

## SUPERANDO LOS 10GB/S Y 2PTB DE ALMACENAMIENTO PARA LOS EXPERIMENTOS DEL ACELERADOR DE PARTÍCULAS LHC (Large Hadron Collider)

### Antecedentes

El Large Hadron Collider, en el Laboratorio Europeo de Física de Partículas (CERN) es el mayor instrumento científico jamás construido. Empezó a funcionar en Noviembre de 2009 y produce aproximadamente 15 Petabytes de datos cada año. Para el procesamiento y análisis de estos datos es necesario generar adicionalmente datos secundarios y simulados. En total, los experimentos del LHC generan entre 50 y 100 Petabytes de datos anualmente. Teniendo en cuenta que la vida útil esperada para estos experimentos es de al menos 10 años, esto sitúa al LHC como el primer experimento científico en la escala del Exabyte. Esta enorme cantidad de datos será analizada por miles de científicos de todo el mundo, lo que supone un reto adicional para la infraestructura informática necesaria. El proyecto Worldwide LHC Computing Grid (WLCG) se inició en 2002 con el objetivo de construir y poner en marcha la infraestructura informática para procesar los datos del LHC. Dicha infraestructura consta de recursos provistos por más de 170 centros informáticos en 34 países, el acceso a los cuales se uniformiza mediante el uso de servicios Grid. En WLCG los centros informáticos se han clasificado jerárquicamente. La “capa cero” de la jerarquía es el CERN, origen de los datos. En la primera capa se encuentran 11 centros con unos requisitos de fiabilidad y eficiencia muy altos: los centros Tier1. La función principal de estos centros es proporcionar servicios de almacenamiento masivo a largo plazo, así como recursos de cálculo para el procesamiento y filtrado masivo de los mismos. España contribuye al LHC con uno de estos centros Tier1, que da soporte a los experimentos ATLAS, CMS y LHCb. El centro Tier1 español es el Port d'Informació Científica (PIC), situado en el campus de la UAB

cerca de Barcelona, y proporciona aproximadamente un 5% de la potencia total de los 11 centros Tier1.

El PIC es un centro científico-tecnológico mantenido mediante un acuerdo de colaboración entre el CIEMAT, la Generalitat de Catalunya, el IFAE y la UAB. Recibe financiación para desplegar el Tier1 a través de proyectos específicos del Plan Nacional de Física de Partículas. El proyecto WLCG está gobernado por un Memorandum de Entendimiento (MoU, por sus siglas en inglés) firmado por el CERN y las instituciones que albergan centros Tier1 y Tier2, representadas por las agencias nacionales de financiación. España, a través del Ministerio con competencias en Ciencia y Tecnología, firmó el MoU en Julio de 2007. El MoU define los niveles de servicio que los centros han de proporcionar así como su capacidad de cálculo y almacenamiento. Los Tier1 son centros informáticos de gran capacidad. De acuerdo con la planificación del LHC, estos 11 centros tienen que proporcionar entre el 30% y el 40% de toda la capacidad del sistema Grid del LHC. Debido a esto, y la alta criticidad de sus servicios, es fundamental que su crecimiento de capacidad se ajuste a los perfiles comprometidos en el MoU. La capacidad actual de almacenamiento en el Tier1 del PIC es de 3500 Terabytes de disco y 5200 Terabytes de cinta magnética.

### Necesidades

El PIC, para mantener el estatus de Tier1 y satisfacer las necesidades del LHC y sus tres experimentos principales (ATLAS, CMS y LHCb), necesita crecer en capacidad de almacenamiento además de elevar las tasas de transferencia, sin obviar la seguridad en los datos.

El objetivo del proyecto de ampliación de almacenamiento se basó en 5 premisas principales:

1. Coste por GB
2. Consumo eléctrico por GB
3. Tasa de transferencia típica
4. Máxima densidad GB/U
5. Capacidad total

## Descripción del proyecto

El proyecto arranca con un trabajo exhaustivo de ingeniería I+D de proyectos. Flytech prepara una serie de sistemas ABASTOR con múltiples configuraciones, para conseguir cumplir las 5 premisas anteriormente citadas.

Desde el primer momento, teníamos la certeza de que el mayor escopo sería conseguir las tasas de transferencia requeridas a un coste extremadamente bajo por GB. Para ello, se realizan múltiples pruebas con herramientas como IOZONE, para ajustar el alineamiento de los Stripe Size de las controladoras RAID y el Stripe Size del file system.

Las controladoras ADAPTEC RAID Hardware incorporaban microprocesadores multicore para el algoritmo X-OR y cache de 512MB por controladora.

Se generan diferentes layouts RAID6, para conseguir diferentes perfiles de prueba, al igual que se realizan diferentes configuraciones de NCQ.

Puesto que el objetivo final de uso era trabajar normalmente con un porcentaje muy alto de ficheros grandes, superiores a la capacidad de cache de las controladoras, se realizan pruebas con hasta 50 streams concurrentes de capacidades desde 2MB hasta 100GB cada uno de ellos, consiguiendo una configuración optima, la cual nos permitía superar los 10,4 GB/s de acceso a disco en lectura/escritura. El utilizar estos

tamaños de ficheros, es para que las caches de los controladores RAID siempre estén saturadas y no pudieran “falsear” los resultados de los test, dando valores más altos de los reales de acceso a disco.

Otro elemento clave era el consumo. Para conseguir un consumo/GB muy bajo, se utilizan chasis de alta densidad de discos, fuentes de alimentación de elevada eficiencia energética y procesadores de bajo consumo. Cada 8U de altura alberga 81 discos de 3TB SATA Enterprise, utilizando el proyecto un total de 64U. Esto habilita una capacidad RAW de 1944TB. Esta capacidad era muy superior a la mínima requerida por el PIC. Para optimizar todavía más la capacidad, los sistemas operativos de cada una de las 8 cabeceras NAS ABASTOR incluyen internamente 2 discos SATAII en raid 1 para el sistema operativo de cada cabecera. Esto se hace de este modo, porque de otro modo, perderíamos la capacidad de 16 discos de 3TB para el servicio (48TB).

La tasa de transferencia hacia los clientes (hosts), se configura en modo redundante, además de disponer de puertos de Backup. Dicho de otro modo, el proyecto incluye 16 puertos a 10Gb/s para el servicio de datos y otros 16 puertos a 10Gb/s para que en caso de caída, la infraestructura de Backup pueda mantener la tasa de transferencia sin degradación del servicio.

Para aumentar la eficiencia y eliminar el mantenimiento de todos los controladores RAID del proyecto (16 de doble canal SAS a 6Gbps), se diseñan sin batería de Backup cache (BBU). En vez de utilizar BBU's se opta por supercondensadores de mantenimiento cero ZMM (Zero Maintenance Module). Esto nos permite habilitar más rápidamente la protección de caches en escritura ante caídas de alimentación y eliminar en el futuro, el mantenimiento de las baterías basadas en Litio.

Para la administración de todo el sistema, cada cabecera NAS ABASTOR incorpora un puerto dedicado a 1Gbps para la consola de gestión, basada en el protocolo IPMI 2.0, permitiendo no sólo el acceso en modo consola, sino también el acceso en modo gráfico al entorno hardware (ventiladores, voltajes, fallos hardware, etc).



### Datos de la solución en producción

- ✓ **Coste por GB = 0,15€**
- ✓ Capacidad total RAW de la solución = 1944TB
- ✓ Tasa típica de la solución = 10,4 GB/s
- ✓ Numero de puertos 10 Gbps de la solución = 32 ópticos multimodo SFP+
- ✓ Numero de puertos 1 Gbps de la solución = 16 de cobre UTP
- ✓ Numero de puertos IPMI 1 Gbps para consola de la solución = 8 de cobre UTP
- ✓ Consumo de la solución por TB = 5,3W
- ✓ Consumo total de la solución = 10304W
- ✓ BTU's = 35167 BTU's
- ✓ Eficiencia energética de las fuentes de alimentación = Superior al 92%
- ✓ Número de cabeceras ABASTOR = 8
- ✓ Número de discos por cabecera ABASTOR= 36 discos para datos y 2 discos para sistema operativo
- ✓ Número de puertos SAS2.0 por cabecera = 4

- ✓ Número de controladores RAID por cabecera ABASTOR = 2
- ✓ Número de JBOD's ABASTOR = 8
- ✓ Número de discos por JBOD ABASTOR = 45 discos
- ✓ Número de puertos SAS2.0 por JBOD = 2
- ✓ Altura en U de cada cabecera = 4U
- ✓ Altura en U de cada JBOD = 4U
- ✓ Número de U usadas por la solución = 64U

### Principales beneficios

- ✓ Increíblemente bajo coste por GB
- ✓ Un ratio de consumo energético por TB muy bajo
- ✓ Altas tasas de transferencia para dar servicio a entornos de superalmacenamiento y supercomputación
- ✓ Ratio de capacidad/U muy elevado (superior a 10TB/U)
- ✓ Utilización de funcionalidades y sistema operativo estándar (LVM - Extendend File System - Linux)
- ✓ Optimización de procesos X-OR con controladores RAID multicore
- ✓ Redundancia muy alta, balanceo de cargas en los puertos de servicio de las cabinas
- ✓ Ahorro de costes al eliminar el sistema KVM (teclado/monitor/ratón), integrado en cada almacenamiento



## Acerca de FLYTECH

Desde 1988, FLYTECH es mayorista de soluciones informáticas, servidores, almacenamientos y seguridad. Como empresa de valor añadido, ofrece alta tecnología tanto a grandes cuentas y administración pública como al canal especializado de distribuidores, con asesoramiento, eficiencia y profesionalidad.

Con el objetivo de ofrecer el mejor servicio y aportar la solución necesaria a la extensa red de clientes, Flytech cuenta con oficinas en Madrid, Barcelona y Noroeste.

En FLYTECH sólo encontrará primeras marcas líderes de mercado como Supermicro, Overland, Adaptec, Nexsan, Panasas, Cherry, Netgear, Kingston, Sanrad, Belkin, etc. El esfuerzo y buen trabajo de Flytech durante más de 20 años ha sido reconocido por importantes fabricantes a nivel mundial.

## Acerca del PIC

El PIC, *Port d'Informació Científica*, fue fundado en 2003 gracias al acuerdo de 4 instituciones:

- ✓ Departament d'Universitats, Recerca i Societat de la Informació (DURSI) de la Generalitat de Catalunya (actualmente Departament d'Economia i Coneixement)
- ✓ Centro de Investigaciones Energéticas, Medioambientales y Tecnológica (CIEMAT)
- ✓ Universitat Autònoma de Barcelona (UAB)
- ✓ Institut de Física d'Altes Energies (IFAE)

Es un centro de excelencia para procesar datos científicos que da soporte a grupos de investigación que realizan proyectos que requieren un alto nivel de recursos de computación para el análisis de datos masivos.

Desde su creación en 2003, el PIC recibió el encargo de la Secretaría de Estado de Investigación para participar como Tier1 español en el proyecto del CERN para proceso de datos del Gran Colisionador de Hadrones (WLCG, Worldwide LHC Computing Grid)

## Personas de contacto

Eduardo Vales Hernandez. Director de Proyectos FLYTECH S.A.

Gerard Bernabeu. Production Coordinator. *Port d'Informació Científica*