DATACORE

KUNDENBERICHT

IT-Grundlagen für die Bioinformatikforschung und die Verarbeitung großer Datenmengen im IBMP

Das Institute of Plant Molecular Biology (IBMP) ist das größte CNRS-Labor im französischen Elsass. An dem der Universität Straßburg angegliederten Institut erforschen über 160 Wissenschaftler, Doktoranden und Studierende unterschiedlicher Nationalitäten Pflanzenentwicklung, Molekularstrukturen und Viruserkrankungen.

Die Herausforderung

Heutzutage werden die meisten wissenschaftlichen Daten in digitaler Form produziert. Darüber hinaus führt die Implementierung neuer Tools wie Next-Generation Sequencing (NGS) zu einem explosionsartigen Wachstum des Datenvolumens. Am IBMP werden jährlich rund 80 TB an Daten generiert. Zudem tragen neue Verfahren wie die Nanopore-Sequenzierung, mit der die Anordnung von Nukleotiden in DNA-Fragmenten bestimmt werden kann, weiter zur inflationären Zunahme der Datenmenge bei. Hinzukommt die Notwendigkeit, diese Daten langfristig aufzubewahren, in der Regel bis zu 15 Jahre lang, um den Vergleich mit aktuelleren Studien zu ermöglichen. Aus diesem Grund müssen die Daten immer abrufbar bleiben.

Bei Betrachtung der Zahl und der unterschiedlichen Herkunft der Forscher am IBMP sowie deren unterschiedlichen Methoden zur Dateiidentifikation wird klar, wie wichtig ein universelles Verfahren für den Datenzugriff ist, das den schnellen Abruf von Daten aus der Datenbank ermöglicht. Die IT-Abteilung und die wissenschaftliche Community am IBMP zogen all diese Faktoren in Betracht, als der **Austausch ihres RAID 6 NAS anstand, der den hohen Anforderungen moderner Sequenzierungsmethoden nicht länger gewachsen war.**

„DataCore Swarm ist für unser Institut die ideale Lösung zur Sequenzierung von Pflanzen-DNA mit modernsten Methoden. Swarm ermöglicht uns die Verarbeitung großer Mengen an Bioinformatikdaten, die über mehrere Jahrzehnte hinweg gesammelt wurden. Damit verbessern sich sowohl unsere Analysefähigkeiten als auch unsere wissenschaftlichen Leistungen.“

Jean-Luc Evrard
Director of the Information System, IBMP

Lösung

Die IT des IBMP wurde einer umfassenden Modernisierung unterzogen, die mit einer Reihe von IT-Transformationen verbunden war. Dazu zählten auch die Einführung der Server- und Speichervirtualisierung sowie die Implementierung einer hochgradig ausfallsicheren Architektur mit 24/7-Verfügbarkeit. Diese Lösung hing an einem VMware-Cluster mit einer Kapazität von 200 TB, unterstützt von der Software-Defined Storage (SDS)-Plattform SANsymphony.

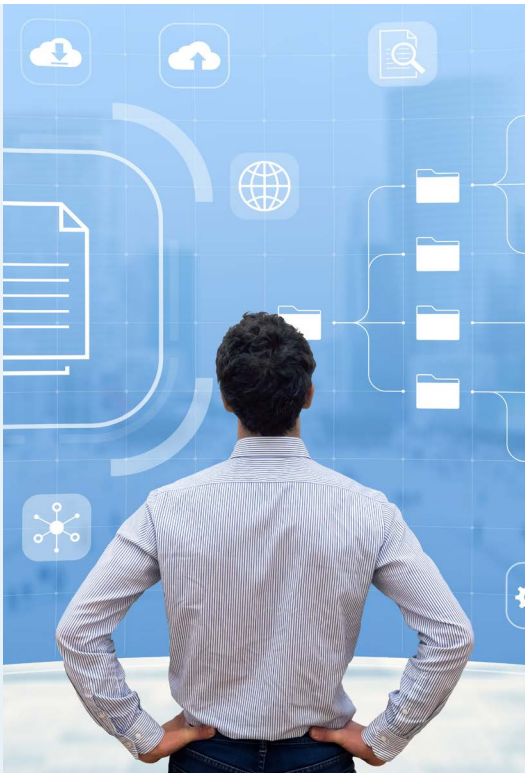
Dieses System hatte sich als äußerst robust erwiesen; allerdings stieß der NAS-Langzeitspeicher im Laufe der Zeit an seine Grenzen. Die betriebliche Wartung wurde mit zunehmender Kapazität immer komplexer und die Festplattenwiederherstellung (nach Ausfällen) nahm unangemessen viel Zeit in Anspruch.

Es wurde also dringend eine Lösung benötigt, die die wachsenden Kapazitätsanforderungen flexibel erfüllen und die steigende Datenflut problemlos bewältigen konnte. Nach Prüfung mehrerer Optionen wurden herkömmliche Lösungen definitiv ausgeschlossen. Es wurde beschlossen, dass **nur eine Objektspeicherlösung mit S3-Zugang** die Anforderungen erfüllen würde, ohne das begrenzte Budget des Instituts zu sprengen.

Nach gründlicher Auswertung der Angebote mehrerer Anbieter wurden zwei Lösungen in die engere Wahl gezogen, eine davon DataCore Swarm. Angesichts der ausgezeichneten Erfahrungen mit dem Support von DataCore, entschied sich das IBMP schließlich für den **Software-Defined Objektspeicher Swarm**.

Ergebnisse

- Eine objektbasierte Speicherarchitektur mit höherer Leistungsfähigkeit als traditionelle Dateisysteme
- Ausgezeichnete Widerstandsfähigkeit gegenüber Ausfällen, ähnlich wie bei SANsymphony (für Blockspeicher)
- Eine einfache, benutzerfreundliche Web-Schnittstelle für die Administration und den Zugriff auf Inhalte (S3/HTTP)
- Ein robustes Speichersystem mit dem wirksamen Schutz der Daten dank Erasure Coding
- Erheblich geringerer Stromverbrauch und somit geringere Stromkosten dank Darkive-Technologie



Langfristige Datenspeicherung mit uneingeschränktem Datenzugriff

Derzeit wird Swarm überwiegend von einem Teil des Bioinformatik-Teams beim IBMP genutzt, dass die größten Datenmengen durch Next-Generation Sequencing (NGS) generiert und verarbeitet. Während die Hardware vollständig betriebsbereit ist, muss die Software noch optimiert werden, um die Migration der Daten zu Swarm zu ermöglichen. Die Integration der Metadaten bei der Datenaufnahme ist ein kritischer nächster Schritt für das IBMP, um den Objektabruf aus der umfangreichen Datenbank zu optimieren. So kann das IBMP sich von den konventionellen heterogenen Namensschemata lösen, die von den verschiedenen Forschern bei der Datenverarbeitung eingeführt wurden und die Suchleistung beeinträchtigten.

Diese Initiative wird eine geraume Zeit in Anspruch nehmen, da das CNRS, das Aufsichtsgremium des Instituts, ein Electronic Laboratory Notebook (ELN) mit einem „digitalen Datensatz“ einführen möchte, der jede Aufnahmesequenz wissenschaftlicher Daten begleiten soll. Da mehrere Laboratorien am Objektspeicher beteiligt sind, muss ausreichend Zeit dafür eingeplant werden, die Anforderungen zu formalisieren, Gespräche zu koordinieren und Erfahrungen innerhalb der ELN-Arbeitsgruppen auszutauschen. In der Zwischenzeit sind die in Swarm gespeicherten Bioinformatik-Daten durch spezielle Visualisierungsserver (wie JBrowse zur Genomidentifikation) bereits für die Nutzer abrufbar, und die komplette Migration in den Objektspeicher läuft über das ELN. Die primäre Datenaufnahme und die Speicherung „heißer“ Daten wird weiterhin von SANsymphony auf Blockspeicher unterstützt, damit alle Dienste den IBMP-Nutzern zuverlässig zur Verfügung stehen.

Wichtigste Punkte der Bereitstellung

- Swarm Objektspeicher-Cluster, bestehend aus 10 Dell PowerEdge-Servern
- Vorläufig lizenziert für 850 TB an nutzbarer Kapazität (von 1,3 PB Rohkapazität insgesamt)
- VMware ESXi für die Servervirtualisierung
- Active-Directory-Integration für das Identitätsmanagement und die Zugriffskontrolle
- 25 Gbps Link und 10 Gbps Glasfaser-Link
- FS-Switches
- iDRAC-Verbindungen zur Überwachung von Remote Machines

*Dies ist eine kurze Zusammenfassung der wichtigsten Punkte der Bereitstellung. Nähere Einzelheiten sind der französischen Fassung dieses Kundenberichts zu entnehmen.

IBMP
Institut de biologie moléculaire des plantes

Das im Oktober 1987 gegründete **Institute of Plant Molecular Biology (IBMP)** ist eine Abteilung des CNRS, die der Fakultät für Life- and Health Sciences (ED 414) der Universität Straßburg angegliedert ist. Als führendes französisches Zentrum für Pflanzenstudien am CNRS umfasst das Institut vier wissenschaftliche Abteilungen, die sich der Erforschung der Biosynthese und der Regulierung bioaktiver Moleküle, Pflanzenviren, regulatorischer Wege für die Entwicklung und pflanzliche Reproduktion sowie der Biogenese von Organellen wie Chloroplasten und Mitochondrien widmen, die essenziell für die Zellenenergieproduktion sind. Das IBMP nutzt hochmoderne technologische Plattformen für die Analyse kleiner Moleküle, Proteinproduktion, DNA-Sequenzierung, Bioinformatik-Analysetools und klimageregelten Treibhäuser oder Wachstumskammern.