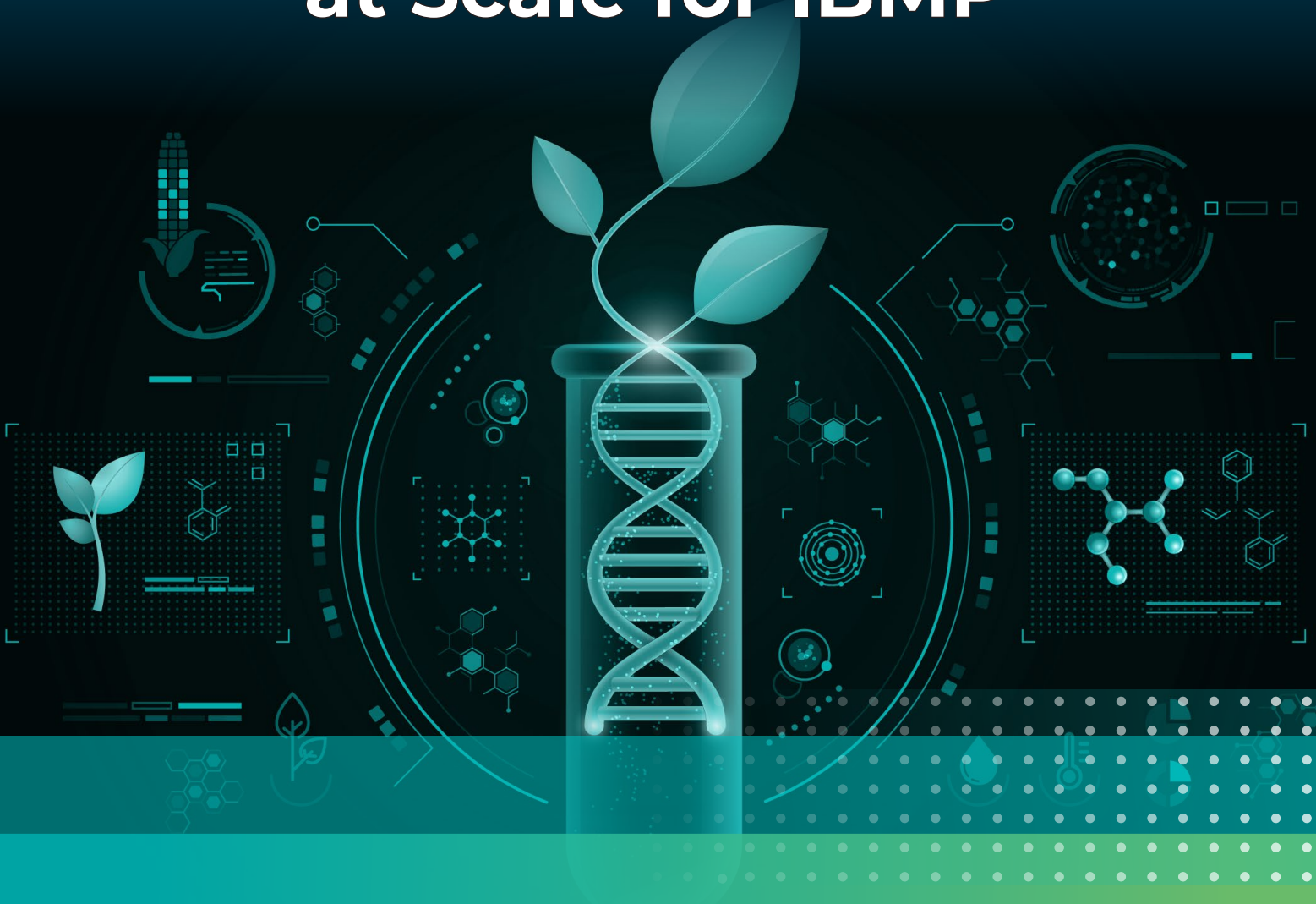DATACORE™

Empowering Bioinformatics Research and Data Handling at Scale for IBMP

The Institute of Plant Molecular Biology (IBMP), the largest CNRS laboratory in Alsace, France. IBMP is associated with the University of Strasbourg, engaging its 160+ researchers, doctoral candidates, and students from various nationalities in studying plant development, molecular structures, and viral diseases.

The Challenge

Today, the production of scientific data in digital form is widespread, and the implementation of new tools like Next-Generation Sequencing (NGS) leads to an explosive growth in data volume. At IBMP, around 80 TB of data is generated annually, and new methods such as nanopore sequencing, used to determine the arrangement of nucleotides in DNA fragments, further contribute to data inflation. Then, there are requirements to preserve this information for the long term, typically up to fifteen years, to enable comparison with more recent studies. Therefore, the data must always remain accessible.

Furthermore, considering the number and diverse origins of researchers at IBMP and their file identification logic, it is crucial to rely on a truly universal methodology of data access that allows rapid retrieval from the database. The IT department and scientific community at IBMP took all these factors into account when considering the **replacement of their RAID 6 NAS, which no longer met the heavy-duty demands of advanced sequencing methods.**

> **"**With DataCore Swarm, our institute takes a big leap forward in our ability to sequence plant DNA using the most advanced methodologies. Swarm provides us with a significant volume of bioinformatics data collected over several decades, which greatly enhances our analytical capabilities and improves our scientific performance.**"**

Jean-Luc Evrard
Director of the Information System, IBMP

Solution

IBMP underwent a comprehensive overhaul of its information system, embracing a range of IT transformations. These included the adoption of server and storage virtualization, as well as the implementation of a highly resilient architecture that is available 24/7. This solution relied on a VMware cluster backed by a software-defined storage (SDS) platform, DataCore SANsymphony, with a capacity of 200 TB.

While this system proved extremely robust, their long-term storage with NAS approach became increasingly outdated over time. Operational maintenance seemed more complex with growing capacity, and disk reconstruction times (in case of failure) were unreasonably long.

It was therefore imperative to find a solution that could handle increasing capacity with agility and effortlessly manage the oncoming data tsunami. After considering various options, traditional solutions were definitively ruled out, and it was determined that **only object storage enabled with S3 access** could meet the requirements and budget constraints of the institute.

Following a thorough evaluation of proposals from various vendors, two solutions, including DataCore Swarm, were being considered. Given their excellent support relationship with DataCore, **Swarm software-defined object storage** emerged as the preferred choice for IBMP.

Results

- Object-based storage architecture that outperforms traditional file systems

- Excellent resilience to failures, similar to SANsymphony (for block storage)

- A simple and accessible web interface for administration and content access (S3/HTTP)

- Robust storage system with effective data protection utilizing erasure coding

- Significant reduction in power consumption and energy costs through Darkive technology

Long-Term Data Storage
with Always-on Access

Swarm

Currently, Swarm is mainly used by a part of the bioinformatics team at IBMP, which generates and manages the largest volumes of data through Next-Generation Sequencing (NGS). While the hardware is fully operational, the software still requires some fine-tuning to facilitate the migration of data into Swarm. Metadata integration during data ingestion is a critical next step for IBMP for optimizing object retrieval from their extensive database. This will allow IBMP to move away from conventional and heterogeneous naming schemes (adopted by researchers handling data) that negatively affect search performance.

This initiative will take time as the CNRS, the institution's supervisory body, aims to deploy

an Electronic Laboratory Notebook (ELN) with a "digital record" accompanying every scientific data ingestion sequence. Since several laboratories share interest in object storage, it is necessary to take the time to express requirements, coordinate discussions, and share experiences within the ELN working groups. In the meantime, the bioinformatics data stored on Swarm is already accessible to users through dedicated visualization servers (such as JBrowse for genome identification), and the complete migration to object storage will be facilitated through the ELN. Primary data ingestion and storage of hot data will continue to be supported by SANsymphony on block storage, which reliably provides all services to IBMP users.

Deployment **Highlights**

- Swarm object storage cluster formed with 10 Dell PowerEdge servers
- Licensed initially for 850 TB of usable capacity (out of 1.3 PB of total raw capacity)
- VMware ESXi for server virtualization

- Active Directory integration for identity management and access control
- 25 Gbps link and 10 Gbps fiber optic link
- FS switches
- iDRAC connections to monitor remote machines

*This is a concise summary of the deployment highlights which is described in detail in the French version of this case study.

IBMP
Institut de biologie
moléculaire des plantes

Established in October 1987, the **Institute of Plant Molecular Biology (IBMP)** is a CNRS unit associated with the University of Strasbourg and affiliated with its Doctoral School of Life and Health Sciences (ED 414). As the leading French center for plant sciences at CNRS, the institute comprises four scientific departments dedicated to studying biosynthesis and regulation of bioactive molecules, plant viruses, regulatory pathways for development and plant reproduction, and the biogenesis of organelles such as chloroplasts and mitochondria, essential for cellular energy production. IBMP employs cutting-edge technological platforms for small molecule analysis, protein production, DNA sequencing, bioinformatics analysis tools, and climate-controlled greenhouses or growth chambers.

1023

Your Data, Our Priority

Discover the Ultimate Flexibility of DataCore Software

DataCore Software delivers the industry's most flexible, intelligent, and powerful software-defined storage solutions for the core, edge, and cloud. With a comprehensive product suite, intellectual property portfolio, and unrivaled experience in storage virtualization and advanced data services, DataCore has helped over 10,000 customers worldwide modernize how they store, protect, and access data. **www.datacore.com**