Organización en la que se ha implantado el proyecto: Centre de Regulació Genòmica
Apoyado por: FLYTECH S.A.

Solución de almacenamiento cluster paralelo para la investigación genómica y proteómetrica, orientado al servicio a 2 supercomputadores con 974 núcleos de proceso y 500 usuarios concurrentes.

Antecedentes/Problemática

El **Centre de Regulació Genòmica (CRG)** es un innovador centro de investigación creado en Diciembre de 2000 y situado en Barcelona, cuyo principal objetivo es ser un centro de referencia internacional en investigación biomédica

Focalizado en los ámbitos de la regulación genómica y proteómica, otros de sus objetivos es atraer y formar científicos destacados a todos los niveles

El **CRG** es un centro de investigación compuesto por 6 programas formados por un máximo de 6 laboratorios. Además de esto, varias de las principales instalaciones del CRG están ofreciendo algunos servicios a la comunidad científica mundial , además de a todo el instituto.

Objetivos

Necesitábamos un sistema de almacenamiento que pudiera soportar la carga de trabajo de nuestros usuarios, los cuales el 25% son estaciones de trabajo, además de poder dar servicio a nuestros dos supercomputadores para los procesos del genoma humano. Como es de todos conocido, los procesos informáticos de la secuenciación del **ADN** y **proteinas** requieren un altísimo poder de cálculo, grandísimas capacidades de almacenamiento y tasas de transferencia sin parangón.

También era imprescindible el soporte a múltiples protocolos, técnicas avanzadas de seguridad en los datos y funcionalidades avanzadas. Se realizaron cálculos para prever la tasa de transferencia y la capacidad máxima, siendo estas de **8GB/s** y **5 Petabytes** (5120 Terabytes)

Cualquier cuello de botella o imposibilidad de crecimiento, habrían llevado al traste el proyecto.

Fases del Proyecto – Recursos empleados

Nuestros usuarios de almacenamiento se pueden dividir en dos categorías, las estaciones de trabajo / equipos de sobremesa y los nodos de cálculo.

En el **Centre de Regulació Genòmica** disponemos de 400 equipos de sobremesa (150 MAC, 200 Windows, 50 Linux) y 100 estaciones de trabajo Linux.

Todos los equipos de sobremesa y todas las estaciones de trabajo están en funcionamiento las 24 horas del día, 7 días a la semana, por lo cual, cualquier impacto en el rendimiento no es asumible y por supuesto impensable, cualquier mínima parada.

La parte crítica a nivel de cliente eran dos supercomputadores de cálculo HPC. El primero está compuesto por 100 nodos (con un total de alrededor de 770 núcleos) dedicados a trabajos con grandes requisitos de memoria, y el segundo está compuesto por 35 nodos (con un total de 204 núcleos) dedicados a trabajos que requieran una gran cantidad de CPU.

Resultados- Beneficios (entidad y/o ciudadano)

Se planteó la necesidad de solventar los 2 puntos calientes: capacidad y rendimiento. La opción más viable era un nuevo sistema de almacenamiento paralelo.

Flytech propuso la solución Panasas ActiveStor 8, que encajaba perfectamente con los requerimientos nuestros, solución compuesta por múltiples cabinas de 4U, cada una de ellas con **40TB de capacidad y 40GB de caché por cabina**, formando todas ellas un clúster de almacenamiento de gran capacidad y altas prestaciones, destinada principalmente a guardar los datos de los clusters de cálculo, los cuales pueden acceder concurrentemente, con unas tasas de transferencia únicas en el mercado. Sirva como ejemplo, cada rack de almacenamiento dispone de 400GB de cache en pre-escritura y 96 microprocesadores trabajando en paralelo para el trasiego de los datos

Con esta solución conseguimos un rendimiento excepcional y una gran capacidad para guardar los datos en un único filesystem. Todas las cabinas trabajan como un solo sistema y al agregar más cabinas, crecemos en capacidad, en velocidad e I/O´s. Este trinomio es fundamental para nosotros.

Actualmente tenemos más de 400.000 GB en producción, más de 100.000.000 ficheros de “datos vivos” en un único espacio de nombres, accesible por todos los Clusters de cálculo y desde todas las estaciones Linux, con un resultado satisfactorio, pudiendo ampliar el ancho de banda de acceso a los ficheros hasta 100GB/s

Conclusiones de la Entidad

La solución nos permite ser más eficientes en nuestros procesos, lo que redunda en una maximización de las tareas de investigación, con el consiguiente avance en los campos genómicos y proteométricos

Gracias al estudio preliminar de los requerimientos y optar por una solución escalable cluster de almacenamiento paralelo, hemos conseguido solventar nuestras necesidades y dar el servicio requerido.

Realizado por Don David Carmargo, director TIC del **Centre de Regulació Genòmica**