

CRAG (Centre for Research in Agricultural Genomics) is a public centre dedicated to cutting-edge scientific research in plant and farm animal biology. To meet the demand for increasing the speed and scalability of its high-performance computing infrastructure, which performs demanding tasks such as sequencing the genome of animals and plants, CRAG needed to replace its disk-based storage environment. In 2019, they installed a Pure Storage® FlashBlade™ solution that helped them to multiply writing speeds by 4x, to reduce full backup times by 60%, and to increase stability, eliminating unforeseen stops due to storage backend overloads.



BUSINESS TRANSFORMATION

CRAG was able to update its HPC environment to achieve the speed, scalability and stability required by the research projects of its users by leveraging the Pure Storage FlashBlade solution.

REGION / COUNTRY

EMEA / Spain

INDUSTRY

Agricultural Genomics

“Other solutions have been adapted to the use of flash technology, but Pure Storage has been built by and for flash storage, and that is the big difference.”

Gonzalo Vera,
Head of Scientific IT at CRAG

STATE-OF-THE-ART AGRICULTURAL GENOMICS RESEARCH

The Centre for Research in Agricultural Genomics (CRAG) is a public consortium formed by the Higher Council for Scientific Research (CSIC), the Agrifood Research and Technology Institute (IRTA), the Autonomous University of Barcelona (UAB) and the University of Barcelona (UB). Since 2011, it occupies modern facilities on the campus of the Autonomous University of Barcelona, where more than 200 researchers and support staff carry out an intense activity of basic and applied research in various programmes. These programmes range from the metabolism and stress response of plants to the genome of domestic animals, and are backed by a state-of-the-art technology platform.

The frontier scientific research carried out in CRAG uses large volumes of data and requires an advanced technical infrastructure. For this they have their own data centre, with a high performance computing (HPC) cluster and several types of interconnected nodes that access a shared storage platform. Every day, a large number of concurrent users demand large storage resources with high speeds, flexibility and scalability. Their needs may vary radically with each new experiment or analysis that is performed.

The previous storage platform from 2012 was totally based on hard drives and no longer satisfied the needs of the centre. As Gonzalo Vera, Head of Scientific IT at CRAG, explained: “It was clear that we had to go from a 100% disk-based solution to a hybrid solution that combined the speed of flash storage, for hot analysis data, with a disk-based solution for the storage of cold or archived data. Since we are a public institution we must keep data for all projects for a period of ten years.”

A SOLUTION THAT FITS PERFECTLY

To know the state of the industry of flash storage technology, CRAG IT managers made a market study in which the [Gartner Magic Quadrant for Solid-State Arrays](#) stood out. Although they evaluated other vendors, Pure Storage’s position as “visionary leader” in Gartner’s Quadrant caught their eyes. So they contacted the managers of the company in Spain to request more information. “We saw that, from the point of view of architecture, a Pure Storage solution was identical to what we were trying to assemble and that is why we got in touch with them. We especially liked that it is totally designed with flash; there is no trace of old disk technologies. Other solutions have been adapted to the use of flash technology, but Pure Storage has been built by and for flash storage and that’s the big difference,” said Vera.

COMPANY:

CRAG
www.cragenomica.es

USE CASE:

- High performance computing (HPC) – genomic sequencing

CHALLENGES:

- Modernize the disk storage platform.
- Maintain the current topology (NFS and KVM).
- Reduce the backup window.
- Ensure highest reliability.

IT TRANSFORMATION:

- Writing speed is now 4 times higher (up to 8 GiB/s vs previous 2.4 GiB/s).
- The backup window has been reduced by 60%.
- New features are available, such as snapshots.

“The saturation limit of the platform is so high that we have not even glimpsed it. We are certain that our storage will not fail.”

Gonzalo Vera,
 Head of Scientific IT at CRAG

Finally, CRAG opted for a [Pure FlashBlade](#) data hub where the entire data repository of the scientific cluster of the centre is stored. Eight 17 TB blades are connected by 40GbE links to a 100GbE network. The platform maintains the existing NFS file system and the mixed workloads generated in the computing nodes of the scientific cluster and the virtual nodes supported by the KVM hypervisor.

By internal policy, scientific IT managers of the centre want to maintain full control over all the hardware and software of the infrastructure so that they depend as little as possible on suppliers. For that reason, CRAG technicians participated directly in the installation of the FlashBlade, which was quick and did not pose any special difficulties. “We set it up in the morning and in the afternoon we connected it to our environment. We took a couple of days to organize the folders, we made a 12-hour stop to synchronize the files and since then the solution is working without problems,” summarized Vera.

SPEED INCREASE AND ELIMINATION OF BOTTLENECKS

The installation of Pure FlashBlade has satisfied the expectations of the CRAG IT department by providing more speed, less latency, new features and, above all, reliability.

Writing speed of the previous solution had a peak of 2.4 GiB/s, while with the Pure Storage platform, speeds of 4 GbB/s have already been reached and the configuration is sized to reach peaks up to 8 GiB/s. This performance has made it possible to eliminate the bottlenecks that appeared when the cluster was under heavy load. It has also made it possible to streamline other tasks, such as backup copies or occupation sampling taking place every two hours to invoice IT services to internal users.

“The latency of the cabin is extremely low even though we work with folders that can have more than 200 million files. Before it took days to know what was inside a directory; now we know it in a matter of minutes. Without touching anything else, we have reduced the time to make a complete backup of our entire environment from the 250 hours that we spent before to just 100 hours, and with a few changes we expect to finish it in less than 24 hours,” explained Vera.

Another advantage is that, with the new storage platform from Pure Storage, CRAG technical managers now have new features that they did not have before. For example, snapshots, instant copies that allow a complete recovery of the data at any time. In addition, Pure Storage specialists have continued to visit CRAG’s team regularly after the implementation to further optimize the solution.

“The most important thing for us is the peace of mind that the solution provides us, because we know that, even if we increase the load of the platform, we will not saturate it,” concluded Vera. “The saturation limit is so high that we have not glimpsed it. We are certain that our storage will not fail.”



info@purestorage.com
www.purestorage.com/customers