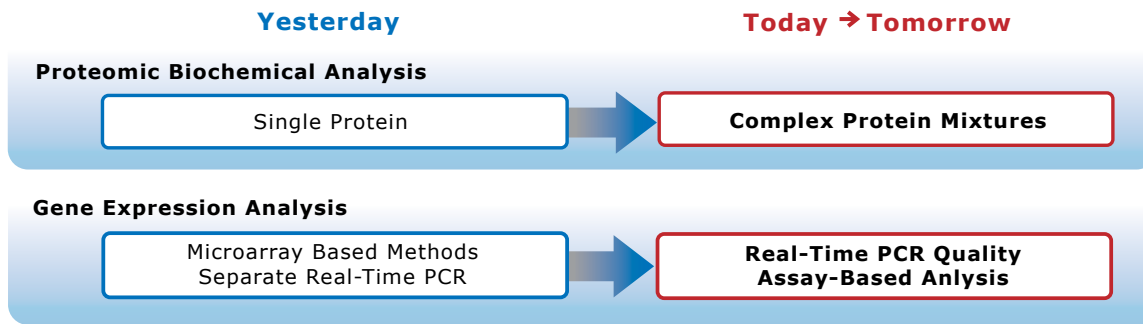
- DNA microarrays for high-throughput genetic and epigenetic sequencing and resequencing
- High-throughput assays for RNA analysis for gene expression modeling
- Mass spectrometers for high-throughput protein research
- World-wide databases for correlating and comparing bio-information to identify genetic & proteomic disease patterns and identify biomarkers, or detectable bio-molecular symptoms, which can be used in disease identification and early disease identification technology



*Next-generation gene sequencing technology, such as Applied Bioscience's SOLiD® high-throughput sequencer, can generate terabytes of data per run and present new challenges in traditional storage environments which are typically not designed to hold petabytes of data and deliver rapid ingest and data analysis performance.*

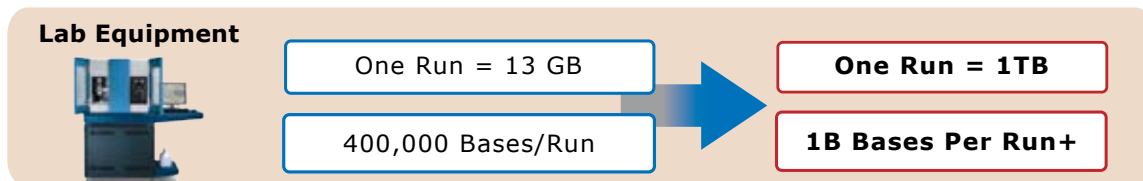
With these tools, scientists use statistical analysis methods and high-performance cluster computing to increase understanding of disease causes, cures and therapies. This understanding leads to eliminating clinical drug trial risk and accelerating the development of personalized medicine to prevent and cure disease.

As the capabilities of these scientific tools increase, scientists are now able to evolve their understanding of biological interactions, this can be seen in many step-changes in measurement capability, including:



## The BioInformation Data Deluge

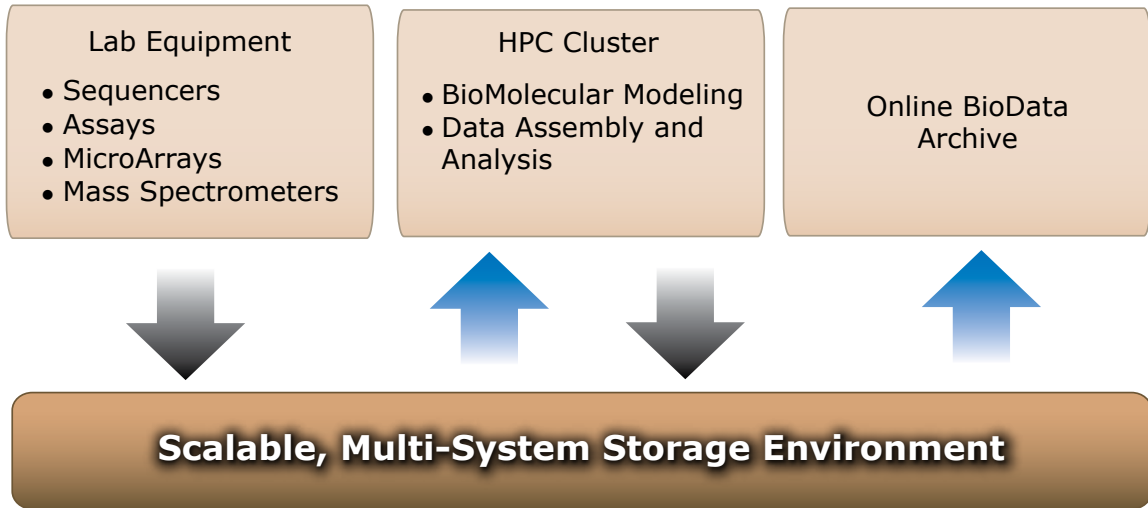
Concomitant with the advances in bio-analysis equipment, and the heightened level of understanding that these machines have ushered in, is the data creation rates which are an order of magnitude greater than they were just 2 years ago. The throughput and output of today's microarrays and spectrometers from companies such as 454, Solexa, Applied Biosystems, Helicos, and Sanger have challenged IT managers with an unprecedented level of I/O delivery and scale-out complexity.



The information creation rates of higher-fidelity bio-analytic equipment have created 3 distinct storage scaling issues in the laboratory:

- **Bandwidth Delivery:** Sequencing centers and spectrometry laboratories are now equipped with system farms that collectively create data at many gigabytes per second. In most cases, this data is ideally stored in a single, scalable storage system to avoid file storage system proliferation and the resulting system management overhead.
- **Transactional Capability:** Once the data is initially ingested into the storage environment, it then needs to be processed and correlated via a cluster supercomputer using popular sequence analysis tools, such as scalable & parallel versions of NCBI's BLAST software. This transactional load has put new demands on file systems that have not been previously seen in bioinformatics as processing capabilities scale from workgroup cluster size (eg. 64 linux nodes in one HPC cluster) to much larger facilities that have 1000s of HPC cluster nodes for bioinformatics analysis.

- **BioData Volume Capacity:** Over time, the ingested bio-information generated from next-gen laboratory equipment needs to be stored online and made accessible for future reprocessing, semantic web availability and additional correlation to future DNA, RNA & protein samples. In large facilities, average online capacity levels have jumped from the terabyte level in 2006 to the multi-petabyte level today. As traditional file storage systems have not been designed to manage petabyte volumes or for data center operational expense (space, power, cooling & system administration; aka OPEX), the cost of managing online archives can be overwhelming to scientific organizations who are not equipped with scalable, efficient IT tools.



*Data access properties within a large-scale bioscience environment.*

To solve these data capacity and performance scaling challenges, IT managers are now looking to next-generation storage technologies with roots in servicing similar workloads within the high-end HPC industry.

## **DataDirect Networks: Worldwide Extreme Storage Provider**

For over 12 years, DataDirect Networks (DDN) has delivered unrivaled bandwidth and optimized storage capacity to the high-end HPC industry. As the life sciences industry evolves to uncover greater amounts of biological data, DDN storage solutions are now becoming the platform of choice for advanced bio-science laboratories. As testament to this, DDN delivered more bandwidth to the November 2008 Top500 than all other storage companies combined.

DDN has specialized in developing storage technology that is purpose-built to resolve the specific challenges associated with unstructured data. With industry leading performance, capacity, energy efficiency and data protection, DDN's storage platforms enable users to scale-out bio-science data infrastructure with far fewer systems, components, administrative overhead and complexity.

Today, leading bioinformatics research facilities have built scalable storage foundations around DDN storage technologies, including:

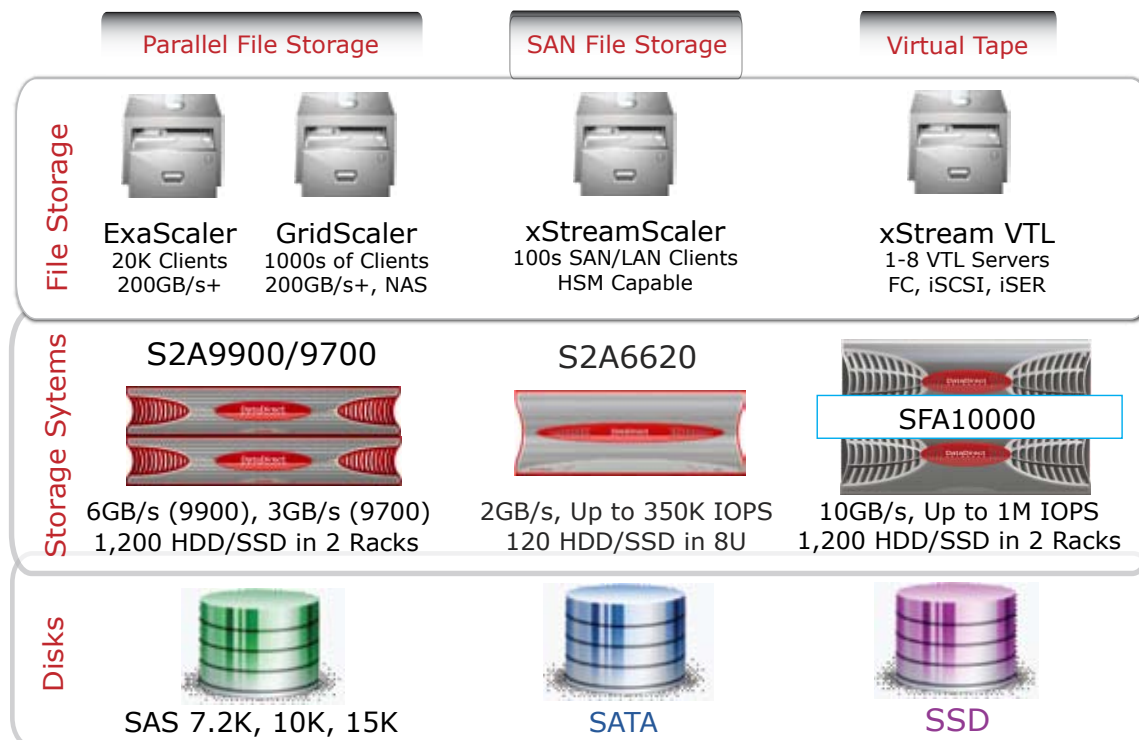
Sangre Wellcome Institute	Bristol Meyers Squibb
Ontario Institute for Cancer Research	The Broad Institute
Ontario Cancer Research Center	European Bioinformatics Institute
Translational Genomics Research Institute	And more...

While founded to solve the HPC I/O challenges found in the world’s largest supercomputing centers – DDN has designed technology which is now demonstrating a proven maturity and broad acceptance in the storage marketplace across a number of data intensive applications, including: backup & archive, media and entertainment, web 2.0, corporate tier 2/3 storage pools and more. Over time, DDN product offerings have evolved to now provide multiple classes of service and capabilities to storage users.

Leveraging DDN’s industry-leading storage arrays, DDN has engineered file storage systems with optimized options for:

- Parallel File Storage: HPC Cluster Storage
- Scalable NAS: High-Performance, High-Capacity NAS Storage
- SAN Storage: Direct I/O for Heterogeneous Streaming SAN Clients
- VTL: High-Speed Virtual Tape Systems with D-MAID Technology

In total, the DDN product portfolio represents a comprehensive file workflow and data lifecycle solution set – designed to simplify production workflows and accelerate and protect applications.



*DataDirect Networks File Storage Product Portfolio*

## **Introducing: The GridScaler Networked File Storage System**

As performance and capacity requirements rapidly grow, the wide adoption of conventional NAS technologies saddles storage managers with volume limitations and performance bottlenecks. Until now, the only solution to these issues has been to purchase and deploy additional quantities of independent NAS systems and distribute the data across independent islands of disparate storage.

DDN has solved tomorrow's file I/O scaling challenges with GridScaler, a networked file storage system. Powered by the world's most scalable POSIX file system, GridScaler is designed to cost-efficiently grow within budgetary and storage management requirements.

Leveraging DDN's award-winning S2A and SFA storage platforms, GridScaler is a scale-out utility that delivers unrivaled data protection, support for 1000s of concurrent GridScaler, NFS and CIFS clients, and provides industry leading storage density and power efficiency. In addition, the GridScaler system supports storage pools, ensuring that data resides and is placed and automatically migrated to the most cost effective tier available.

### *Future-Proof your Research Facility:*

Scaling up to 240GB/s of equivalent read and write performance, the GridScaler features parallelized performance to power up over 2000 concurrent file system clients. This scalability enables a low solution entry-point and much more scalability than competing storage technologies. As such, life sciences data centers can consolidate storage networks and volumes while also being prepared for the next generation of high-throughput bioinformatics data.

### *No-Compromise Data Protection:*

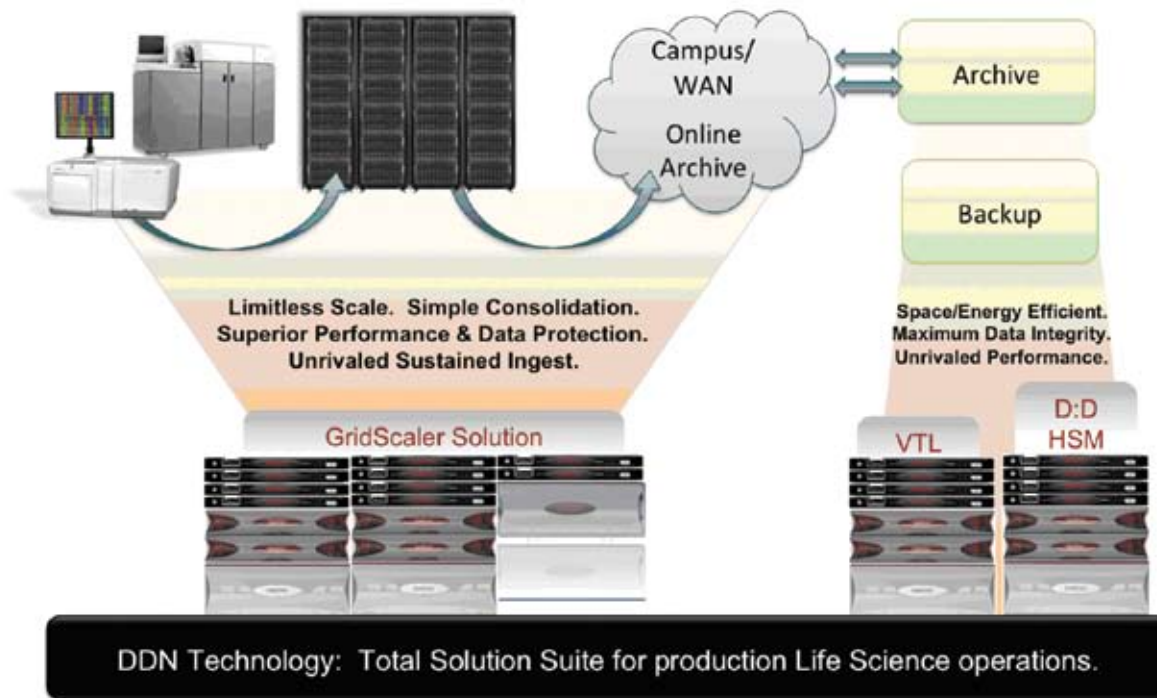
DirectOS, SATAAssure and DirectRAID 6 features are available to safeguard data integrity and ensure availability and integrity while retaining full system performance even during drive error and recovery events. Intelligent snapshot, replication & backup integration are built-in for easy data management.

### *Ready for the BioData Explosion:*

GridScaler holds petabytes of data within a single file system volume. Ultra-dense DDN storage technology ensures that even large-scale capacity growth demands less space/power/cooling than other storage technologies. The GridScaler system stores more capacity per rack than any other NAS storage system available today.

### *Supports Mission Critical Sequencing Runs and Processing Events:*

The GridScaler architecture eliminates all single-points-of-failure and delivers maximum storage uptime. Coupled with DDN's industry-leading performance protection intelligence, GridScaler is a dependable storage foundation that can deliver predictable bandwidth even while undergoing storage self-healing.



## Summary

Today's researchers are challenged with unprecedented storage scaling challenges. Conventional file storage and protection tools are becoming no longer appropriate for meeting the needs of large-scale life sciences facilities.

Combined with DDN solutions for archiving and backup – the GridScaler solution serves as the life sciences storage foundation for laboratory equipment ingest as well as HPC data assembly and analysis. Within a single, unified storage environment, researchers can now scale all laboratory services and eliminate application I/O bottlenecks.

DDN, leveraging a rich history in solving HPC file storage challenges on many of the world's largest supercomputers, has applied innovative technology – proven with over 200 Petabytes deployed world-wide – to solve life science challenges across the world and deploy scalable infrastructure for today's and tomorrow's "omics" data sets.

© 2009, DataDirect Networks, Inc. All Rights Reserved. DataDirect Networks, the DataDirect Networks logo, Silicon Storage Architecture and S2A are trademarks of DataDirect Networks. All other trademarks are the property of their respective owners.

**DataDirect**<sup>™</sup>  
N E T W O R K S

[ddn.com](http://ddn.com)

1.800.TERABYTE

9351 Deering Ave.  
Chatsworth, CA 91311  
USA