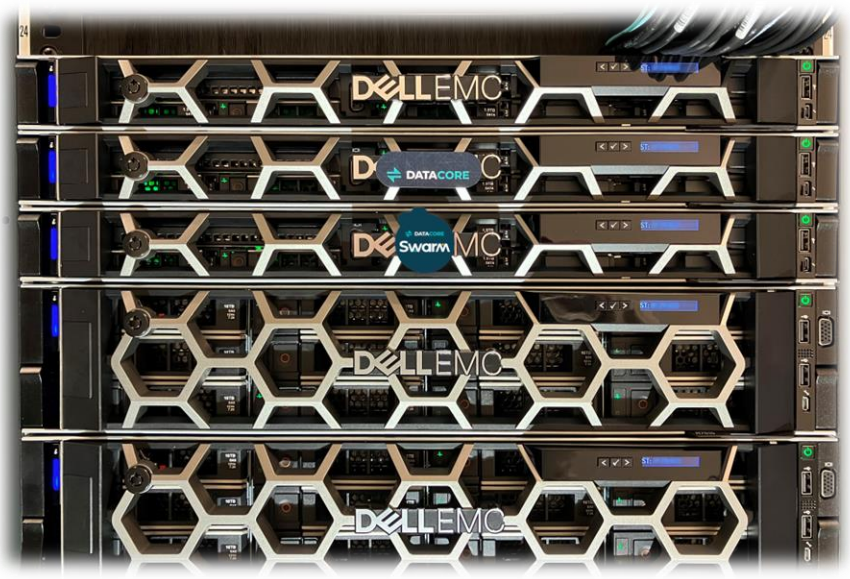

IBMP
Institut de biologie moléculaire des plantes

DataCore Swarm engrange les données bio-informatiques moissonnées par l'IBMP

L'Institut de biologie moléculaire des plantes (IBMP), plus gros laboratoire CNRS d'Alsace, associé à l'Université de Strasbourg, mobilise ses 160 chercheurs, doctorants et étudiants de toutes nationalités dans l'étude du développement des végétaux, de leurs structures moléculaires et de leurs maladies virales. Un travail qui génère une masse considérable de données conservées en mode objet et consultables dans un système DataCore Swarm.

LE DEFI

Aujourd'hui, la production de données scientifiques sous forme numérique est généralisée et la mise en œuvre de nouveaux outils comme le séquençage de nouvelle génération (NGS) induit une croissance explosive de leur volumétrie. À l'IBMP ce sont déjà quelques 80To de données par an qui sont générées et les nouvelles méthodes, comme celle dite des nanopores, utilisée pour déterminer la disposition des nucléotides dans des fragments d'ADN, sont des facteurs d'inflation supplémentaire de leur production. À cette croissance volumétrique, s'ajoutent des contraintes temporelles puisqu'il est indispensable de conserver ces informations sur le long terme, jusqu'à une quinzaine d'années en moyenne, pour pouvoir les consulter et les comparer avec des études plus récentes ce qui signifie qu'elles doivent demeurer disponibles à tout instant. Enfin, compte tenu du nombre et de l'origine des chercheurs passant par l'IBMP, représentant jusqu'à 50 nationalités et des logiques différentes d'identification de fichiers, il est indispensable de s'appuyer sur une méthodologie vraiment universelle permettant des « fouilles » approfondies et rapides dans la base de données. Tous ces paramètres ont alors été pris en compte par la DSI et la communauté scientifique de l'IBMP lorsqu'il a été envisagé, en 2021, de remplacer le NAS en Raid 6 qui servait jusque-là pour la conservation des données sur le long terme mais ne répondait plus aux contraintes nouvelles générées par les méthodes avancées de séquençage.

LA SOLUTION

Le système d'information d'IBMP a été entièrement rénové en 2015 sur la base de quelques principes telle la virtualisation des serveurs, comme du stockage, avec la mise en œuvre d'une architecture redondante et disponible en 24/7. Cette solution repose sur un cluster sous VMWare adossé à un système de software-defined storage (SDS) de 200To redondés en temps réel DataCore SANsymphony. Ce système s'est avéré extrêmement robuste, mais le principe du NAS de stockage long terme s'est révélé de plus en plus dépassé au gré du temps : son maintien opérationnel s'est complexifié avec les augmentations de capacités tandis que les délais de reconstruction en cas de panne de disque devenaient déraisonnables. Il était donc impératif de trouver une solution à la fois capacitive, agile, et permettant d'anticiper le tsunami de données qui s'annonçait. Plusieurs consultations et analyses prospectives ont permis d'écarter définitivement les solutions traditionnelles et de déterminer que seul le stockage objet dit S3 (Simple Storage Service) était apte à répondre aux critères du cahier des charges et aux contraintes budgétaires de l'Institut. Un tour des propositions des constructeurs a finalement mis en concurrence deux solutions dont Swarm, solution tout juste arrivée dans le giron de DataCore, une entreprise avec qui l'IBMP entretenait une relation de support.

LES RÉSULTATS

Un système de stockage robuste dont la méthode de protection de données par dissémination de fragments (erasure coding) est particulièrement efficace et dont le mode objet surpasse définitivement le classique système de gestion de fichiers (file system).

Une excellente résilience vis-à-vis des pannes à l'instar du comportement de SANsymphony. Une interface Web simple et abordable, plutôt orientée administrateur.

Une réduction significative de la consommation électrique et donc de la facture énergétique grâce à la technologie Darkive

Un stockage de données longue durée toujours accessibles à tout moment

Pour confirmer le choix de Swarm, les niveaux des performances du système ont été vérifiés par une phase d'essai à distance, avec simulation de pannes, sur un serveur basé chez DataCore France à Paris. Des tests ont également permis de valider l'intégration logicielle avec l'Active Directory et le déploiement des droits d'accès. La solution Swarm de DataCore a ensuite été installée sur site début 2022, par l'équipe du SI sur un ensemble de dix serveurs Dell, trois R6515 en tête du cluster pour supporter les services et sept R7515 pour le stockage proprement dit, tous sous contrat de maintenance de sept ans, la virtualisation étant assurée par des ESXi VMware. Le déploiement logiciel a, quant à lui, été effectué directement par DataCore. Ces matériels sont interconnectés par des liens redondants à 25 Gbps transitant par un switch FS S5860-48SC, lui-même en liaison avec le cœur de réseau par une fibre optique 10 Gbps. Un second petit switch FS S3700-24T4F sert aux liaisons iDRAC pour la surveillance des machines à distance. L'architecture retenue devrait d'ailleurs favoriser la migration future du dispositif vers le datacenter du campus. La pérennité logicielle de la solution est assurée par une licence « à vie » pour 850To de stockage, sur le 1,3Po brut disponible, et un contrat de maintenance de 3 ans. L'investissement représente une enveloppe 145 k€ HT.

Swarm est pour l'heure principalement utilisé par une partie de l'équipe de bio-informatique, celle qui génère et gère les plus gros volumes de données par séquençage NGS. Le matériel est donc complètement opérationnel tandis que la partie logicielle nécessite encore des mises au point pour que l'intégralité des informations produites à l'IBMP migrent dans Swarm. Pour cela, il faut finaliser la méthode d'intégration des métadonnées dès l'ingestion des données dans le système, processus indispensable pour optimiser la « fouille » (data mining) dans cette imposante base et ne plus dépendre d'un classique processus de nommage, nécessairement hétérogène, vu la diversité d'origine des chercheurs et donc pénalisant en termes de performance de « fouille ». Ce travail prend du temps car le CNRS, tutelle de l'établissement, souhaite déployer un Cahier de Laboratoire Electronique (CLE), avec une « fiche numérique » qui doit accompagner chaque séquence d'ingestion de données scientifiques. Plusieurs laboratoires ayant les mêmes préoccupations et un même intérêt pour le stockage objet, il faut prendre le temps d'exprimer les besoins, de coordonner les réflexions et de partager les expériences au sein des groupes de travail sur ce CLE. En attendant, des données de bio-informatique stockées sur Swarm sont déjà accessibles via des serveurs de visualisation dédiés (Jbrowse pour l'identification de génomes), l'intégralité devant être poussée sur le stockage objet par l'intermédiaire du CLE. En amont, l'ingestion primaire et le stockage des données chaudes se fait toujours sur SANsymphony qui fournit sans défaillir l'ensemble des services aux utilisateurs de l'IBMP.

Avec DataCore Swarm, notre institut fait un bond en avant dans sa capacité à séquencer l'ADN des végétaux selon les méthodologies les plus en pointe. Cet outil met à notre disposition un volume très conséquent de données bio-informatiques récoltées sur plusieurs dizaines d'années, ce qui démultiplie nos capacités d'analyse et améliore finalement notre performance scientifique

Jean-Luc Evrard, directeur du système d'information de l'IBMP

Conclusion

DataCore Swarm consolide la capacité d'acquisition de l'IBMP qui s'accorde parfaitement à l'air du temps, celui de la science ouverte, des référentiels nationaux et internationaux de stockage centralisé et organisé de données et d'un mode de fonctionnement et de partage « full web ».

***

A propos de l'IBMP : Créé en octobre 1987, l'Institut de biologie moléculaire des plantes (IBMP) est une unité propre du CNRS associée à l'Université de Strasbourg et affiliée à son Ecole doctorale des sciences de la vie et de la santé (ED 414). Premier centre français du CNRS en sciences du végétal, l'institut compte 4 départements scientifiques qui se consacrent à l'étude de la biosynthèse et de la régulation de molécules bioactives, à l'étude de virus végétaux, à l'exploration des voies de régulation pour le développement, la reproduction des plantes et leur adaptation à leur environnement, enfin à l'étude de la biogenèse des organites, chloroplastes et mitochondries, indispensables à la production d'énergie des cellules. Pour accomplir ces travaux, l'IBMP met en œuvre des plates-formes technologiques de pointe (analyse de petites molécules, production de protéines, séquençage d'ADN, d'outils d'analyse bio-informatique ainsi que des serres ou logettes climatisées.

https://www.ibmp.cnrs.fr/

A propos de DataCore : DataCore Software est un développeur de solutions de classe mondiale destinées à relever les plus grands défis du stockage. Depuis plus de 15 ans, les entreprises de médias et de divertissement les plus exigeantes font confiance à la solution Swarm de DataCore pour préserver et protéger leurs pétaoctets de contenus haute résolution. Rapide, sécurisé, abordable et évolutif, DataCore Swarm accélère la monétisation de contenus pour les professionnels du multimédia.

www.datacore.com

Contacts presse
the messengers
Elodie Antoine / Jennifer Lepreux
datacore@themessengers.fr
06 34 42 63 81 / 06 34 42 62 86

DATACORE