Preservación de Datos en el marco del LHC

Memoria presentada por

Ibán Cabrillo Bartolomé

para optar al grado de Doctor por la Universidad de Cantabria

Dirigida por

Dra. Alicia Calderón Tazón

Dr. Jesús Marco de Lucas

Santander, Enero de 2020

Declaración de Autoría

Dra. Alicia Calderón Tazón, Doctora en Ciencias Físicas y Profesora de la Universidad de Cantabria,

y

Dr. Jesús Marco de Lucas, Doctor en Ciencias Físicas y Profesor de Investigación del Consejo Superior de Investigaciones Científicas,

CERTIFICAN que la presente memoria

Preservación de Datos en el marco del LHC

ha sido realizada por Ibán Cabrillo Bartolomé bajo nuestra dirección en el Instituto de Fisica de Cantabria, para optar al título de Doctor por la Universidad de Cantabria. Consideramos que esta memoria contiene aportaciones científicas suficientemente relevantes como para constituir la Tesis Doctoral del interesado.

En Santander, a 8 de Enero de 2020,

Dra. Alicia Calderón Tazón

Dr. Jesús Marco de Lucas

Índice general

Capítulo 1

Introducción

El objetivo de esta tesis doctoral es recoger las ideas y el trabajo desarrollado en un tema actual de gran interés, tanto técnica como científicamente: **la preservación de datos científicos**.

Este trabajo ha sido realizado en el Instituto de Física de Cantabria (IFCA), centro mixto de la Universidad de Cantabria (UC) y del Consejo Superior de Investigaciones Científicas (CSIC), donde varios grupos desarrollan su investigación en el contexto de colaboraciones internacionales, en áreas como física de partículas, astrofísica o medio ambiente y que conllevan la gestión de grandes volúmenes de datos.

Dentro del complejo ciclo de vida que requiere la gestión de los datos científicos[1], esta tesis se centra en la etapa de "preservación". Preservación entendida en un sentido más amplio que el del simple almacenamiento seguro de datos a largo plazo: en el curso del desarrollo de esta tesis el concepto de preservación se ha extendido para incluir, primero la preservación del software asociado a la gestión de dichos datos, y después la preservación de un entorno virtual que permita garantizar la reproducibilidad del sistema de gestión, con el objetivo último de lograr lo que entendemos como preservación del conocimiento (ver fig:1.1).

Esta tesis aborda además la implementación técnica, que ha sido realizada en el Centro de Procesado de Datos (CPD) del IFCA, de cuyo sistema de gestión de datos soy responsable desde el año 2009.

El desarrollo se ha extendido a lo largo de varias fases: test, adaptación, integración y despliegue de diferentes soluciones técnicas; como un sistema global

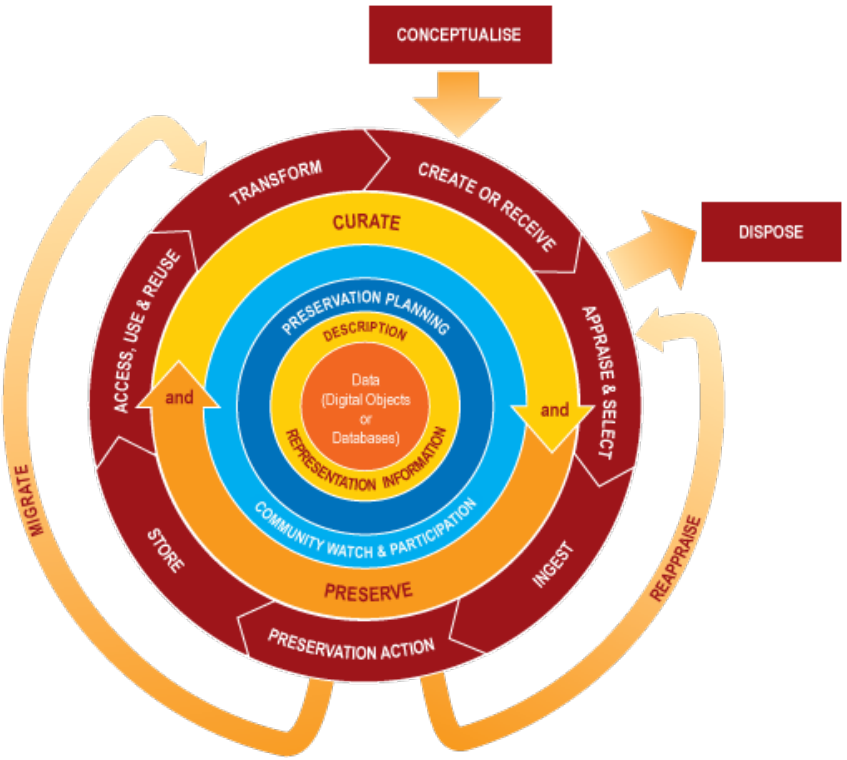


FIGURA 1.1: Ciclo de vida de los datos.
[2]

de almacenamiento de datos paralelizado sobre diferentes tipos de redes y servicios, que se extiende hasta el sistema de archivado en robots de cintas magnéticas, empleando soluciones de almacenamiento distribuido en "e-Infraestructuras" "Grid" y/o "Cloud".

El trabajo se ha desarrollado principalmente en el marco de la colaboración del experimento Compact Muon Solenoid (CMS) del Gran Colosionador de Hadrones (LHC) del Laboratorio Europeo de Física de Partículas (CERN). Como miembro de esta colaboración desde el año 2005, he tenido la satisfacción de contribuir a la implementación del sistema de computación distribuido que permitió procesar la gran cantidad de datos recogidos en este detector, participando en hitos como la identificación de la señal de producción del Bosón de Higgs. Hay que tener en cuenta que el análisis de datos en LHC ha requerido el mayor sistema de almacenamiento de datos científicos distribuido a escala mundial, con una capacidad cercana a los

FIGURA 1.2: Centro de Procesado de Datos del IFCA.

200PB. El IFCA participa en la colaboración CMS aportando un nodo, denominado Tier-2, con una capacidad que ha llegado a los 2PB de almacenamiento y 2.500 cores de procesamiento.

Como responsable del almacenamiento y preservación de datos en el CPD del IFCA (ver fig:1.2), durante la puesta en producción de este nodo, se han realizado múltiples actividades, entren ellas, el diseño de una infraestructura computacional singular, integrando soluciones de almacenamiento y preservación de datos con el supercomputador ALTAMIRA, nodo de la Red Española de Supercomputación (RES), Instalación Científico Técnica Singular (ICTS) de la Universidad de Cantabria, instalado y operado desde el IFCA.

Este desarrollo, ha posibilitado que los usuarios de la colaboración CMS, pudieran acceder a recursos de Computación de Alto Rendimiento (HPC) de forma puntual, incrementando su capacidad de computo.

Gracias a esta infraestructura, y a la colaboración en varios proyectos europeos e internacionales en tecnologías Grid y Cloud, se ha podido desarrollar e implementar algunas de las técnicas necesarias para procesar y gestionar datos, que se han aplicado en el resto proyectos de investigación en los que el IFCA participa.

Esta tesis se presenta estructurada en cuatro capítulos:

- En este primer capítulo se presenta el contexto de las grandes colaboraciones y consorcios internacionales, necesario para entender la complejidad de las infraestructuras computacionales en este tipo de entornos colaborativos a gran escala.

- El segundo capítulo se centra en el análisis en las comunidades sobre las que se han integrado las técnicas descritas en esta tesis. Se describirán los detectores del LHC, la implementación del sistema de flujo y procesado de datos, y especialmente, la solución desarrollada en los proyectos Worldwide LHC Computing Grid (WLCG) y Infraestructure Grid Europea (EGI) a la que he contribuido como responsable del nodo del IFCA.

- El tercer capítulo detalla el sistema de gestión de datos de CMS, y en particular en un nodo Tier-2, así como los diferentes procesos que garantizan la calidad de los datos. En este capítulo se mostrará también la integración desarrollada entre este centrp Tier-2 del IFCA y el nodo de supercomputación ALTAMIRA.

- Por último, el cuarto capítulo se centra en la preservación de datos la iniciativa Preservación de Datos para Física de Altas Energías (DPHEP). Se describen varios ejemplos de preservación y se muestra la política de preservación datos definida por la colaboración CMS. Se detallan diferentes implementaciones; desde el archivado de datos a nivel de bit, hasta la preservación del entorno completo, utilizando tecnologías computacionales actuales, desarrolladas desde el IFCA,y en las que he participado activamente.

Como hemos señalado anteriormente, mediante la "preservación de datos", nos estamos refiriendo a un concepto mucho más amplio que el simple hecho del almacenamiento seguro a largo plazo. Nos centraremos en la preservación del conocimiento adquirido, de manera que un agente externo sea capaz de recuperarlo: poder repetir todos los pasos dados en un análisis hasta llegar a la publicación de los resultados. La preservación nos abre nuevas posibilidades: reutilización de datos, nuevos análisis o incluso entrenamiento de algoritmos de aprendizaje automático.

En las diferentes disciplinas científicas podemos ver este proceso como una retroalimentación, a la hora de preparar y/o diseñar nuevos experimentos. Además posibilita la replicación de un resultado siguiendo unos pasos dados, certificando la reproducibilidad de los resultados obtenidos. Esta es una de las principales razones por las que una buena política de preservación de datos debe de estar presente en todo experimento científico. Es importante que esta planificación se lleve a cabo mientras el experimento se encuentra en su máximo apogeo, contando así con los recursos necesarios, tanto humanos como materiales que, dependiendo del tamaño del experimento, pueden ser más o menos cuantiosos.

1.1. Grandes colaboraciones Científicas

La imagen de científico de principios de siglo poco tiene que ver con la actual. Esta distorsionada imagen de científico solitario, que podía tener validez hasta hace una pocas décadas ha evolucionado rápidamente. Con el desarrollo de las nuevas tecnologías en el ámbito de las comunicaciones, los científicos de todo el mundo comenzaron a establecer colaboraciones, no solo dentro de sus propios centros de investigación, sino entre partes distantes del planeta, agrupándose por intereses comunes.

Con el paso del tiempo crece la necesidad de abordar desafíos cada vez mucho mayores, que den respuesta a las grandes preguntas que la ciencia se plantea en nuestro tiempo, como la formación del universo, el surgimiento de la vida en la tierra, la localización de mundos distantes o la composición de la materia. Este tipo de colaboraciones que hace unas décadas aglutinaba a unas pocas personas, de ámbitos reducidos, se ha transformado con el paso del tiempo en grandes colaboraciones formadas por cientos o incluso miles de físicos, ingenieros y otro tipo de personal de alta cualificación, distribuidos por diferentes regiones del mundo.

Esto ha sido posible, en parte gracias al desarrollo de las llamadas “nuevas tecnologías”, en particular el uso de Internet y más específicamente servicios como World Wide Web (WWW), el correo electrónico y algún tiempo después la videoconferencia. Estos avances tecnológicos han posibilitado una democratización en el acceso al conocimiento, ya que el intercambio de información puede fluir de

manera rápida de una a otra parte del globo, y prácticamente cualquier persona puede acceder a la misma.

Aunque el desarrollo de estas tecnologías introdujo un avance significativo en la formación de colaboraciones, todavía quedaba el desafío de poder construir grandes infraestructuras, muy complejas desde el punto de vista técnico y de un coste realmente elevado, imposibles de financiar para la mayoría de los países en solitario, pero necesarias para poder abordar los grades problemas científicos.

Estos retos técnicos y científicos situados en la frontera del conocimiento, fueron el impulso que las colaboraciones necesitaban para establecerse como entidades supranacionales, y ademas han permitido devolver lo aportado por la sociedad civil, en forma de nuevas técnicas y recursos tecnológicos.

Una de los principales retos de este tipo de agrupaciones es la gestión distribuida de los datos, en el caso de la colaboraciones de carácter científico, son los datos generados.

A continuación mostraremos algunas de las colaboraciones, más relevantes dentro su sus respectivos campos, para que una vez mostrado el contexto, podamos entender la problemática de la preservación en este tipo de entornos.

1.1.1. Física de Altas Energías (HEP)

Dentro de los diversos tipos de colaboraciones científicas actuales, una de las más productiva y proclive a la formación de grandes colaboraciones es Física de Altas Energías (HEP). Desentrañar los misterios de la composición de la materia y los orígenes de universo, son algunos de los propósitos de esta comunidad. Para poder responder estas preguntas, y simular las condiciones del origen del Universo primigenio, enormes experimentos han tenido que ser diseñados y construidos durante años. Estas máquinas son los "Aceleradores y Detectores de Partículas". Existen varios en funcionamiento, pero el más grande construido hasta la fecha es el Gran Colisionador de Hadrones (LHC) que se encuentra junto a la frontera entre Suiza y Francia. El LHC fue construido por una de las colaboraciones más importantes del mundo: CERN

El CERN fue inicialmente fundado por doce países europeos en 1954. Podríamos decir que hoy en día es un importante modelo de colaboración internacional y

uno de los laboratorios científicos más importantes del mundo. Actualmente el CERN se encuentra constituido por 23 estados miembros, los cuales aportan financiación y participan en la toma de decisiones de la organización. Además hay otros 28 estados participantes que no son miembros de pleno derecho. Ocho de estos estados y organizaciones tienen el estatus de observadores y pueden participar en los consejos. El presupuesto anual aproximado de CERN es de unos 750 millones de euros.

El objetivo principal del CERN es entender las interacciones entre las fuerzas elementales. Décadas de experimentos se han llevado acabo para poder conseguir este objetivo[3].

En 30 años, el CERN ha tenido dos infraestructuras principales, es decir dos aceleradores de partículas, ambos instalados sobre la misma estructura civil. El primero conocido como Large Electron-Positron collider (LEP) y el segundo, actualmente operativo, el LHC.

- **El LEP**: Era un acelerador de partículas (e^-) -(e^+) circular con una longitud de unos 27Km y construido a unos 100m bajo la ciudad de Ginebra. Estuvo en funcionamiento desde 1989 hasta el año 2000. LEP albergaba cuatro experimentos ALEPH, DELPHI, L3 y OPAL, en las cuatro zonas de colisión distribuidos a los largo de la circunferencia.

- **EL LHC**: Basado en la infraestructura inicial del LEP, y usando el túnel de 27km de longitud ya construido, el LHC fue diseñado para colisionar dos haces de protones (p^+). Cuenta con cuatro puntos de colisión, donde están instalados los experimentos principales para la detección de partículas: Los experimentos Compact Muon Solenoid (CMS) y A Toroidal LHC ApparatuS (ATLAS) son de propósito general, mientras que A Large Ion Collider Experiment (ALICE) trabaja con iones pesados y Large Hadron Collider beauty (LHCb) se diseñó para experimentos relacionados con el quark-b.

Uno de los principales propósitos por los que se construyó el LHC, fue producir y detectar la partícula conocida como Bosón de Higgs, que estaba formulada por el modelo estándar de la Física de partículas.

La búsqueda del Bosón de Higgs[4] en el CERN comenzó en finales de la década de 1980, en el LEP. Los experimentos en el colisionador Tevatron en Fermilab[5] en

los EEUU, también comenzaron a buscar el Bosón de Higgs en la década de 1990. La gran dificultad inicial era que la teoría no predecía la masa de la partícula y era posible que pudiera encontrarse en cualquier lugar dentro de un amplio rango. LEP se cerró en año 2000 para dar paso al LHC, y así los experimentos del LHC retomaron la búsqueda de dicha partícula en el año 2010, ya con la indicación más precisa de que dicha partícula se encontrase entre los 114Gev-130 GeV.

El CERN, el 4 de julio de 2012, anunció que los dos experimentos de carácter general CMS y ATLAS confirmaban la observación de una nueva partícula "consistente con el Bosón de Higgs", pero que se necesitaría más tiempo y datos para confirmarlo. Así e1 14 de marzo de 2013 el CERN, anunció la observación de una nueva partícula. La manera en que interactuaba con otras partículas, así como sus propiedades cuánticas, junto con la medida de sus interacciones, indicaban fuertemente que era un Bosón de Higgs (ver fig:1.3) del modelo estándar con una masa de $\sim 125 GeV$. Todavía permanece la cuestión de si es el Bosón de Higgs postulado por el modelo estándar u otro de los varios bosones predichos en algunas teorías que van más allá del modelo estándar.

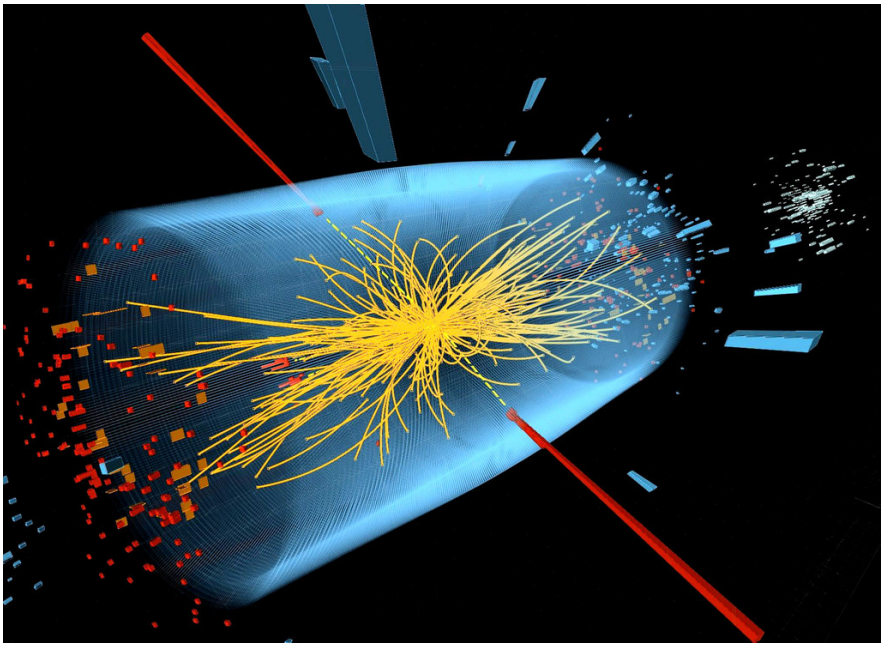


FIGURA 1.3: Evento de desintegración del Higgs a dos fotones observado en el experimento CMS.

El 8 de octubre de 2013 se concedió a Peter Higgs, junto a François Englert, el Premio Nobel de física "**Por el descubrimiento teórico de un mecanismo que contribuye a nuestro entendimiento del origen de la masa de las partículas subatómicas, y que recientemente fue confirmado gracias al descubrimiento de la predicha partícula fundamental por los experimentos ATLAS y CMS en el Colisionador de Hadrones del CERN**".

Los diferentes periodos de toma de datos del LHC, se denominan "**RUNs**", y abarcan diferentes espacios de tiempo, mientras que los tiempos de parada entre las tomas de datos son los "**LSs**" (Long Shutdown), y son empleados para diversas tareas como calibraciones, mejoras en los detectores, en el software de los experimentos y en la propia infraestructura del LHC.

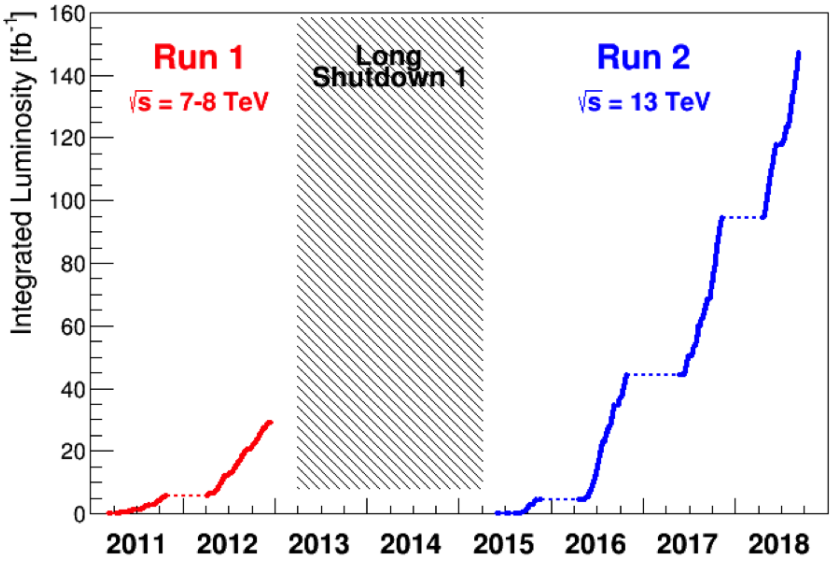


FIGURA 1.4: Comparativa de luminosidades entre el Run 1 y el Run 2.

- La primera toma de datos "**Run 1**" tuvo lugar entre los años 2011 y 2013 (ver fig:1.4), logrando una energía en el centro de masas de 7∼8Tev y una luminosidad máxima de $1{,}5 \cdot 10^{32} cm^{-2} s^{-1}$. Durante este "Run", tuvo lugar el anuncio de la observación de una partícula consistente con el bosón en la región de masas alrededor de ∼125GeV, con una significación estadística de 5σ.

- Durante el "**Run 2**", la segunda fase de toma de datos en el LHC, que tuvo lugar entre 2015 y 2018, se alcanzó una energía en el centro de masas de 13Tev y se observaron diez mil millones de colisiones. Siguen las investigaciones encaminadas a descubrir los pormenores del comportamiento del Bosón de Higgs. Las mediciones de ATLAS y CMS señalan hasta ahora que los ritmos de producción y desintegración observados son compatibles, dentro de la actual incertidumbre estadística, con los previstos por el modelo estándar.
- El "**Run 3**", la tercera fase de toma de datos, está prevista para 2021-2023.
- A partir del año 2026 entraremos en una fase de "alta luminosidad" (HL-LHC) "**Run 4**", que puede arrojar conocimiento sobre "nueva física" más allá del modelo estándar y la materia oscura.

Los recursos computacionales comprometidos por el CERN con los experimentos, ATLAS, CMS, ALICE y LHCb, para el año 2019, ascienden a 1.270.000 CPU (HEP-SPEC06), 97PB de almacenamiento en Disco y 272PB de almacenamiento en cinta magnética.

1.1.2. Astrofísica

Si hablamos de colaboraciones que se han desarrollado en el contexto de la Astrofísica, algunas de las más notables que podemos mencionar son la Observatorio Europeo Austral (ESO) y la Agencia Espacial Europea (ESA).

1.1.2.1. La Organización Europea para la Investigación Astronómica en el Hemisferio Sur (ESO)

Es una organización inter-gubernamental con 16 Estados Miembros junto con el estado anfitrión, Chile y Australia como socio estratégico. La sede central de la ESO se encuentra en Garching, cerca de Munich, en Alemania.

El 6 de noviembre de 1963, el Gobierno de Chile y el ESO acordaron ubicar el observatorio astronómico en Chile, convirtiéndose en la principal organización astronómica inter-gubernamental de Europa, consiguiendo el que sería el observatorio astronómico más productivo del mundo. ESO proporciona a los astrónomos

instalaciones de investigación de vanguardia gracias al apoyo de Alemania, Austria, Bélgica, Dinamarca, España, Finlandia, Francia, Irlanda, Italia, Países Bajos, Polonia, Portugal, República Checa, Suecia, Suiza y el Reino Unido, junto con el país anfitrión, Chile.

La misión principal de ESO, es proporcionar a los astrónomos y astrofísicos instalaciones punteras que les permitan desarrollar ciencia de vanguardia en las mejores condiciones, mediante la construcción y operación del conjunto de telescopios terrestres más potentes del mundo. La contribución anual a ESO de los estados miembros asciende a unos 200 millones de euros, y cuenta con alrededor de 700 empleados. Ofrece numerosas posibilidades en transferencia y generación de subproductos de alta tecnología, junto con numerosas oportunidades de contratos, constituyendo un espectacular escaparate para la industria europea[6].

En Chile ESO opera desde las oficinas centrales en Vitacura, además de tres centros de observación: La Silla, Paranal y Chajnantor. La ESO posee algunas de las infraestructuras científicas más complejas creadas por el ser humano, una muestra de ellas sería:

- **El New Technology Telescope (NTT)**: Con 3.58 metros se inauguró en 1989. Abrió nuevos caminos para la ingeniería y el diseño del telescopio y fue el primero en el mundo en tener un espejo principal controlado por computadora. El espejo principal es flexible y su forma se ajusta activamente durante las observaciones, para preservar una calidad de imagen óptima. La posición del espejo secundario también se controla activamente en tres direcciones. Esta tecnología, desarrollada por ESO, es conocida como óptica activa y ahora se aplica a todos los telescopios modernos más importantes, como el "Very Large Telescope" en Cerro Paranal y el futuro "Extremely Large Telescope".

- **ALMA**: Es el telescopio más poderoso del mundo para estudiar el Universo en longitudes de onda submilimétricas y milimétricas, en el límite entre la luz infrarroja y las ondas de radio largas. No utiliza los espejos brillantes y reflectantes de los telescopios de luz visible e infrarroja. Se compone de muchas "antenas" similares a grandes antenas parabólicas metálicas. ALMA consta de 66 antenas, 54 de ellas con platos de 12 metros de diámetro y 12

más pequeñas, con un diámetro de 7 metros cada una.

- **El Extremely Large Telescope (ELT)**: con un espejo principal de 39 metros de diámetro, será el ojo más grande del mundo en el cielo cuando esté operativo a principios de la próxima década. El ELT abordará los desafíos científicos más grandes de nuestro tiempo, como el rastreo de planetas similares a la Tierra alrededor de otras estrellas en las "zonas habitables" donde la vida podría existir. El diseño del telescopio en sí es revolucionario y se basa en un nuevo esquema de cinco espejos que resulta en una calidad de imagen excepcional. El espejo primario consta de casi 800 segmentos, cada uno de 1,4m de ancho, pero de solo 50mm de espesor. El diseño óptico requiere un inmenso espejo secundario de 4.2m de diámetro, más grande que los espejos primarios de cualquiera de los telescopios de ESO en La Silla. Los espejos adaptativos se incorporan a la óptica del telescopio para compensar la falta de claridad en las imágenes estelares introducidas por la turbulencia atmosférica.

1.1.2.2. La Agencia Espacial Europea (ESA)

La Agencia Espacial Europea (ESA) tiene como misión configurar el desarrollo de la capacidad espacial europea. La Agencia Espacial Europea (ESA) está compuesta por 22 Estados Miembros. La coordinación de los recursos económicos e intelectuales de sus miembros permite llevar a cabo programas y actividades de mayor alcance que los que podría realizar cualquier país europeo individualmente.

La ESA elabora el Programa Espacial Europeo. Los programas de la Agencia se diseñan con el fin de conocer más a fondo la Tierra, el entorno espacial que la rodea, el Sistema Solar y el Universo, así como para desarrollar tecnologías y servicios basados en satélites y fomentar la industria europea. La ESA también trabaja en estrecha colaboración con organizaciones espaciales no europeas. Tiene su sede en París y desde allí se toman las decisiones sobre futuros proyectos. No obstante, también dispone de centros distribuidos en el resto de Europa, cada uno con sus respectivas competencias. Además, dispone de oficinas de coordinación en Estados Unidos, Rusia y Bélgica, una base de lanzamientos en la Guayana francesa, y estaciones de aterrizaje y seguimiento en diversas partes del mundo[7].

La ESA participa en importantes misiones, relevantes para el conocimiento del universo, algunos de ellas son:

- **GAIA**: Gaia creará un mapa tridimensional extraordinariamente preciso de más de mil millones de estrellas en toda nuestra galaxia y más allá, mapeando sus movimientos, luminosidad, temperatura y composición. Este gran censo estelar proporcionará los datos necesarios para abordar una enorme gama de problemas importantes relacionados con el origen, la estructura y la historia evolutiva de nuestra galaxia. Las estimaciones de adquisición de datos científicos por GAIA será de unos 100TB aproximados, que una vez procesados alcanzarán aproximadamente 1PB de datos.

- **JUICE**: JUpiter ICy Moons Explorer, primera misión del programa "Cosmic Vision" 2015-2025 de la ESA. Planificado para su lanzamiento en 2022 y su llegada a Júpiter en 2029, pasará al menos tres años haciendo observaciones detalladas del gigante planeta gaseoso Júpiter y tres de sus lunas más grandes, Ganímedes, Calisto y Europa.

- **Athena**: Telescopio de rayos X, avanzado para Astrofísica de alta energía, diseñado para abordar el "El universo caliente y energético". Intentará responder a dos preguntas clave:

  - ¿Cómo se ensambla la materia ordinaria en las estructuras a gran escala que vemos hoy?
  - ¿Cómo crecen y dan forma los agujeros negros al Universo?

  Para abordar la primera pregunta, será necesario mapear las estructuras de gas caliente en el Universo, específicamente el gas en cúmulos y grupos de galaxias, y el medio intergaláctico; determinar sus propiedades físicas y rastrear su evolución a través del tiempo cósmico. Para responder a la segunda pregunta, los agujeros negros supermasivos (SMBH) deben identificarse, incluso en entornos oscuros, y se deben entender las entradas y las salidas de materia y energía a medida que crecen los agujeros negros. El 27 de junio de 2014, Athena fue seleccionada como la segunda misión de clase "Large" en el plan "Cosmic Vision" 2015–25 de la ESA. La Fase A hasta

finales de 2018, finaliza con las Revisiones de Requisitos Preliminares del Instrumento (IPRR). A continuación, la Fase B1 se extenderá hasta el Q3/2019, finalizando con la Revisión de Formulación de la Misión (MFR). Se espera la adopción de la misión por parte del Comité para el Programa Científico (SPC) de la ESA en la segunda mitad de 2021, lo que llevará a su lanzamiento aproximadamente a principios de la década de 2030.

1.1.3. Ciencias de la Tierra y Medioambientales

Dentro de las ciencias medioambientales, podemos diferenciar dos disciplinas científicas principales: las relacionadas con la biología y las ingenierías medioambientales, y las meteorológicas.

Algunas podríamos decir que tienen cientos de años de antigüedad, ya que el ser humano, desde los albores de la agricultura, siempre ha estado pendiente de la climatología. Pero es en los últimos años, cuando están siendo unos de los focos de mayor actividad científica, debido a la más que establecida relación causal entre la actividad humana y el hoy tan obvio cambio climático. Esta causa/efecto es muchas veces directa y observable a simple vista, como la evolución de los ecosistemas de flora y fauna en relación con la actividad humana, ya sea industrial o turística, mientras que otras en cambio, menos evidente a corto plazo, como los diferentes cambios climatológicos, cambios en corrientes transoceánicas, pluviosidad, olas de calor, todas ellas enmarcables en un futuro no tan lejano.

1.1.3.1. Meteorología

Las comunidades relacionadas con las ciencias meteorológicas, se han desarrollado muy rápidamente durante los últimos años. Gracias a los avances computacionales, hoy en día gozamos de predicciones realmente precisas. Esto facilita la toma de decisiones en el día a día de la actividad humana, relativas a ámbitos tan dispares como el turismo, la agricultura o la movilidad.

Una de las principales comunidades de este ámbito es The Earth System Grid Federation (ESGF). La ESGF se formó a partir de la colaboración entre grupos de diferentes instituciones internacionales como el Department of Energy (DoE), la National Aeronautics and Space Administration (NASA), el National Oceanic

and Atmospheric Administration (NOAA), National Science Foundation (NSF) y Laboratorios internacionales como el Max Planck Institute for Meteorology (MPI-M), el German Climate Computing Centre (DKRZ), la Australian National University (ANU), el Australian National Computational Infrastructure (NCI), el Institut Pierre-Simon Laplace (IPSL) o el Centre for Environmental Data Analysis (CEDA)[8]. La ESGF es una colaboración internacional que desarrolla, implementa y mantiene infraestructura de software para la gestión, difusión y análisis de resultados de modelos y datos de observación sobre el cambio climático a nivel mundial, a través del Intergovernmental Panel on Climate Change (IPCC).

El IPCC, fue creado por el Programa de las Naciones Unidas para el Medio Ambiente (ONU Medio Ambiente) y la Organización Meteorológica Mundial (OMM). Determina el estado del conocimiento sobre el cambio climático. Identifica dónde hay acuerdo en la comunidad científica sobre temas relacionados con el cambio climático y dónde se necesita más investigación. Los informes se redactan y revisan en varias etapas, lo que garantiza la objetividad y la transparencia. El IPCC no realiza su propia investigación. Los informes del IPCC son neutrales, relevantes para las políticas, pero no prescriptivos de las políticas. Los informes de evaluación son un aporte clave en las negociaciones internacionales para abordar el cambio climático.

El Coupled Model Intercomparison Project (PCMDI) tiene como misión desarrollar métodos y herramientas mejorados para el diagnóstico y la inter-comparación de prototipos de circulación general (GCM) que simulan el clima global. La necesidad de un análisis innovador de las simulaciones climáticas de GCM es evidente, ya que se desarrollan modelos cada vez más complejos, mientras que los desacuerdos entre estas simulaciones y su relación con las observaciones climáticas siguen siendo significativos y poco conocidos. La misión de PCMDI exige que se trabaje tanto en proyectos científicos, como en tareas de infraestructura. Los proyectos científicos actuales se centran en apoyar la inter-comparación de modelos, desarrollando un banco de pruebas de parametrización, identificando retroalimentaciones robustas, en observaciones y patrones, diseñando métodos estadísticos sólidos para la detección de causas atribuibles al cambio climático. La Tareas de infraestructura en curso incluyen:

- Desarrollo de software para gestión de datos.

- Visualización y computación.
- Organización de conjuntos de datos de observación para la validación del modelo.
- Documentación consistente de las características del modelo climático.

Entre los papeles de PCMDI, está la responsabilidad de liderar la ESGF que almacena y distribuye conjuntos de datos de escala terrestre, de múltiples simulaciones de modelos climáticos globales de atmósfera oceánica acoplada. El análisis exhaustivo de estas simulaciones por parte de miembros de la comunidad climática internacional, proporcionará una base científica importante para los "Informes de Evaluación del Cambio Climático". El volumen de datos estimado para el PCMDI6/IPCC-AR6 en 2020, es de 100PB y 280 millones de ficheros.

ESGF, gestiona la primera base de datos descentralizada para el manejo de datos de ciencia climática, formada por varios Petabytes de datos distribuidos en docenas de sitios federados en todo el mundo. Es reconocida como la infraestructura líder para la gestión y el acceso a grandes volúmenes de datos distribuidos para la investigación del cambio climático. Compatible con el PCMDI, cuyos protocolos permiten las evaluaciones periódicas realizadas por el IPCC.

Utilizando un sistema de nodos "P2P" distribuidos geográficamente (ver fig:1.5), administrados independientemente pero unidos por protocolos e interfaces comunes, la comunidad ESGF posee la principal colección de simulaciones, datos de observación y re-análisis para la investigación del cambio climático. Su principal objetivo es, desarrollar y facilitar el acceso a datos y procedimientos para el análisis de datos climáticos y facilitar los avances en la Ciencia del Sistema Terrestre.

20 de los más potentes supercomputadores del TOP500[10], están dedicados a realizar predicciones meteorológicas. Algunos de los desafíos computacionales más importantes, a los que se enfrentan estas colaboraciones son:

- Gestionar y distribuir de forma eficiente, una cantidad ingente de datos producidos (desde Petabytes a Hexabytes).
- Crear un sistema compatible con las diferentes herramientas de análisis empleados por los clientes finales.
- Federar las múltiples políticas de acceso local.

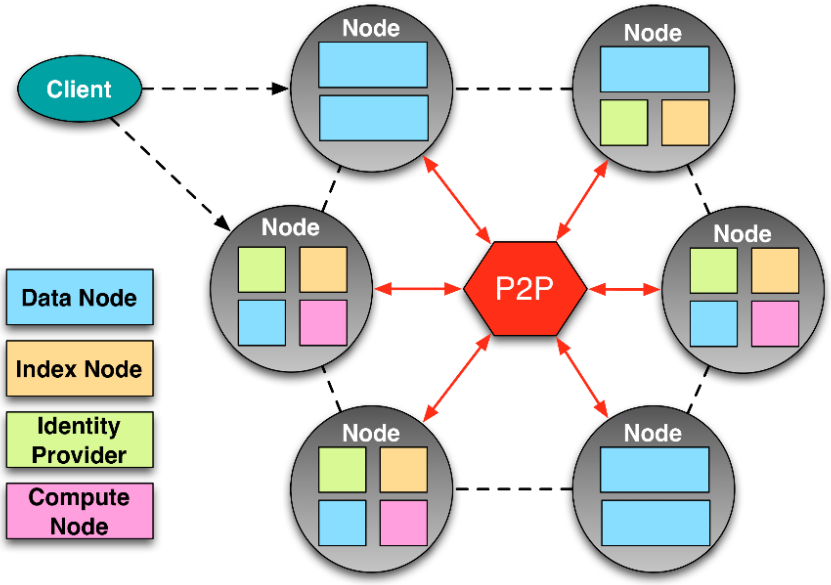


FIGURA 1.5: Modelo Peer to Peer de distribución de datos ESGF. [9]

1.1.3.2. Biodiversidad

En el campo de la biodiversidad, tenemos la colaboración LifeWatch ERIC. Es un Consorcio Europeo de Infraestructura que ofrece instalaciones y plataformas de investigación en e-Ciencia a científicos que buscan aumentar el conocimiento y profundizar en comprensión y la organización de Biodiversidad, las funciones y servicios del Ecosistema, para ayudar a la sociedad a enfrentar los desafíos del futuro[11].

LifeWatch, trata de comprender la evolución y las funciones de la biodiversidad. Aborda factores globales como el clima, la presión demográfica, la contaminación, el consumo de suelo, los responsables de la pérdida continua de diversidad biológica y el correcto funcionamiento de los diversos ecosistemas, con impacto directo en el bienestar y desarrollo de la sociedad actual.

Esto requiere un análisis, tanto de los impactos, como de las decisiones de gestión en un rango de escalas espaciales y temporales:

- Observación y monitorización de datos de ecosistemas.
- Experimentos de laboratorio.

- Gestión de datos medioambientales (distribución, almacenamiento, archivado) de presencia/ausencia de especies, meteorológicos, físicos, químicos biológicos.
- Establecimiento de estándares para garantizar la inter-operabilidad.
- Modelado preciso de la dinámica y procesos de los ecosistemas.

LifeWatch busca comprender las complejas interacciones entre las especies y el medio ambiente, aprovechando los sistemas de computación de alto rendimiento. Para ello emplea herramientas computacionales de vanguardia, Grid, Cloud, Big Data, y Virtual Research Environments que integran distintas aplicaciones o funcionalidades para crear entornos de investigación (laboratorios), con distintas necesidades de recursos. El desarrollo de herramientas avanzadas de modelado para implementar medidas de gestión destinadas a preservar la vida en la Tierra, precisa combinar una amplia gama de herramientas y recursos de TIC con un profundo conocimiento de la materia. La misión de LifeWatch es ser un proveedor mundial de contenido y servicios "de primera clase" para la comunidad de investigación de Biodiversidad ofreciendo:

- Oportunidades para el desarrollo científico a gran escala.
- Captura de datos acelerada con nuevas tecnologías innovadoras.
- Toma de decisiones basadas en el conocimiento para la biodiversidad y la gestión de los ecosistemas.
- Programas de formación, divulgación y sensibilización.

1.1.4. Ciencias Sociales y Bibliotecas Digitales

Las ciencias sociales, debido a su composición intrínseca, alejada de las ciencias aplicadas y la experimentación, han sido posiblemente las últimas en emplear de forma generalizada este tipo de medios tecnológicos. Si bien es verdad que en su ámbito no es necesario la operación de infraestructuras tecnológicas de vanguardia, el uso de estos recursos ha cambiado la forma en la que accedemos a plataformas distribuidas de enseñanza, realizamos paseos virtuales por alguno de

los museos más famosos del mundo o accedemos a sistemas de archivado de todo tipo de información. La cantidad de datos y recursos empleados en este tipo de colaboraciones no es despreciable.

1.1.4.1. DARIAH

Es un consorcio dedicado a mejorar y apoyar la investigación y la enseñanza en Artes y Humanidades. Desarrolla, mantiene y opera una infraestructura en apoyo de las prácticas de investigación basadas en las TIC, ayudando a los investigadores a la hora de construir, analizar e interpretar recursos digitales. Al trabajar con comunidades de práctica, DARIAH reúne actividades individuales de arte digital y humanidades, y escala sus resultados a nivel europeo. Preserva, proporciona acceso y difunde la investigación que surge de estas colaboraciones y asegura que se sigan los estándares metodológicos y técnicos.

DARIAH se estableció como un Consorcio Europeo de Infraestructura de Investigación (ERIC) en agosto de 2014 y actualmente, tiene 17 miembros y varios socios cooperantes en once países no miembros. Su carácter supranacional e interdisciplinario, proporciona valor a sus miembros y partes interesadas a través de la validación y el intercambio de datos, servicios y herramientas, proporcionando oportunidades de capacitación y educación al permitir una organización jerárquica en torno a las necesidades de investigación emergentes. Promueve el desarrollo de métodos de investigación en las artes y las humanidades, documentando el estado del arte, apoyando la preservación y curación de datos de investigación con un enfoque en desafíos particulares que incluyen diversidad, procedencia, colecciones multimedia y granularidad, actuando como coordinador e integrador para una comunidad diversa.

DARIAH[12] opera a través de las redes europeas de los "Virtual Competence Centers" (VCC). Cada uno de los cuatro VCC es inter-disciplinario, multi-institucional, internacional y se centra en un área específica de especialización, así:

- **VCC1**: Mantiene un entorno digital que permite compartir datos y herramientas desarrollados por la comunidad y garantiza la calidad, la permanencia y el crecimiento de los servicios técnicos para las artes y las

humanidades.

- **VCC2**: Es el enlace de investigación y educación actuando como interfaz de enlace con las comunidades de investigación y enseñanza.
- **VCC3**: Se ocupa de la gestión de contenido académico en las diversas etapas, desde la creación, la curación y la difusión, hasta la combinación de recursos digitales académicos y resultados para su reutilización posterior.
- **VCC4**: Se enfoca en abogacía, en las artes y en las humanidades.

Dentro de esta estructura, DARIAH cuenta con más de 20 grupos de trabajo dinámicos para integrar los servicios nacionales en categorías operativas específicas.

1.1.4.2. INSPIRE

Los laboratorios de física CERN, Deutsches Elektronen-Synchrotron (DESY), SLAC National Accelerator Laboratory (SLAC) y Fermilab crearon un nuevo Sistema de Información Científica para física de alta energía llamado INSPIRE. INSPIRE-HEP es una biblioteca digital de acceso abierto para el campo de la física de alta energía (HEP). Es la combinación la base de datos SPIERS y la tecnología de biblioteca digital Invenio[13] desarrollada por el CERN. Proporciona una visión para la gestión de la información en otros campos de la ciencia, al interactuar con otros proveedores de servicios como arXiv[14], "Particle Data Group" y el Sistema de Datos de Astrofísica de la NASA.

INSPIRE-HEP es mucho más que un simple repositorio, proporciona información como métricas de citas, gráficos extraídos de documentos, notas de experimentos internos, además de herramientas para que los usuarios mejoren metadatos como el "crowdsourcing" para la desambiguación de autores. Mediante HEPNames gestiona un directorio integral de personas involucradas en instituciones de física de alta energía. Dispone de varias bases de datos con más de 7.000 institutos relacionados al campo HEP vinculando, sobre cada instituto, todos los documentos que INSPIRE-HEP tiene asociados con la institución, así como una lista de personas obtenidas desde HEPNames[15]. También dispone de una colección de reuniones, conferencias, lista de académicos con trabajos de interés para la comunidad en física de alta energía, física nuclear y astrofísica.

Para cuantificar la complejidad de las colaboraciones científicas descritas anteriormente, podemos observar la siguiente tabla (ver tab:1.1). Son colaboraciones muy prolíficas desde el punto de vista científico, manejan presupuestos muy elevados, con de miles colaboradores, y gestionan enormes y complejas infraestructuras. Todas ellas unidas por el nexo común de las ingentes cantidad de datos que generan, y la problemática de como abordar su gestión y preservación.

Tabla 1.1: Resumen de las principales Colaboraciones científicas.

	Colaboraciones	Vol. datos	Personal	Art/año	Pto
Física de alta Energía (HEP)	CERN (Atlas,CMS) (Alice,LHBc)	$\sim 10^2$ PB	$\sim 10^3$	$\sim 3*10^2$	$\sim$750M€
Astrofísica	ESO,ESA	$\sim 10^2$ PB	$\sim 10^3$	$\sim 10^3$	$\sim$6000M€
Ciencias de la Tierra y Mediambientales	ESGF,LifeWatch	$\sim 10^2$ PB	$\sim 10^3$	$\sim$2*10	—
Ciencias Sociales y Bibiotecas Digitales	DARIAH,INSPIRE	$\sim 10^2$ TB	$\sim 10^2$	$\sim$2*10	—

1.2. Organizaciones virtuales

Tal como se comentaba en la introducción de este capítulo 1, la revolución en las tecnologías de la información (TIC), así como los servicios que fluyen a a través de Internet, WWW, e-mail, redes sociales, videoconferencia, han permitido que se lleven cambios en la forma en que las organizaciones e individuos se relacionan. Esto cambios son forzados por fenómenos estructurales como la globalización, la economía o la libre competencia de mercado. Los individuos y las organizaciones responden a estos fenómenos estructurales implementando estrategias que les permitan competir en el nuevo entorno. Pero las estrategias que se pueden implementar están limitadas a la tecnología disponible.

Durante la mayor parte del siglo XX, los fenómenos estructurales estaban basados en tecnologías de coordinación de la era industrial como el tren, el teléfono y sus derivados o más tarde la computadora. Con estos medios, las compañías eran capaces de administrar grandes organizaciones de forma centralizada.

Con la llegada de las nuevas tecnologías basadas en Internet, el paradigma de las organizaciones cambia. La información fluye de manera instantánea a un bajo costo, independientemente de la distancia. El valor de la centralización y la burocracia disminuye.

Gracias a las TIC, organizaciones pequeñas pueden acceder a grandes bancos de información, antes solo disponibles para las grandes organizaciones.

La organización del nuevo siglo se llama Organización Virtual. Definimos Virtual, como algo que no es tangible. Una definición posible de organización virtual es "una red temporal de organizaciones independientes unidas por la tecnología para compartir conocimientos, costos y objetivos comunes". Suelen tener organización distribuida geográficamente, aunque pueden tener una sede central, y su trabajo es coordinado principalmente, por medios de comunicación electrónica.

Las condiciones siguientes, deberían ser extrapolables para cualquier tipo de Organización Virtual (VO), ya sea del sector público, como puede ser el caso de colaboraciones científicas, o del sector privado, en el caso de organizaciones empresariales e incluso de tipo mixto:

- Las distintas unidades que componen la organización deben estar distribuidas y en ningún caso emplazadas en el mismo ámbito geográfico.
- Uso de las TIC para la interacción entre las unidades que la conforman.
- Flexibilidad sin abandonar la eficacia.
- Confianza entre las distintas partes que forman la VO, know-how, recursos, etc.
- Jerarquía horizontal o formalización, autocontrol y autoregulación. Ninguna parte se impone sobre las demás.
- Explotar las competencias distintivas, que cada parte puede aportar en forma de recursos, habilidades y conocimiento.
- La habilidad de aprender y de cambiar, a partir de la cooperación.

1.2.1. Roles

La gestión interna de la VO usa políticas basadas en roles. Los Roles consisten en un conjunto de tareas y competencias relacionadas con la creación, operación y desarrollo de la VO y su estructura. El rol puede ser asumido por una sola persona o una entidad organizativa dentro de la estructura de la VO.

Diferentes usuarios pueden tener diferentes roles, o incluso un usuario determinado puede tener varios roles diferentes dentro de cada VO a la que pertenece. Estos roles son asignados por los administradores de la VO en función de las necesidades que un usuario tiene para poder desarrollar su labor.

1.3. Foro Estratégico Europeo sobre Infraestructuras de Investigación (ESFRI)

Es un Foro que desempeña un papel clave en la formulación de políticas sobre infraestructuras de investigación en Europa. Está compuesto por delegados nacionales designados por ministros de investigación de países de la UE y países asociados con Horizonte 2020. También incluye un representante de la Comisión. Es un organismo autorregulado, que funciona por consenso y generalmente se reúne cuatro veces al año.

El desarrollo de su trabajo se basa en establecer una hoja de ruta europea para las infraestructuras de investigación en los próximos 10-20 años, estimulando su implementación, apoyando un enfoque coherente basado en la estrategias sobre infraestructuras de investigación en Europa. Debe actuar como incubadora de nuevas iniciativas, realizando el seguimiento de la implementación del proyecto, evaluando e implementando el Espacio Europeo de Investigación (EEI).

Capítulo 2

Gestión de una Infraestructura de Datos para el LHC

El Gran Colosionador de Hadrones (LHC) es, en la actualidad, la instalación científica más grande construida por el hombre. Más de 5.000 físicos y 3.000 ingenieros de todo el mundo han participado en su diseño y operación. Es la máquina con el mayor campo magnético creado hasta el momento, entre 8∼9T, con ∼9.300 imanes, que se utilizan para curvar la trayectoria y acelerar las partículas con carga eléctrica, alcanzando velocidades muy cercanas a la velocidad de la luz (99.9999991 %). Dado que la intensidad del campo magnético es proporcional a la cantidad de corriente eléctrica, para lograr campos magnéticos de 9T de potencia, necesitamos usar imanes superconductores. Para ello, se elimina la resistencia al paso de la corriente eléctrica enfriandolos, estableciendo una temperatura de operación en el LHC de 1.9 Kelvin.

En el LHC chocan dos hadrones, que pueden ser protones (p+) o iones de Pb (82+), pudiendo generar en el punto de colisión, una energía máxima de 14TeV (7TeV para cada haz). Durante el "**Run 2**", 2015-2018, la energía máxima alcanzada ha sido de 13TeV. Los hadrones se lanzan en paquetes, espaciados entre si 7,5 metros aproximadamente, dando lugar a unos 600 millones de colisiones por segundo. Las partículas chocan a altas energías dentro de los cuatro detectores instalados, creando nuevas partículas que se desintegran de formas complejas, y a medida que cruzan las capas de los detectores, dejan un rastro de señales.

Los detectores registran el paso de cada partícula, convirtiendo los caminos y las energías de las partículas en señales eléctricas, combinando la información para crear un resumen digital: el evento "de la colisión". El tamaño de los datos en bruto, por evento, es de alrededor de un millón de bytes (1Mb). El volumen de datos del LHC fue aproximadamente de 70PB durante el año 2018.

El flujo de datos de los cuatro experimentos para el "**Run 2**" ha sido aproximadamente de:

- ALICE: 4GB/s (Pb-Pb)
- ATLAS: 800MB/s∼1GB/s
- CMS: 600MB/s
- LHCb: 750MB/s

2.1. Experimentos y Detectores del LHC

En la parte introductoria de esta tesis, vimos que la infraestructura del LHC es compartida por varios experimentos diferentes (ver fig:2.1), dos de ellos de propósito general y otros dos dedicados al estudio de procesos físicos más concretos.

La existencia de dos experimentos que tienen el mismo propósito como son A Toroidal LHC ApparatuS (ATLAS) y Compact Muon Solenoid (CMS), no tiene otra finalidad que corroborar los nuevos descubrimientos de manera transparente e inequívoca. En otras palabras, tomando dos caminos y planteamientos diferentes, obtener datos y resultados estadísticamente independientes, que nos lleven a la misma conclusión.

Los científicos de ATLAS y CMS intentan demostrar si el modelo estándar de la física de partículas, es satisfactorio, o si debemos que prepararnos para una física más allá del modelo estándar (SM).

Los detectores de partículas son dispositivos muy complejos, no solo desde el punto de vista de la ingeniería, sino por toda la física involucrada en su diseño. La forma más común es el diseño en capas (ver fig:2.2), cada una de ellas planteada para la máxima eficiencia posible en la detección de las diferentes partículas creadas en las colisiones. Podemos diferenciar cuatro capas principales:

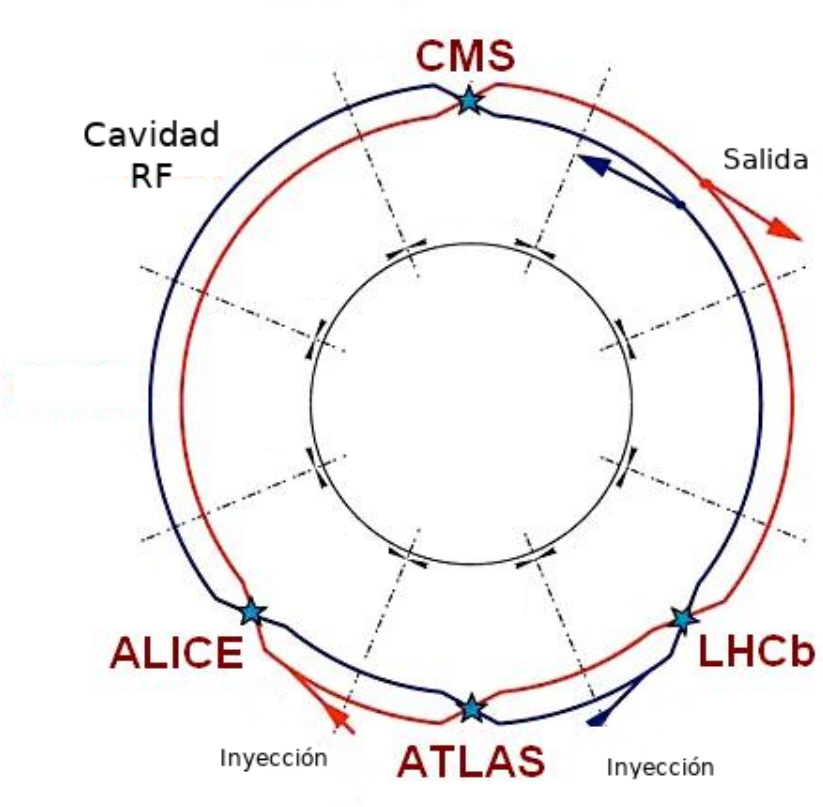


FIGURA 2.1: Descripción del LHC.
[16]

- **Detector Interno**: Mide la trayectoria de las partículas de forma muy precisa y es capaz de estimar de forma muy precisa el momento de las partículas cargadas gracias a su curvatura debida al campo magnético.

- **Calorímetro**: Mide con precisión las energías de fotones, electrones o hadrones (protones, neutrones, piones y kaones) que la atraviesan.

- **Imanes Internos**: Este imán permite demás determinar la relación entre la masa y la carga de las partículas que lo atraviesan a partir del análisis de la curvatura.

- **Espectrómetro de Muones**: Mide con precisión el momento de los muones, que son capaces de atravesar las partes internas del detector. Los muones son partículas cargadas como los e^+ y los e^-, pero son 200 veces más pesadas.

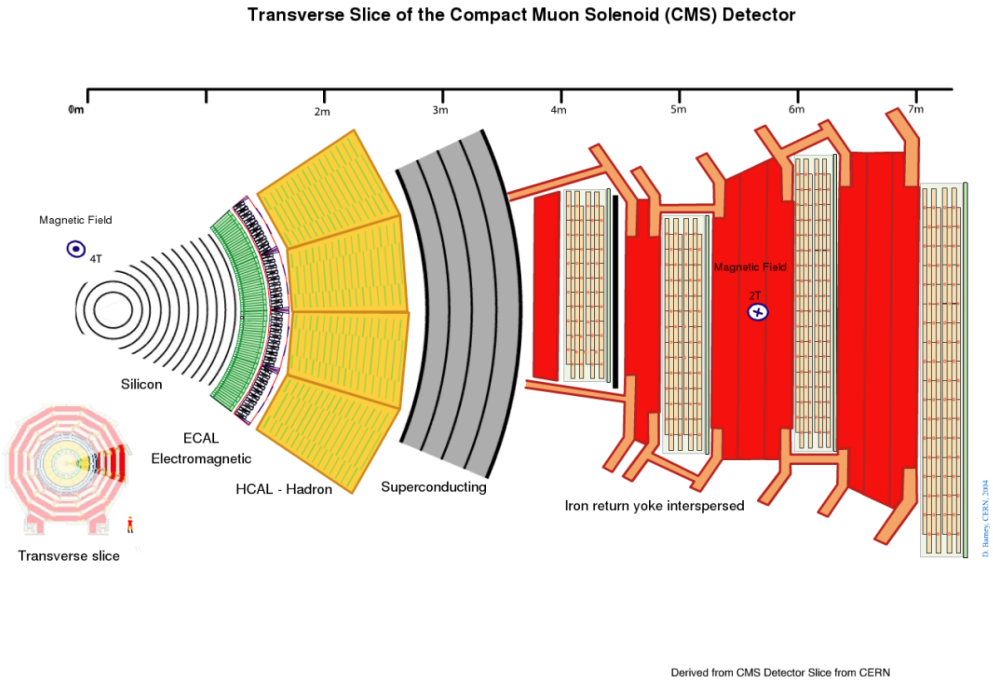


FIGURA 2.2: Modelo en capas del detector CMS.

Aún compartiendo el mismo modelo de capas, cada experimento ha diseñado su propio detector con diversas implementaciones, dando lugar a detectores de diferentes, tamaños, masas y geometrías. A continuación presentaremos las cuatro colaboraciones principales, sus detectores, las tecnologías que han empleado para su desarrollo y la tasa de datos que es generada por cada uno de ellos[17].

2.1.1. ATLAS

La colaboración de ATLAS está formada por unos 3.000 científicos de 182 instituciones en 32 países. Su detector, es uno de los dos más utilizados debido a su propósito general. Sus dimensiones son 44m de largo y 25m de diámetro, siendo el más grande y voluminoso de los cuatro. Su diseño sigue el modelo de capas de otros detectores (ver fig:2.3), pero con algunas peculiaridades.

- **El Detector Interno**: Es la parte central del detector, es muy compacto y altamente sensible. Se compone de tres sistemas diferentes de sensores; detector de píxeles, rastreador de semiconductores (SCT) y rastreador de

radiación de transición (TRT), todos inmersos en un campo magnético paralelo al eje del haz. El detector interno mide la dirección, el momento y la carga de partículas cargadas eléctricamente producidas en cada colisión p^{+}-p^{+}.

- **Los Calorímetros**: Para medir la energía de una partícula, el calorímetro debe lograr que esta interactúe, dando lugar a una cascada de nuevas partículas, y absorbiendo la mayoría de ellas generadas en la colisión, para ello está compuesto de materiales "pasivos" o "absorbentes" de alta densidad. El sistema de calorimetría ATLAS está formado por dos calorímetro; el calorímetro electromagnético (ECAL) de argón líquido (LAr) y el hadrónico (HCAL) con acero como material absorbente. El calorímetro electromagnético mide la energía de electrones y fotones a medida que interactúan con la materia, mientras que el calorímetro hadrónico toma muestras de la energía de los hadrones a medida que interactúan con los núcleos atómicos. Los calorímetros detienen la mayoría de las partículas conocidas, con excepción de muones y neutrinos.

- **Imán**: ATLAS ofrece un sistema híbrido de cuatro imanes superconductores: un solenoide central rodeado por dos toroides extremos (End-cap) y un sistema toroidal "de barril" (BT). Las dimensiones de este sistema magnético son 20m de diámetro y 26m en longitud. Con sus cerca de 2GJ de energía almacenada, es realmente el imán superconductor más grande del mundo. El solenoide central, de 5.5 toneladas de peso, 2.5m de diámetro y 5.3m de largo, proporciona un campo magnético axial de 2T en el centro del área de "tracking" de ATLAS. Dado que este solenoide precede al calorímetro electromagnético, su espesor debe ser el mínimo posible para permitir la máxima respuesta del calorímetro. Contiene 9km de cables superconductores enfriados por helio líquido, por los que circula una corriente eléctrica de 8000A. ATLAS posee también un enorme sistema magnético toroidal superconductor (Barrel Toroid - BT) con unas dimensiones de 25m largo y 22m de diámetro. Este sistema toroidal proporciona el campo magnético para las áreas de detección muónica. El toroide está compuesto por 8 estructuras de 25m x 5m por donde circulan corrientes superconductoras de 20.500A.

- **Espectrómetro de Muones**: Su tarea es la detección de muones. Tiene 4.000 cámaras de muones individuales y cuatro tecnologías diferentes. Funciona de manera parecida al detector interno, curvando la trayectoria de los muones para poder identificar su momento, tiene menor precisión espacial y un volumen mucho mayor. Los muones son indicativos de muchos procesos físicos.

- **Sistema de Trigger**: ATLAS puede procesar hasta 1.700 millones de colisiones p^+-p^+ por segundo, con un volumen de datos combinado de más de $60 \cdot 10^6 MB/s$. El "Trigger System", selecciona aquellos eventos con características distintivas que los hacen interesantes para los análisis físicos, reduciendo así el volumen de datos a 800MB/s aproximadamente.

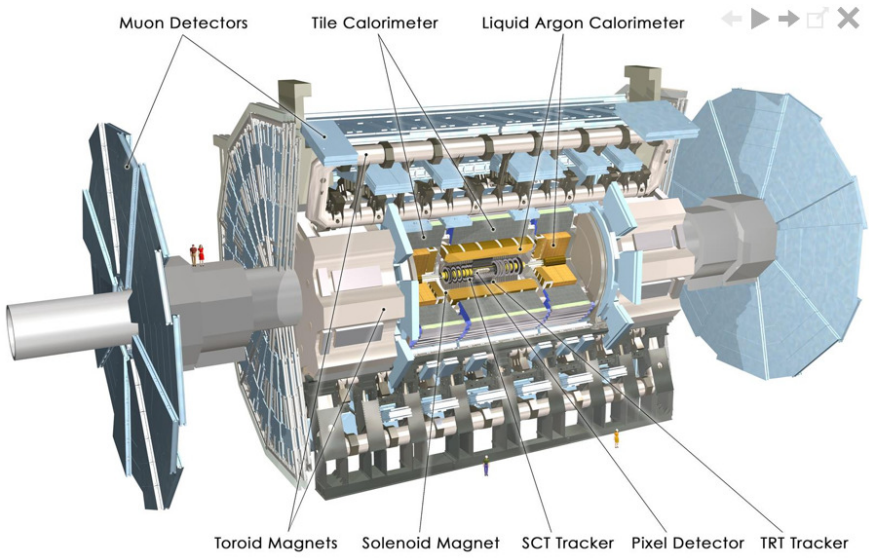


FIGURA 2.3: Detector ATLAS.
[18]

2.1.2. CMS

La colaboración de CMS reúne a miembros de la comunidad Física de Altas Energías (HEP) de todo el mundo en un desafío para promover el conocimiento de la humanidad de las leyes más básicas de nuestro Universo. CMS tiene más

de 4.000 físicos de partículas, ingenieros, informáticos, técnicos y estudiantes de alrededor de 200 institutos y universidades en más de 40 países. La colaboración opera y recopila datos de uno de los detectores de partículas de propósito general (ver fig:2.4) del LHC del Laboratorio Europeo de Física de Partículas (CERN). Colaboradores de todo el mundo han diseñado y fabricado los componentes en sus institutos locales, que luego fueron llevados al CERN para el ensamblaje final. Es el más pesado de los 4 detectores con unas 12.500T.

- **El Detector Interno**: Registra los caminos tomados por las partículas cargadas al determinar sus posiciones en varios puntos clave. Puede reconstruir los caminos de muones de alta energía, electrones y hadrones, así como ver las huellas que provienen de la descomposición de partículas de vida muy corta. Esta hecho totalmente de silicio.

- **Los Calorímetros**: El calorímetro electromagnético (ECAL), es la capa más interna de los dos. Está formado por cristales de tungstato de plomo, que es más pesado que el acero inoxidable, pero con un toque de oxígeno, confiriéndole una forma cristalina muy transparente. El ECAL "centellea" cuando los electrones y fotones lo atraviesan, produciendo luz en ráfagas rápidas, cortas y bien definidas. Unas fibras ópticas especiales recogen esta luz y la alimentan en cajas de lectura donde los fotodetectores amplifican la señal. La cantidad total de luz registrada, es una medida de la energía de la partícula. El calorímetro hadrónico (HCAL) se encarga de detectar partículas más penetrantes, sometidas a la interacción fuerte. Usará los mismos métodos que el ECAL, pero con un material más denso (Bronce o Acero) para detener todas las partículas que el calorímetro electromagnético no pudo.

- **El Imán**: El imán está formado por tres secciones: la bobina superconductora, el tanque de vacío y el núcleo de hierro. La bobina produce el campo axial, mientras que el núcleo es el responsable del retorno del flujo magnético en la parte exterior del solenoide. Este retorno del flujo es el que conforma el conjunto de líneas de fuerza que llenan el detector en todo su volumen paralelamente al eje, y que curvarán las trayectorias de las partículas cargadas que se produzcan en las colisiones en el centro del detector. El imán

solenoidal de CMS, es su parte más característica. Está formado por una bobina cilíndrica de fibras superconductoras y puede generar un campo magnético de casi 4T.

- **Espectrómetro de muones**: La detección de muones es una de las tareas más importantes de CMS (una de las desintegraciones del Bosón de Higgs produce cuatro muones). En total hay 1400 cámaras de muones: 250 tubos de deriva (DT) y 540 cámaras de tira catódica (CSC) que rastrean las posiciones de las partículas, mientras que 610 cámaras de placas resistivas (RPC) forman un sistema de "trigger" redundante, que rápidamente decide mantener los datos muon adquiridos o no. Debido a las muchas capas de detectores y las diferentes especialidades de cada tipo, el sistema es naturalmente robusto y capaz de filtrar el ruido de fondo.

- **Sistema de Trigger**: Se producirán unos $40 \cdot 10^6$ colisiones por segundo. Las "firmas" de cada partícula son analizadas por sistemas electrónicos veloces que guardarán aquellos eventos, unos 100∼150 por segundo, que muestran indicios de nuevas partículas o eventos, reduciendo así el volumen de datos a 600MB/s aproximadamente.

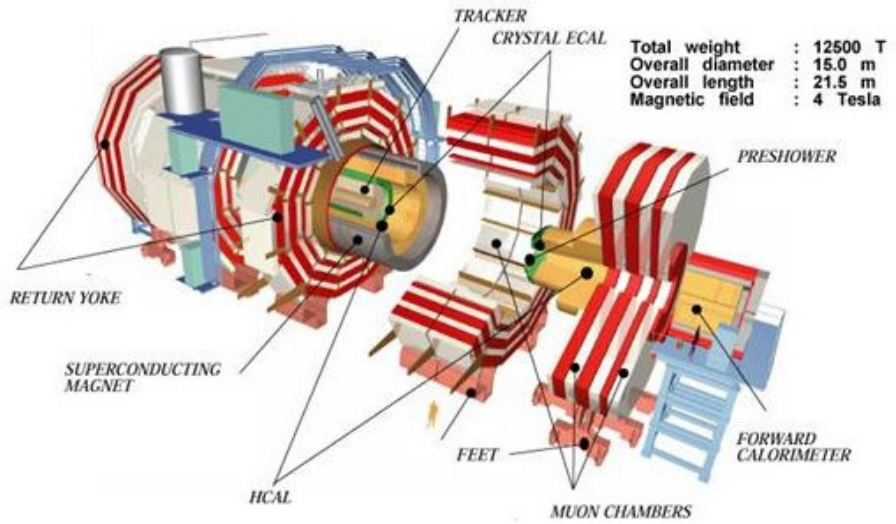


Figura 2.4: Detector CMS.
[19]

2.1.3. ALICE

A Large Ion Collider Experiment (ALICE) involucra a más de 1500 físicos, ingenieros y técnicos, incluidos alrededor de 350 estudiantes graduados, de 154 institutos de física en 37 países de todo el mundo. ALICE intenta buscar respuestas a cuestiones fundamentales como:

- ¿Qué le sucede a la materia cuando se calienta a 100.000 veces la temperatura en el centro del Sol?
- ¿Por qué los protones y los neutrones pesan 100 veces más que los quarks de los que están hechos?
- ¿Se pueden liberar los "quarks"dentro de los protones y neutrones?

Es un experimento dedicado a la física nuclear. Durante aproximadamente tres semanas al año, el LHC colisiona núcleos atómicos pesados (ver fig:2.5). Estas colisiones tienen suficiente energía para generar un plasma de quark-gluón, un estado de la materia que los científicos creen que existió en el universo temprano, justo después del Big Bang. El aumento de la luminosidad, en 2021 y luego en el proyecto High-Luminosity LHC (HL-LHC), abrirá una gama de posibilidades y desafíos, permitiendo estudiar fenómenos raros y realizar mediciones de alta precisión, arrojando luz sobre la termodinámica, la evolución y el flujo del Quark-Gluon Plasma (QGP).

- **El Detector Interno**: Consta de tres secciones:
  - Silicon Pixel Detector (SPD) y Silicon Drift Detector (SDD).
  - Silicon Strip Detector (SSD) y Time Projection Camera (TPC).
  - Transition Radiation Detector (TRD).

  Mide en muchos puntos el paso de cada partícula que lleva una carga eléctrica y proporciona información precisa sobre la trayectoria de la misma.
- **Calorímetro**: Mide la energía de las partículas y determina si tienen interacciones electromagnéticas o hadrónicas. La luz emitida por un objeto caliente, nos informan sobre la temperatura del sistema. Los detectores de

cristales de tungstato de plomo del Photon Spectrometer (PHOS), permiten medir con una gran precisión una región delimitada, mientras que el Photon Multiplicity Detector (PMD) y en particular el EMCal realizan medidas en un área más amplia.

- **Los Imanes Internos**: Un solenoide magnético genera campo magnético de 0,5T que curva las trayectorias de las partículas cargadas.

- **El Espectrómetro de Muones**: El espectrómetro de muones presenta un absorción frontal muy gruesa. Está formado por un complejo filtro de muones adicional sobre una pared de hierro de 1,2m de espesor. Los candidatos a muon son seleccionados de las trazas que penetran los filtros.

- **Sistema de Trigger**: El sistema de adquisición necesita equilibrar su capacidad para registrar el flujo constante de eventos de gran tamaño, generados como resultado de colisiones centrales, con la capacidad de seleccionar y registrar procesos raros de sección transversal. Estos requisitos resultan en un evento agregado que genera una necesidad de almacenamiento de hasta 4GB/s, más de 1PB de datos anuales.

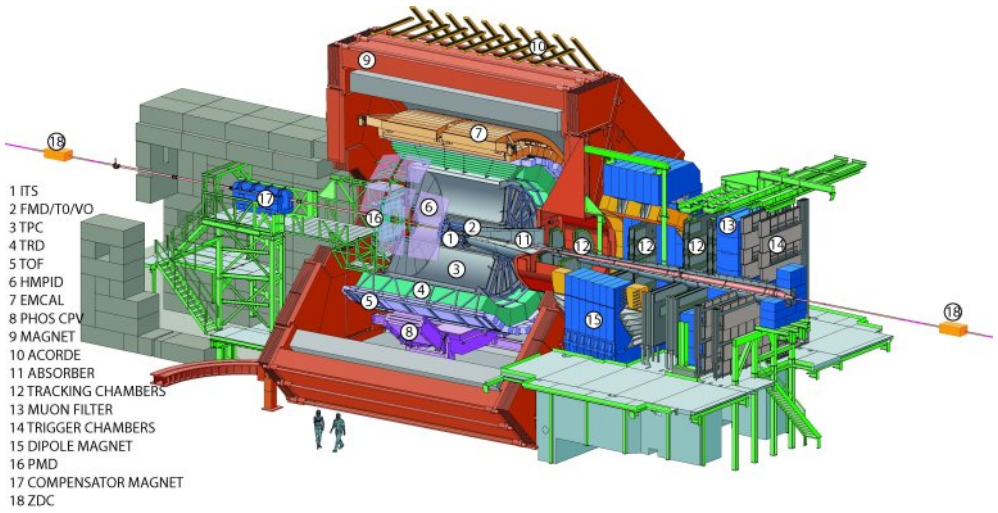


Figura 2.5: Detector Alice.
[20]

2.1.4. LHCb

Formada por aproximadamente 1.260 físicos de partículas, ingenieros, informáticos, técnicos y estudiantes de alrededor de 74 institutos científicos,representando a 16 países. Sus estudios intentan arrojar luz sobre la asimetría materia-antimateria. No sigue el modelo de capas de sus hermanos mayores (ver fig:2.6), si no que se trata de un espectrómetro formado por subdetectores planos perpendiculares al haz incidente, muy diferente en este aspecto de otros experimentos con forma cilíndrica como ATLAS o CMS. Las partículas con contenido de quarks b se producen en ambos sentidos, de forma simétrica respecto del punto de colisión.

- **El Detector Interno**: Consta de una serie de cuatro grandes estaciones rectangulares, cada una de las cuales cubre un área de aproximadamente $40m^2$. Emplea tecnologías de detección diferentes. El "Traker de silicio" se coloca cerca de la linea del haz. Las partículas cargadas chocan con los átomos de silicio, liberando electrones y generando una corriente eléctrica, indicando el camino de la partícula original. El "Traker" externo está formado por miles de tubos llenos de gas. Cada vez que pasa una partícula cargada, ioniza las moléculas de gas y produce electrones.

- **Calorímetro**: Ambos calorímetros tienen una estructura tipo sándwich, con capas alternas de placas de metal y plástico. Cuando las partículas golpean las placas de metal, producen lluvias de partículas secundarias. Estas, a su vez, excitan las moléculas de poliestireno dentro de las placas de plástico, que emiten luz ultravioleta. La cantidad de UV producida es proporcional a la energía de las partículas que llegan al calorímetro.

- **Imán**: El enorme imán del experimento consta de dos bobinas, ambas con un peso de 27 toneladas, montadas dentro de un marco de acero de 1.450 toneladas. Las partículas normalmente viajan en línea recta, pero la presencia de un campo magnético hace que las trayectorias de las partículas cargadas se curven y examinando la curvatura, es posible calcular el momento y así establecer su identidad.

- **Espectrómetro de Muones**: Ubicado en el extremo más alejado del detector, el sistema de muones comprende cinco "estaciones" rectangulares, que

aumentan gradualmente de tamaño y cubren un área agregada de 435m^2. Cada estación contiene cámaras llenas de una combinación de tres gases: dióxido de carbono, argón y tetrafluorometano. Los muones que pasan reaccionan con esta mezcla, y unos electrodos detectan las interacciones que se producen.

- **Sistema de Trigger**: El detector registra aproximadamente 10 millones de colisiones de protones por segundo. El "Trigger, selecciona alrededor de 1 millón de eventos por segundo para su posterior procesamiento, mientras descarta información de los 9 millones restantes, generando una necesidad de almacenamiento de unos 750MB/s.

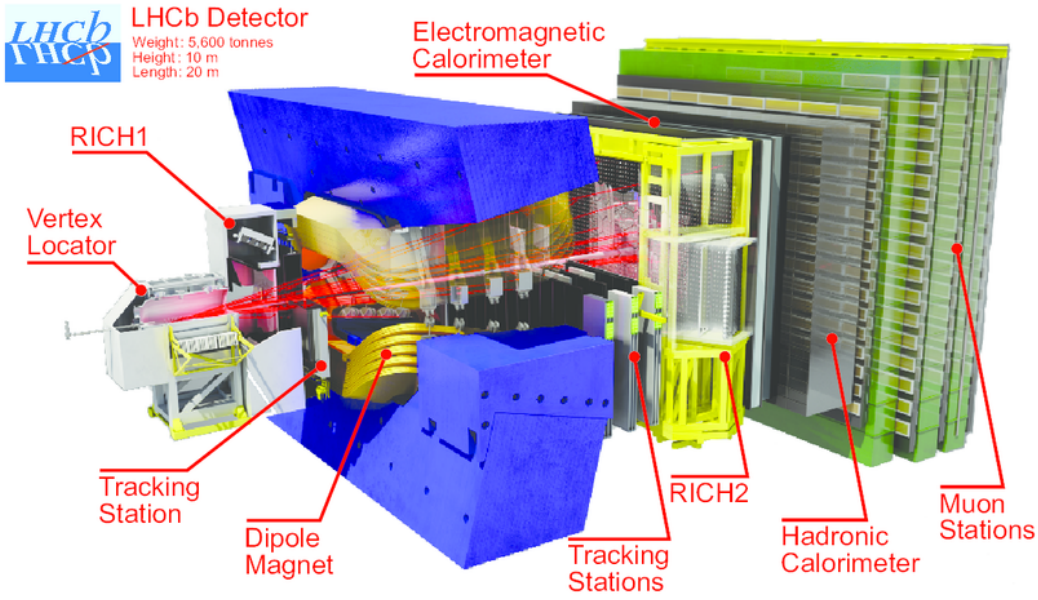


FIGURA 2.6: Detector LHCb.
[21]

2.2. Tasas de datos y flujo de trabajo

Los experimentos de colisión de partículas producen grandes cantidades de datos. Los haces en el LHC están formados por paquetes de protones, separados entre si 25 nanosegundos ($25{\cdot}10^{-9}s{\cdot}300{,}000\frac{km}{s}{\cdot}10^{3}\frac{m}{km} = 25{\cdot}10^{-9}s{\cdot}3{\cdot}10^{8}\frac{m}{s} = 7,5m$), y cada uno de ellos contiene más de 100 mil millones de protones. Con estos números, 2.556 es el número máximo posible de paquetes que se pueden alcanzar en el túnel de 27km, con el método de preparación que se usa actualmente. Los paquetes de partículas que se inyectan al LHC se preparan y aceleran mediante una cadena de cuatro aceleradores. Desde 2018, se establece un nuevo método para agrupar y dividir los paquetes permitiendo que las partículas se puedan juntar aún más. Con un número igual de protones, el diámetro del haz se ha reducido en un 40 %. Grupos más densos, significan una mayor probabilidad de colisiones en el centro de los detectores. Este éxito ha llevado a un nuevo récord de luminosidad para el LHC de $1{,}58 \cdot 10^{34}cm^{-2}s^{-1}$. La luminosidad mide el número de colisiones potenciales por segundo y por unidad de área y esta nueva cifra supera las expectativas iniciales definidas para los diseños originales del LHC, que esperaban alcanzar un máximo de $1 \cdot 10^{34}cm^{-2}s^{-1}$. Una mayor luminosidad implica más colisiones y por lo tanto más datos generados por los experimentos.

Las partículas chocan en los detectores LHC aproximadamente 1 billón de veces por segundo, esto significa aproximadamente un Petabyte (PB) de datos de colisión por segundo. Sin embargo, tales cantidades de datos son imposibles de registrar para los sistemas informáticos actuales y, por lo tanto, son filtrados por los experimentos, manteniendo solo los más "interesantes". Estos datos filtrados de LHC son enviados al Data Center (DC) del CERN, donde se realiza la reconstrucción inicial de los datos y donde se archiva una copia al almacenamiento de cinta a largo plazo. Incluso después de la drástica reducción de datos realizada por los experimentos, el DC del CERN procesa aproximadamente 70PB de datos.

La cantidad de datos procesados aumentará significativamente cuando el CERN entre en alta luminosidad, estimada para el año 2026 (ver fig:2.7). Se esperan aproximadamente 20 veces los datos almacenados en disco a finales de 2018 y será necesario un aumento de un orden de magnitud en los recursos informáticos requeridos para manejar esta cantidad[22]. En consecuencia, se están

batiendo récords en muchos aspectos en la adquisición de datos, dando lugar a niveles excepcionales de uso tanto en los recursos de procesamiento, como de almacenamiento. Para enfrentar estos desafíos, la infraestructura informática en general de los centros involucrados en cada una de la comunidades vistas anteriormente, ha de mejorar de forma radical en dos aspectos:

- Capacidad de los sistemas de almacenamiento: Volumen y Rendimiento.
- Transferencia de datos: incremento en la capacidad de ancho de banda.

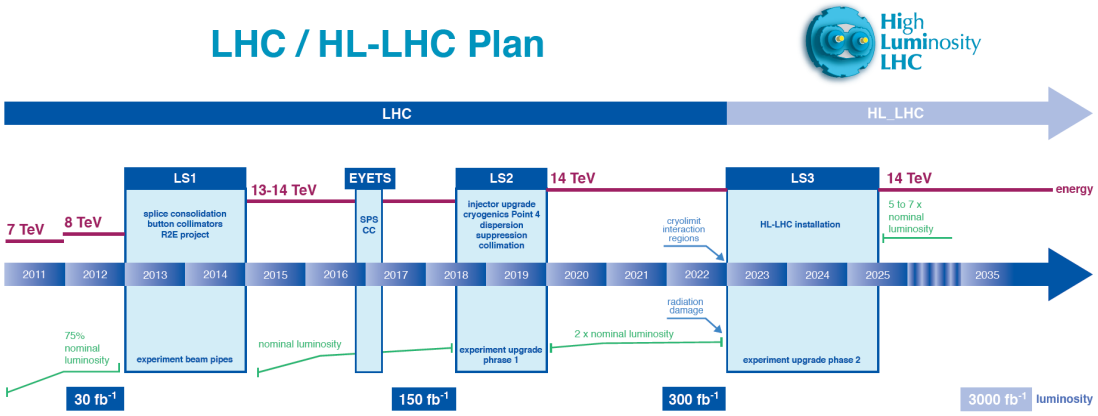


FIGURA 2.7: Plan de incremento de luminosidad para el LHC.

2.3. El entorno Grid

La computación Grid, es un paradigma de computación distribuida. Es una implementación computacional con más de 20 años de vida. Nació ante la necesidad de abordar la gestión y el procesado de ingentes cantidades de información para los que una sola institución no estaba preparada, ni en requerimientos materiales, ni en humanos, ni en los económicos.

La solución fue desarrollar un infraestructura distribuida capaz de agregar y acceder a estos recursos, de forma transparente, independientemente de su localización geográfica. Desde el punto de vista del usuario, los recursos computacionales aportados por cada institución participante, son accesibles como si de un único

sistema se tratase. Como vemos esta definición encaja perfectamente en la forma de trabajo definida para las Organizaciones Virtuales (VOs).

2.3.1. Organizaciones virtuales en entornos computacionales

Como hemos establecido en anteriormente (ver 1.2), la Organización Virtual (VO) es un grupo de personas, si nos referimos a un entorno computacional, hablaremos de científicos o investigadores, con intereses y requisitos comunes, que necesitan trabajar en colaboración y/o compartir recursos, datos, software, CPU y espacio de almacenamiento, independientemente de su ubicación geográfica. Se unen a un VO para acceder a recursos y satisfacer estas necesidades, después de acordar un conjunto de reglas y Políticas que rigen sus derechos de acceso y seguridad.

En el caso de la Infraestructure Grid Europea (EGI) de la que hablaremos en capìtulos posteriores, tanto la creación de una nueva VO, como la asignación de roles o el alta de nuevos usuarios es un proceso bien definido[23].

La responsabilidad del proceso del ciclo de vida una VO, recae en dos figuras:

- **El gestor de la VO (VM)**: responsable de iniciar el proceso de registro.
- **El supervisor de la VO**: persona delegada del equipo de operaciones de EGI para manejar el proceso en nombre del proyecto EGI y responsable de la aprobación de las solicitudes de registro de VO.

Los principales estados en los que se puede encontrar una VO son:

- **Nuevo** : estado inicial cuando se solicita la creación de VO. Se asigna automáticamente.
- **Producción**: estado objetivo de una VO. El supervisor de VO se lo da manualmente a un VO como resultado de este procedimiento.
- **Suspendido**: este estado se ingresa cuando el VO ya no tiene información válida. Este estado puede ser temporal o previo a la cancelación del registro de VO. La intervención manual es necesaria para poner un VO en este estado.

- **Eliminado**: estado para VO que se ha terminado.

Los cambios de estado manuales solo pueden ser realizados por personas registradas en el rol de Supervisor de VO en el portal de operaciones de EGI. La autenticación de los usuario y la asignación de roles se realiza mediante el servicio Virtual Organization Membership Service (VOMS), del que hablaremos con más detalle posteriormente en este capítulo.

2.3.2. Proyecto de computación Grid del LHC

En el año 2000 se comenzó a diseñar el modelo computacional del los experimentos del LHC. Fue dentro del proyecto MONARC[24], donde se describe por primera vez la estructura jerárquica del modelo "Tiering" (Tier-0, Tier-1, Tier-2) que sigue vigente hoy en día, aunque con algunas modificaciones.

En el año 2001, el proyecto LHC Computing Grid (LCG) fue aprobado oficialmente por el consejo del CERN. El desarrollo de este proyecto conjuntamente con los diferentes experimentos, dio lugar a los Technical Design Report (TDR), para el servicio de computación del LHC, en el año 2005. Estos TDRs describen los modelos computacionales necesarios para comenzar a operar los diferentes experimentos en las primeras etapas de la toma de datos.

La puesta en marcha del LHC y la adquisición de los primeros datos reales, "**Run 1**" (2009-2012), demostró que los diseños iniciales presentes en los primeros TDR habían sido acertados, pero también mostró que debían evolucionar, después de 2 años de parada técnica, ante la proximidad del "**Run 2**" (2015-2018) y el incremento de la luminosidad del haz de colisión. Este aumento de luminosidad presentaría nuevos desafíos computacionales dentro de la comunidad Worldwide LHC Computing Grid (WLCG).

2.3.3. La colaboración Worldwide LHC Computing Grid (WLCG)

El propósito principal del WLCG es, a partir de las conclusiones obtenidas dentro del proyecto LCG, proveer de los recursos computacionales necesarios para procesar, analizar y almacenar los datos obtenidos en los diferentes experimentos.

Para poder conseguir estos objetivos, la colaboración, a través de diferentes proyectos ha desarrollado, mantenido y evolucionado una infraestructura computacional a lo largo de estos años. Debido a que el ámbito de esta tesis doctoral es la preservación de datos, nos centraremos en la parte del "middleware" relacionada más directamente con el movimiento y almacenamiento de los mismos. Aún así, para tener una visión general del modelo computacional del WLCG, describiremos de forma general los servicios y el flujo de trabajo característico en un entorno Grid.

2.3.4. Componentes Técnicos en WLCG

Como hemos comentado anteriormente, este modelo ha sufrido muchos cambios desde sus inicios, hace casi 20 años, por lo que el modelo presentando en esta tesis se centra en componentes activos durante el desarrollo de la misma.

Llegados a este punto, estableceremos el concepto de "middleware", como una capa de software que se sitúa entre un sistema operativo y las aplicaciones que se ejecutan en él. Básicamente, funciona como una capa de traducción oculta para permitir la comunicación y la administración de datos en aplicaciones distribuidas.

A pesar de las diversos cambios que el Grid ha experimentado a lo largo de los años, es posible definir la estructura básica de sus servicios (ver fig:2.12), así como sus componentes principales, como veremos a lo largo de esta sección.

2.3.4.1. UI: User Interface

La User Interface por norma general, se trata de una máquina Linux a la que se accede mediante protocolo ssh y dispone del "middleware" instalado, necesario para que un usuario pueda comunicarse con los diversos servicios dentro de la infraestructura Grid, utilidades LCG/Grid File Access Library (GFAL), envío de trabajos, acceso al almacenamiento, autenticación, etc. Es la puerta de acceso al entorno Grid.

2.3.4.2. BDII: El Berkeley Database Information Index

El Berkeley Database Information Index (BDII), es una instancia que contiene información sobre un servicio de red, y generalmente se implementa en el propio

servicio. Consiste en una base de datos OpenLDAP que se actualiza periódicamente mediante un proceso que se ejecuta en paralelo, y que obtiene información sobre el Servicio Grid de una o más fuentes de información (estáticas y dinámicas). Los servicios deben publicar los datos según el modelo de datos descrito para GLUE 1.3 y GLUE 2.0[25]. La información registrada por el recurso BDII local/site es diferente para cada tipo de "middleware".

La información que se publica se puede diferenciar en dos tipos: la estática, que depende simplemente del tipo de servicio que se está publicando (CE, SE, VOMS...), y la dinámica, que depende de los recursos variables en cada momento para cada VO publicada. El servicio BDII trabaja en dos niveles diferentes:

- **Site BDII**: Recoge toda la información publicada por los servicios dentro de un mismo sitio según los esquemas GLUE, y los presenta para ser adquiridos por el Top BDII.

- **Top BDII**: Recoge, agrega y presenta la información disponible en cada Site BDII. Consultando los Top BDII, tanto los usuarios como las herramientas Work Load Management System (WMS) usadas por las VOs, son capaces de acceder a toda la información publicada por cada sitio y planificar sus flujos de trabajo.

2.3.4.3. Servidor VOMS

El Servidor VOMS es el encargado de la autenticación de las credenciales aportadas por el usuario. Tanto los recursos como los participantes en los experimentos se agrupan dentro de las anteriormente mencionas VOs; para poder tener acceso a estos recursos es necesario tener los permisos adecuados. Una vez admitidos dentro de la VO, se nos asignarán una serie de roles dentro de la misma, y dependiendo del desarrollo de nuestro trabajo tendremos mayor o menor acceso a los recursos de la misma. Estos permisos son añadidos al proxy (seudo-cert) delegado por el servidor VOMS en forma de "roles". Este proxy tiene una duración determinada que varía entre las 12h y las 168h; durante este intervalo de tiempo podremos acceder a estos recursos, una vez terminado, deberemos volver a renovar nuestras credenciales.

2.3.4.4. WMS: Work Load management System

Es una herramienta que cada VO ha diseñado para facilitar el trabajo de sus usuarios. Es la encargada del envío de los trabajos de forma automatizada, muchas veces de forma masiva, así como de la configuración de salida de los mismos, definiendo el sitio de almacenamiento permanente y seleccionando el Tier/Institución asignada para este usuario. El WMS soluciona muchos problemas de uso de recursos informáticos distribuidos, a veces inestables, dentro de la infraestructura Grid. En particular, ayuda a la gestión de las actividades del usuario final en grandes organizaciones virtuales, como los experimentos de LHC.

Proporciona una alta eficiencia en los trabajos de los usuarios, ocultando la heterogeneidad de los recursos informáticos subyacentes a cada tecnología. Los usuarios especifican al menos un ejecutable, y/o una serie de argumentos, para ejecutarse en el nodo de cómputo. Pero los trabajos no son enviados directamente a los elementos de computo, su descripción y requisitos se almacenan en la base de datos del WMS (utilizando el formato "JDL", lenguaje de descripción de trabajo) y se agregan a una cola de tareas de trabajos con requisitos iguales o similares. Los trabajos comenzarán a ejecutarse cuando su "JDL" sea recogido por un "pilot job". Los "pilot jobs" se envían a recursos informáticos. Después del inicio, verifican el entorno de ejecución y comprueban la descripción del recurso (SO, capacidad, espacio en disco, software, etc.) La descripción de los recursos se presenta al servicio, que elige el trabajo de usuario más apropiado de la cola de tareas, prepara su entorno de ejecución y ejecuta la aplicación del usuario.

Para los usuarios, toda la maquinaria interna del WMS y los "pilot job" está completamente oculta. Ven todos los recursos informáticos operados por el WMS como un único sistema de lotes. Dentro de esta categoría tendríamos diferentes herramientas que cumplen las funciones descritas anteriormente, dependiendo de la VO que la ha implementado: Panda (ATLAS), CRAB + glideinWMS (CMS), DIRAC (LHCb) y Alien (ALICE).

2.3.4.5. Argus: Servicio de autorización

El servicio de autorización de "Argus" (ver fig:2.8) ofrece decisiones de autorización coherentes para diferentes servicios distribuidos, como interfaces de usuario o portales. El servicio se basa en el estándar XACML[26] y utiliza políticas de autorización, con ellas, se determina si un usuario tiene permiso o no para realizar una determinada acción en un servicio en particular. El servicio de autorización de "Argus" se compone de tres agentes principales:

- **Policy Administration Point (PAP)**: Tiene como finalidad, proporcionar las herramientas para crear políticas de autorización, organizarlas localmente y distribuir las políticas entre PAP remotos (regional Argus).
- **Policy Decision Point (PDP)**: Implementa el motor de autorización y es responsable de la evaluación de las solicitudes de autorización con respecto a las políticas XACML del PAP.
- **Policy Enforcement Point Server (PEP)**: Asegura la coherencia e integridad de las solicitudes de autorización recibidas de los clientes PEP. También proporciona bibliotecas de cliente PEP, facilitando la integración y la inter-operabilidad con otros servicios o componentes de European Middleware Initiative (EMI).

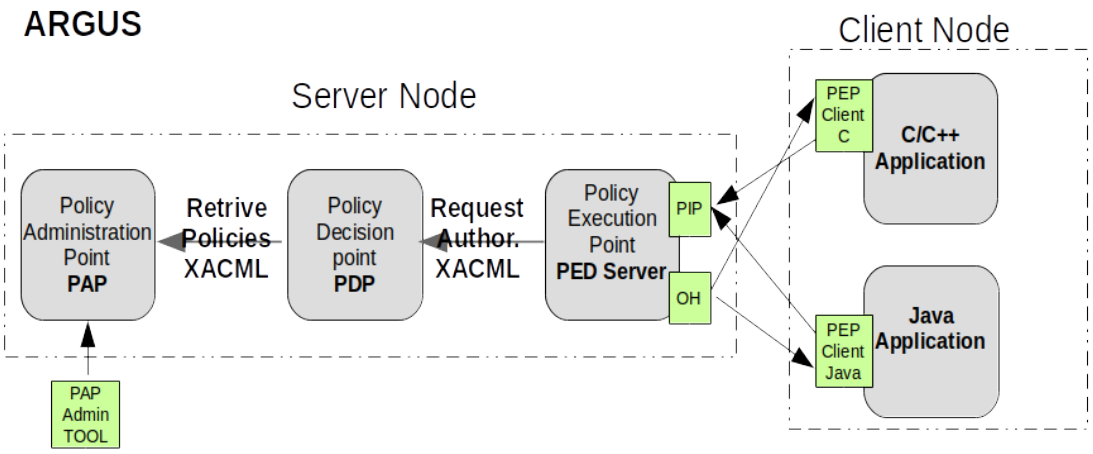


FIGURA 2.8: Descripción del servicio Argus.

2.3.4.6. CE: El Computing Element

Es el punto de acceso a los recursos de procesamiento proporcionados por cada centro de computación. Se encarga de enviar las tareas de cómputo al planificador local y de llevar una trazabilidad, proporcionando los cambios de estado a la capa Grid. El CE es el nexo de unión entre el entorno Grid y los recursos informáticos locales de cada sitio. La mayoría de los trabajos que llegan a un CE, no son enviados directamente por usuario, sino son "pilot jobs" recepcionados desde herramientas de envío (ver 2.3.4.4). Podemos encontrar diferentes implementaciones de CEs, las más comunes a día de hoy son:

- **CreamCE**: El servicio Cream es un servicio simple y liviano para la operación de gestión de trabajos a nivel de "Computing Element" (CE). El Cream acepta solicitudes de envío de tareas, que se describen con el mismo lenguaje "JDL" utilizado para describir los trabajos enviados al WMS y otras solicitudes de administración, como por ejemplo la cancelación o supervisión. CREAM puede ser utilizado por el WMS, mediante el servicio ICE, o a través un cliente "cli", donde un usuario final puede enviar trabajos directamente a un Cream CE (gLite tools) . Los desarrolladores del "middleware" del CreamCE anunciaron un End of Life (EoL) situado a principios del año 2020 por lo que los centros Grid con instancias de este "middleware" en producción, deben planificar la sustitución de este servicio.
- **HTCondorCE**: Es una configuración especial del software HTCondor[27] diseñada para ser una solución de pasarela de trabajo. El componente principal es el "Job Gateway", que es responsable de gestionar los trabajos entrantes, autorizarlos y delegarlos al sistema de colas local, para su ejecución. Sus principales características son:
  - Escalabilidad: HTCondor-CE es capaz de soportar cargas de trabajo de sitios grandes.
  - Herramientas de depuración: HTCondor-CE ofrece muchas herramientas para ayudar a solucionar problemas con trabajos.
  - Encaminamiento como configuración: el mecanismo de HTCondor-CE para transformar y enviar trabajos se personaliza a través de variables

de configuración, lo que significa que las personalizaciones persistirán en las actualizaciones y no implicarán la modificación de los componentes internos del software para encaminar las tareas.

- **ArcCE**: El ARC Compute Element (CE) es un front-end Grid convencional. Es capaz de realizar muchas acciones diferentes, como publicar información sobre sí mismo, realizar actividades de autorización de usuario basadas en credenciales de Grid y políticas locales, mapeos de usuarios Grid a cuentas locales, aceptar trabajos Grid de usuarios autorizados en diferentes lenguajes, JSDL o ARC xRSL, descargar archivos de entrada adicionales de los almacenes de Grid en nombre del usuario autorizado, conversión de descripciones de trabajo Grid a scripts para envío de tareas al sistema de colas local. Soporta gran variedad de sistemas de colas, Condor, PBS, Slurm, SGE, etc. Prepara el entorno de ejecución de la aplicación pre-instalada, si se especifica en la descripción del trabajo, a través de Run Time Environments (RTE). En su interacción con el sistema de colas:
  - Proporciona salidas de trabajo a solicitud del usuario.
  - Proporciona información del estado del trabajo.
  - Puede interrumpir trabajos a solicitud del usuario.
  - Puede reiniciar trabajos fallidos a solicitud del usuario.
  - Limpia las sesiones de usuario.

Como ventaja con respecto a las opciones anteriores, es capaz de recopilar información contable (registro de uso) sin usar un software externo.

2.3.5. SE: El Storage Element

Se encarga de proporcionar recursos de almacenamiento al entorno Grid. Gestiona el espacio de almacenamiento de las VOs y proporciona accesibilidad a los datos mediante diferentes protocolos, Xroot, file, Gsiftp, de forma que los trabajos enviados sean capaces de tener éstos datos como entrada de sus ejecuciones. Si bien es verdad que pequeños datos podrían adjuntarse al "JDL" anteriormente mencionado, en entornos tan exigentes es posible que los datos de

entrada sean de ordenes de magnitud cercanos al Gigabyte, haciendo inviable el envío de trabajos con datos adjuntos. La segunda de las funcionalidades es la de ser receptores de la salida de los trabajos de los usuarios, ya sea de forma temporal o definitiva. A continuación veremos algunas de las diferentes implementaciones de Storage Element (SE)'s más comunes en los entornos WLCG.

2.3.5.1. dCache

dCache (ver fig:2.9) es una solución de almacenamiento distribuido. Organiza el almacenamiento agregado de todos los equipos, para que el almacenamiento combinado pueda ser usado, de modo que los usuarios finales no conocen el nodo físico donde se almacenan sus datos. Simplemente ven una gran cantidad de espacio de almacenamiento. Debido a que los usuarios finales no necesitan saber en qué equipos están almacenados sus datos, se pueden migrar de una computadora a otra sin ninguna interrupción del servicio, también se pueden realizar operaciones en caliente, como agregar o quitar servidores del clúster de almacenamiento dCache en cualquier momento. dCache admite la solicitud de datos de un sistema de almacenamiento terciario. Normalmente estos sistemas se refieren al empleo de cintas magnéticas en lugar de discos, que deben cargarse y descargarse utilizando un brazo robótico. La razón principal para usar el almacenamiento terciario es la mejor rentabilidad medida en €/GB, archivando una gran cantidad de datos en hardware bastante económico. A su vez, la latencia de acceso para los datos archivados es significativamente mayor, ya que el dato se encuentra "nearline". Soporta diferentes protocolos, permitiendo a los usuarios el acceso a los datos. Se implementan de forma modular, lo que permite que dCache soporte una mayor tasa de peticiones añadiendo máquinas de FrontEnd (FE) adicionales.

Una de sus características más notables es la migración de datos calientes. dCache detectará cuándo se solicitan archivos de forma reiterada, es decir con mucha frecuencia. dCache generará duplicados de estos archivos populares distribuidos en su batería de nodos, distribuyendo la carga entre varias máquinas, aumentando así el rendimiento. El flujo de datos dentro de dCache se controla cuidadosamente. Los datos entrantes y salientes son agrupados para que utilicen recursos designados, garantizando un mejor rendimiento y mejorando la experiencia del usuario

final[28]. Su principal desventaja es la complejidad de implementación frente a otras soluciones.

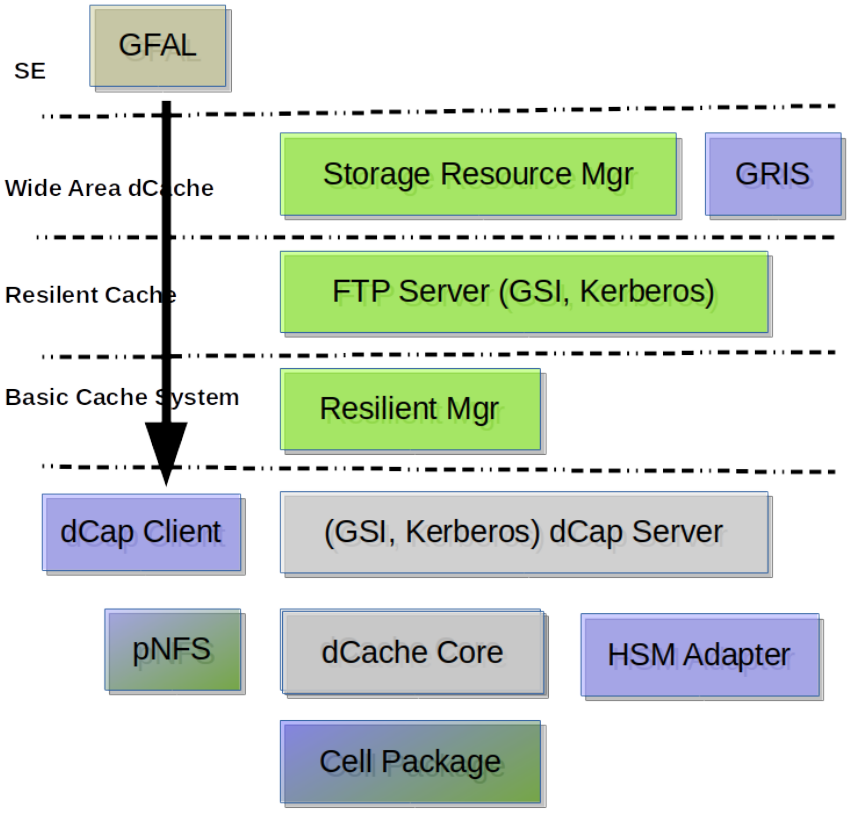


FIGURA 2.9: Descripción del servicio dCache.

2.3.5.2. Storage Resource Manager (StoRM)

StoRM es un servicio de administración de almacenamiento (SRM) ligero, escalable, flexible, de alto rendimiento, independiente del sistema de archivos. Es un sistema de almacenamiento genérico, basado en disco, que cumple con la interfaz estándar de SRM. Actualmente la versión 2.2. StoRM proporciona capacidades de gestión de datos en un entorno Grid, para compartir, acceder y transferir datos entre centros heterogéneos y distribuidos geográficamente. Funciona contra casi cualquier BackEnd de almacenamiento POSIX (ext3, ext4, xfs, btrfs, etc), básicamente cualquier sistema de ficheros soportado en las distribuciones de Linux.

Pero también es capaz de emplear las ventajas de los sistemas de almacenamiento de alto rendimiento, basados en sistemas de archivos Clúster, como "Spectrum Scale" de IBM o Lustre de Sun Microsystems, que admiten llamadas I/O POSIX nativas al sistema de archivos y directorios compartidos.

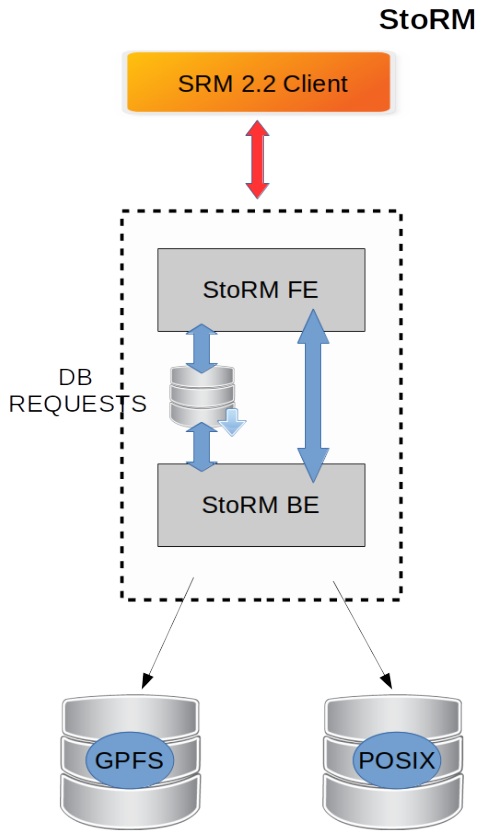


FIGURA 2.10: Descripción del servicio StoRM.

StoRM (ver fig:2.10) es capaz de identificar la ubicación física de los datos solicitados sin consultar ningún servicio de base de datos, usando un esquema XML que describe el espacio de nombres de almacenamiento, así como los parámetros de entrada; el identificador lógico y los atributos SRM. StoRM se basa en la estructura del sistema de archivos subyacente para identificar la posición de los datos físicos. Tres son sus componentes principales, denominados FrontEnd (FE), BackEnd (BE), y una base de datos utilizada para almacenar solicitudes SRM y todos los metadatos StoRM[29].

2.3.5.3. Disk Pool Manager (DPM)

Disk Pool Manager (DPM), ofrece una forma sencilla de crear un elemento de almacenamiento basado en disco compuesto por múltiples servidores de disco. DPM (ver fig:2.11) admite una gran cantidad de los protocolos de acceso a datos y metadatos, como HTTP, Xroot, SRM, gridFTP y RFIO. En el caso de protocolos flexibles como HTTP y Xroot, se enfoca en brindar las características avanzadas que mejoran el rendimiento de las aplicaciones de análisis. Su principal ventaja es la facilidad de instalación y configuración, con un bajo esfuerzo de mantenimiento, sin dejar de proporcionar todas las funcionalidades requeridas para una solución de almacenamiento en red[30].

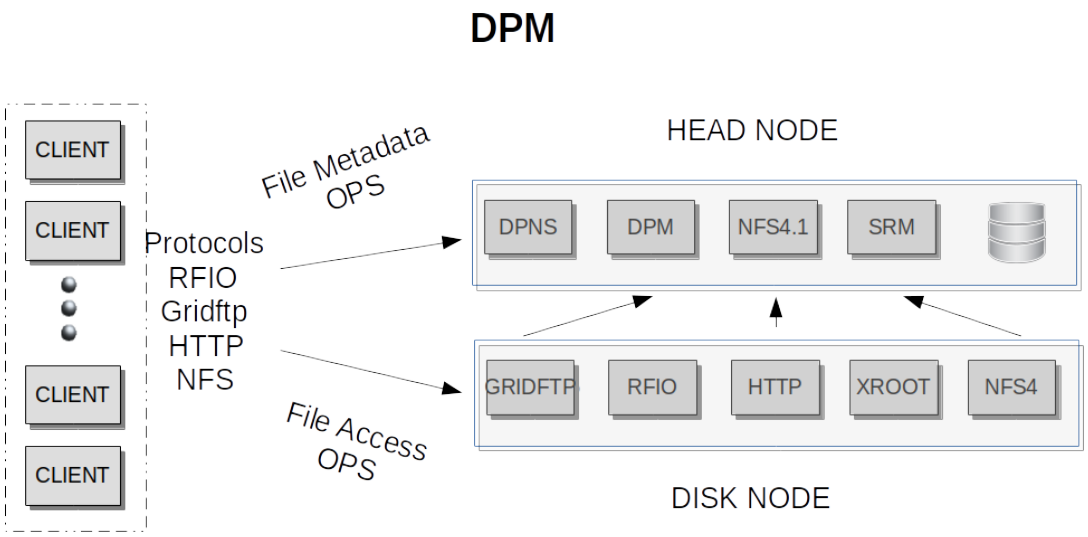


FIGURA 2.11: Descripción del servicio DPM.

Una vez descritos los elementos más comunes, en la figura 2.12 se muestra la conceptualización de la conectividad de los diferentes servicios para un flujo de trabajo Grid.

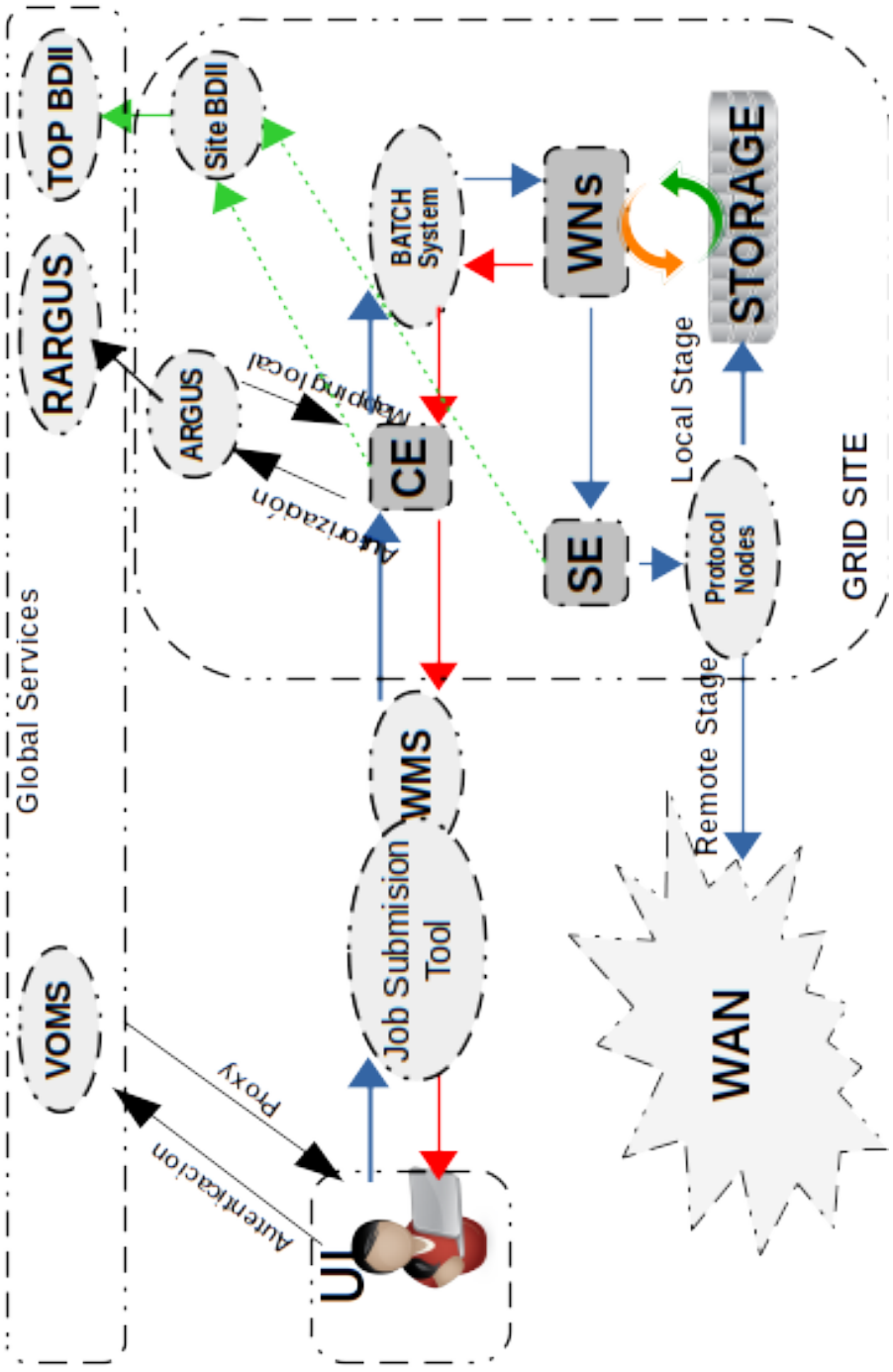


FIGURA 2.12: Conectividad de servicios en un entorno Grid año 2019.

2.3.6. UMD: El Middleware Grid

La Unified Middleware Distribution (UMD) es el conjunto integrado de componentes de software aportados por los proveedores de tecnología y empaquetados para su implementación como servicios de producción en EGI[31]. UMD continúa la labor iniciada por los proyectos gLite 3.1 y 3.2, operativos hasta el 1 de Mayo del 2013.

UMD4 es la distribución UMD actual, año 2019, compatible con la distribución de Linux CentOS7, SL6 y Ubuntu. UMD3 seguirá recibiendo actualizaciones críticas o de seguridad para SL6, mientras que UMD3/SL5 y UMD3/Debian actualmente no son compatibles.

Durante un largo periodo de tiempo, la herramienta empleada para realizar la instalación y configuración de estos componentes ha sido Yaim[32]. Consiste en un conjunto de scripts y funciones que, mediante la definición de una series de variables generales dependientes del sitio Grid y otras particulares propias del servicio, es capaz de instalar y configurar, los diferentes "middlewares" Grid, empleando los repositorios de UMD. Actualmente esta instalación y/o configuración emplea módulos de puppet[33] o ansible[34] para realizar esta tarea[35] sobre los diferentes "midlleware" Grid, auque algunos como StoRM todavía emplean Yaim.

2.3.7. EGI

La Infraestructura Grid Europea (EGI) es una infraestructura computacional federada creada para proporcionar servicios informáticos avanzados para investigación e innovación. La infraestructura EGI está financiada con fondos públicos y comprende cientos de centros de datos y proveedores de servicios distribuidos por todo el mundo. Ofrece una amplia gama de servicios para computación, almacenamiento, datos y soporte, con acceso a más de 1.000.000 núcleos y 740PB en disco y cinta (ver fig:2.13), que son proporcionados por los proveedores Cloud y centros de datos federados de EGI. Estos servicios son accesibles hoy en día a través del EGI Marketplace[36].

La Fundación EGI, también conocida como EGI.eu, coordina la infraestructura en nombre de los participantes del Consejo de EGI: infraestructuras computacionales nacionales y organizaciones europeas de investigación inter-gubernamental

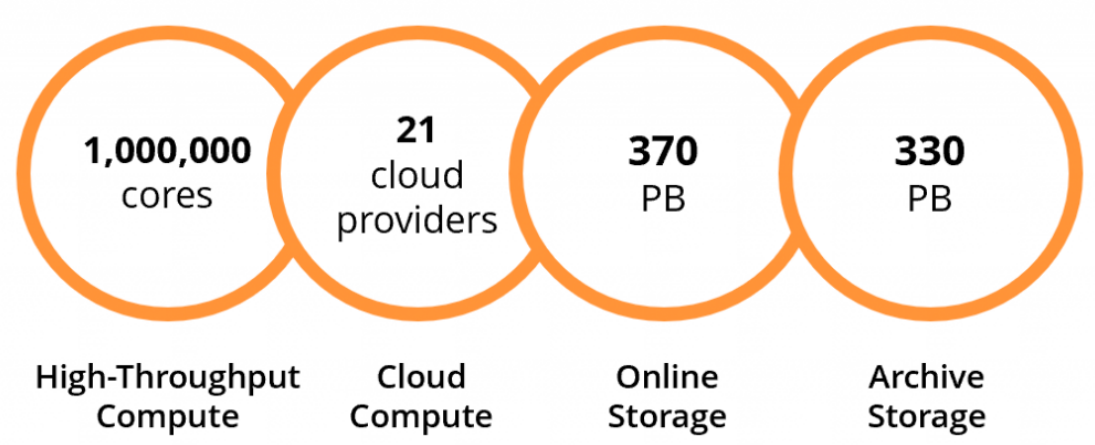


FIGURA 2.13: Recursos accesibles a través del marketplace de EGI a Noviembre del 2019.
[37]

(EIRO). La Fundación EGI no tiene ánimo de lucro y se estableció en Amsterdam en 2010 bajo la ley holandesa. Está gobernada por el Consejo EGI. Las operaciones diarias son supervisadas por la Junta Ejecutiva. Los siete miembros de la Junta Ejecutiva son nombrados por el Consejo EGI por períodos de dos años, y trabajan estrechamente con los directores administrativos y técnicos en temas operativos, técnicos y financieros.

2.3.7.1. EGI Fedcloud

Durante los últimos años, las infraestructuras Grid han comenzado a evolucionar hacia entornos mucho más versátiles y adaptables a las nuevas necesidades computacionales, solicitadas por los nuevos proyectos. Los entornos Cloud son la solución perfecta para cubrir esta nuevas necesidades. La EGI Fedcloud participa en una serie de proyectos financiados con fondos europeos, con muy diversas características y finalidades. Siendo algunos de ellos son:

- **EOSC-HUB**: European Open Science Cloud (EOSC)-HUB, reúne a un amplio grupo de proveedores de recursos computacionales, nacionales e internacionales para crear el "HUB". El HUB consiste en un punto de contacto centralizado, donde los investigadores e innovadores europeos pueden descubrir, acceder, usar y reutilizar un amplio espectro de recursos, dedicados

a la investigación basada en datos. EOSC-HUB reducirá significativamente la fragmentación de las instalaciones de IT en Europa. El Sistema de Integración y Gestión de Servicios (SIAM) es el responsable de garantizar, que todos los proveedores de servicios EOSC proporcionen un servicio continuo, cumpliendo con las obligaciones contractuales contraídas con las organizaciones de clientes.

- **EOSC-Synergy**: contribuirá a EOSC mediante la expansión de las capacidades de recursos y el desarrollo de capacidades humanas para usar EOSC. En la práctica, esto significará más recursos de cómputo y almacenamiento disponible para los investigadores, y más conjuntos de datos y herramientas para ampliar las vías de investigación en las áreas como, Ciencias de la Vida, Ciencias Ambientales y Astrofísica. La Observación de la Tierra está representada por centros de investigación de Portugal, Francia y España, y por un socio industrial líder, INDRA. El proyecto actuará como puente de EOSC hacia América del Sur, estableciendo un vínculo con los experimentos brasileños Cloud y Astrofísica. EGI contribuirá en tareas de comunicación y difusión, gestión de la innovación y actividades de desarrollo de políticas.

Para desarrollar con éxito los proyectos anteriormente mencionados, la EGI Fedcloud, proporciona servicios, cada uno de ellos con diferentes características, para adaptarse a la necesidades.

2.3.7.2. Computación Cloud

Brinda la capacidad de implementar y escalar máquinas virtuales bajo demanda. Ofrece recursos computacionales garantizados en un entorno seguro y aislado con acceso Interfaz de Programación de Aplicaciones (API) estándar, sin la sobrecarga de administrar servidores físicos[38].

La computación Cloud ofrece la posibilidad de seleccionar dispositivos virtuales pre-configurados (por ejemplo, CPU, memoria, disco, sistema operativo o software) de un catálogo replicado en todos los proveedores de nube de EGI. La computación Cloud proporciona la posibilidad de:

- Ejecutar cargas de trabajo intensivas en cómputo y datos (tanto por lotes como interactivas).

- Hospedar servicios de larga duración, como servidores web, bases de datos o servidores de aplicaciones.
- Crear entornos de prueba y desarrollo desechables, en máquinas virtuales.
- Escalar necesidades infraestructurales bajo demanda. Seleccionar configuraciones de máquinas virtuales (CPUs, RAM, disco), y entornos de aplicación, para satisfacer sus necesidades.
- Administrar recursos de Cloud de computo de manera flexible, con capacidades integradas de monitorización y contabilidad.

Dentro de las diferentes implementaciones de Infraestructura como Servicio (IAAS) en entornos Cloud de código abierto, posiblemente la más extendidas sea OpenStack. Mediante una serie de servicios, OpenStack es capaz de dotar de gran cantidad de propiedades a los nodos presentes en la infraestructura de forma intuitiva, a través de una interfaz de gestión web.

Al igual que hemos descrito una serie de servicios Grid, podemos definir también los servicios básicos de una infraestructura cloud (ver fig:2.14):

- **Keystone**: Autentica y autoriza todos los servicios de OpenStack.
- **Neutron**: Crea y gestiona las capas de red de las IAAS "instanciadas" sobre OpenStack.
- **Nova**: Es una herramienta de gestión integral y acceso para los recursos computacionales en OpenStack. Gestiona la planificación, la creación y la eliminación de las máquinas virtuales sobre la infraestructura. También es capaz de interactuar con los diferentes servicios de OpenStack.
- **Horizon**: Mediante una interfaz gráfica, nos muestra todas las opciones de OpenStack. Desde esta interfaz, podemos acceder y gestionar los diferentes servicios sobre los que estemos autorizados, y que estén soportados por nuestra infraestructura Cloud. No es la única implementación de este servicio, pero si la más usada.

- **Cinder**: Este Servicio se centra en la gestión del almacenamiento tradicional. Nos presentará accesos a dispositivos de disco (bloque) que estén accesibles desde nuestro Cloud mediante diferentes BanckEnds como LVM, Ceph o Spectrum Scale por ejemplo.

- **Glance**: Es el servicio de gestión de imágenes de originales diferentes sistemas operativos, o imágenes con softwares pre-instalados. También almacena y recupera imágenes del disco de la máquina virtual ("snapshots").

- **Swift**: Es un servicio de almacenamiento de objetos que almacena y recupera objetos de datos no estructurados utilizando una API RESTful. Asegura la integridad y replica los objetos según el ratio definido, 1:3 usualmente, por los diferentes elementos de almacenamiento definidos en la infraestructura.

- **Manila**: Proporciona la administración de archivos compartidos, NFS y CIFS como un servicio central para OpenStack. Soporta varios BackEnd de almacenamiento (LVM, Spectrum Scale, NetApp, etc). Proporciona el acceso a sistemas de ficheros desde las máquina ejecutadas sobre la infraestructura Cloud.

2.3.7.3. Contenedores Cloud

Brinda la capacidad de implementar y escalar contenedores bajo demanda. Proporciona recursos computacionales garantizados en un entorno seguro y aislado, con acceso API estándar, sin la sobrecarga de administrar los sistemas operativos. Sus principales ventajas son:

- Aprovisionamiento bajo demanda.

- Ambiente ligero para un rendimiento maximizado.

- Interfaz estándar para implementar en múltiples proveedores de servicios.

- Interoperable y transparente.

- Elimina la fricción entre los entornos de desarrollo y operaciones.

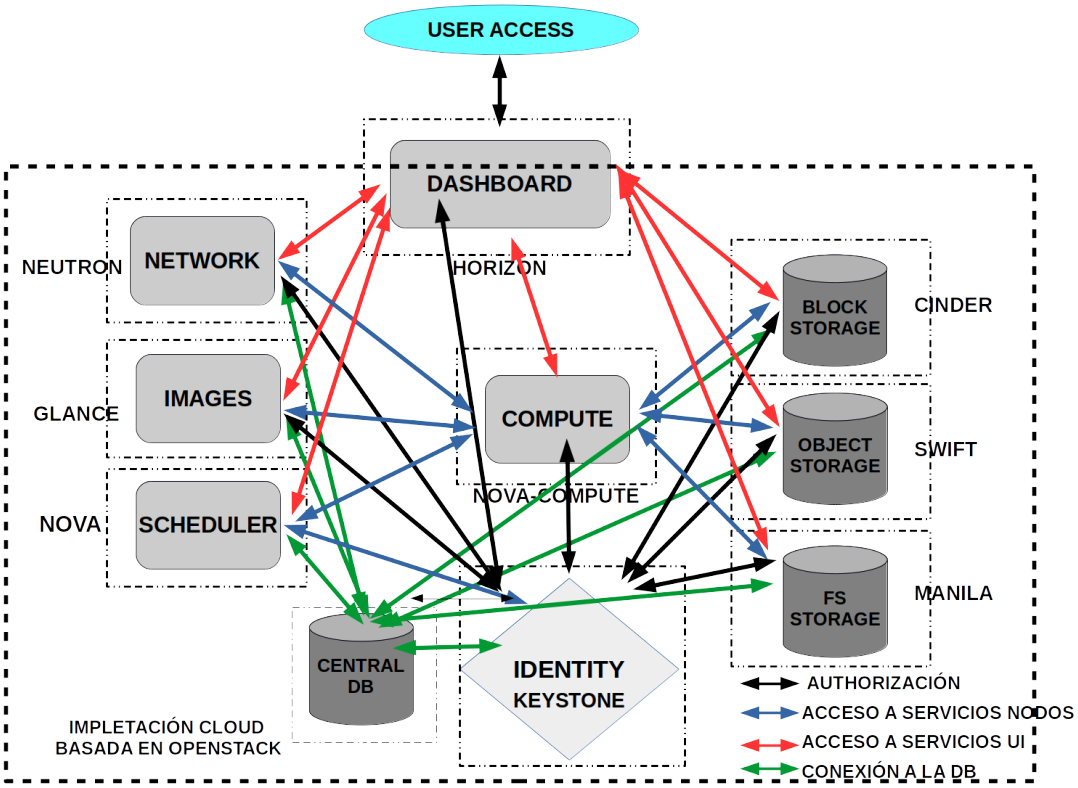


FIGURA 2.14: Despliegue de servicios para una infraestructura Cloud basada en la suit OpenStack.
[37]

2.3.7.4. Infraestructuras de Formación en entornos Cloud

Mediante este tipo de entornos, podemos proporciona recursos de computación y almacenamiento basado en Cloud para eventos de formación, entrenamiento y/o prácticas. Es útil en la organización de tutoriales, talleres, cursos de formación en línea, o el despliegue de plataformas de aprendizaje.

Tiene la capacidad de implementar imágenes de máquinas virtuales personalizadas según las necesidades específicas de cada comunidad, beneficiándose de la fácil implementación y reutilización de los materiales del curso en la infraestructura, como por ejemplo en el desarrollo de entornos de prácticas para estudiantes.

EGI proporciona sus entornos de formación bajo infraestructura computacional y de almacenamiento de alta calidad. Algunas de sus aplicaciones son:

- Cursos específicos de objetivos y valor agregado para comunidades científicas.
- Acceso a demanda fácil de usar y mejoras en la oferta de formación.
- Permite una fácil implementación de cursos y su reutilización.

2.4. El entorno Grid del IFCA

El Instituto de Física de Cantabria (IFCA) lleva varias décadas posicionado en España como referente dentro de las tecnologías de computación distribuida. Su primer clúster se instaló a mediados de la década de los 90, con herramientas software por entonces muy rudimentarias y sistemas operativos bastante básicos (Redhat5/6) y complejos de mantener e instalar. El IFCA participó de forma activa en los primeros proyectos que implementaban este tipo de tecnologías como, CrossGrid[39] y DataGrid[40]. Fue miembro de la organización EGEE, precursora de la anteriormente mencionada EGI (ver 2.3.7), aportando su infraestructura.

Desde su inicio, los científicos del IFCA formaron parte del experimento CMS, con un grupo de trabajo especializado en las secciones de alineamiento y en el grupo de computación orientada a objetos. En el año 2005, comenzó el proyecto del plan nacional[41] "**Centro de procesado de datos de colisiones del LHC: Tier 2 federado para el experimento CMS**". Esto proporcionó al IFCA los recursos necesarios para establecer de forma estable un centro de procesado de datos, basado en tecnologías Grid, que sería el germen de la actual infraestructura computacional y del grupo de Computación Avanzada del IFCA.

La implementación de la infraestructura Grid en el IFCA está basada en algunos de los servicios explicados en el punto 2.3.4. Como hemos visto con anterioridad, dentro de las tecnologías Grid, existen diferentes "middleware" que ofrecen similares funcionalidades a la hora de establecer un servicio. La motivación del IFCA para seleccionar una solución entre las diferente opciones, se debe a la naturaleza multi-disciplinar de los experimentos y/o proyectos en los que participa. El IFCA tiene que dar servicio a múltiples VOs, con ello, la gestión de la infraestructura se vuelve más compleja. Hay que adaptar la infraestructura para que cada proyecto pueda alcanzar sus objetivos, y son tecnologías Cloud (ver

2.3.7.1) las que pueden gestionar estos espacios que las tecnologías Grid no son capaces de cubrir.

A continuación procederemos a explicar de la forma más concisa posible el entorno de trabajo Grid, implementado por el IFCA, así como las motivaciones que han llevado a estas elecciones.

Dentro de este capítulo, estudiaremos aquellos servicios dentro de la infraestructura donde el IFCA ha tenido capacidad de elección, centrándonos en dos servicios fundamentales para el desarrollo de esta tesis: el Computing Element (CE) (ver 2.3.4.6) y el SE (ver 2.3.5).

2.4.1. Implementación del Computing Element (CE) en el IFCA

El Servicio de CE podemos dividirlo en tres partes claramente diferenciadas, la primera de ellas, el propio "middleware" de CE, del que ya hemos hablado con anterioridad, la segunda es el sistema de colas local, y la tercera, los nodos de cómputo o "workernodes", como se conocen dentro del argot computacional.

2.4.1.1. El Middleware CE

El Centro de Procesado de Datos (CPD) del IFCA, ha empleado tres tipos de "middleware" significativos a los largo de dos décadas. Inicialmente el lcg-ce, más tarde el CreamCE y durante el desarrollo de esta tesis, se está realizando la migración a ArcCE, ante el inminente EoL del CreamCE, previsto para finales del año 2019. Nos centraremos en estos dos últimos como pasado, presente y futuro.

- **CreamCE**: Desarrollado y mantenido por el Instituto Nazionale di Fisica Nucleare (INFN), ha estado operativo durante aproximadamente los últimos 12 años instalado en producción en la mayoría de los sitios Grid de mundo. Uno de sus problemas principales es que sus agentes están programados en java, esto provoca que el servicio consuma una gran cantidad de recursos, teniendo que ser monitorizado de forma regular para evitar un mal funcionamiento de los agentes. La infraestructura de este servicio en el IFCA consta de tres máquinas virtuales con 10GB de RAM y 6 CPUs, dos de

ellas dedicadas en exclusiva a la VO de CMS y la otra para el resto de VOs soportadas. El Cream, es capaz de comunicarse con varios sistemas de colas locales, entre ellos Torque/PBS, Maui y SGE.

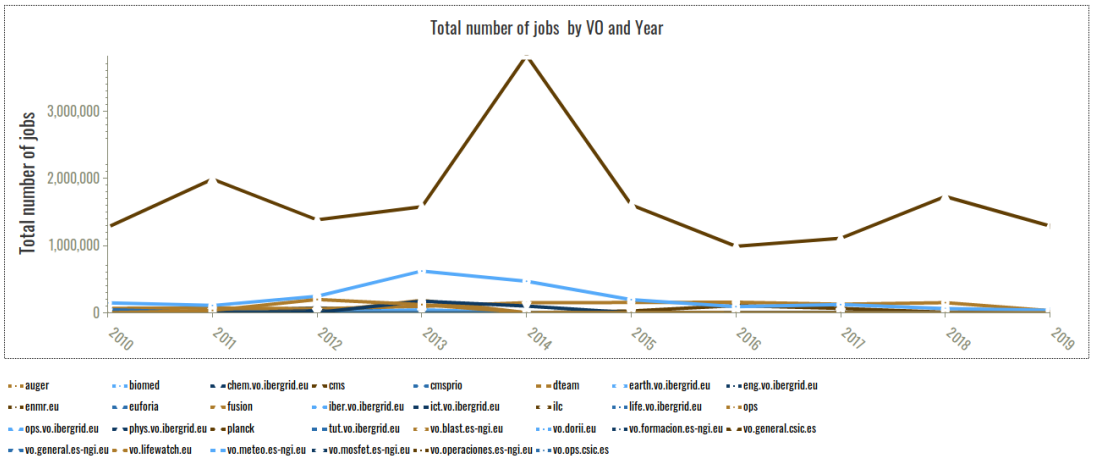


FIGURA 2.15: Trabajos por VO desde 2010 usando CreamCE. [42]

En líneas generales es un servicio que ha dado un buen resultado, pero la ejecución de servicios java, como "Tomcat" o "Blab", provoca que sea un servicio pesado durante ciclos de trabajo elevados. Esta fue una de la principales razones de la multiplicidad de este servicio en tres máquinas virtuales divididas por VO. Como se observa en la figura 2.15 la VO CMS es la más intensiva pudiendo ejecutar varios miles de trabajos de forma consecutiva durante largos periodos de tiempo. A comienzo del 2019, se anuncia su EoL para files del mismo año, el INFN dejará de dar soporte y mantener (parches de seguridad, actualizaciones,etc) el citado software, por lo que se insta a los diversos centros dentro de la comunidad Grid, con este "middleware" en producción, a migrar a otro con soporte para los próximos años.

- **ArcCE**: Después del anuncio del fin del soporte para el "middleware" de CreamCE, el IFCA, sopesando las posibilidades decidió que el ArcCE (ver fig:2.16) era la mejor opciòn. Está diseñado específicamente para implementarse en un sitio que soporte múltiples VOs. La otra posibilidad, HTCon-

dorCE, también permite este tipo de configuración multi VO pero ArcCE es más versátil, pudiendo implementar varios sistemas de colas diferentes, Son of Grid Engine (SGE), Slurm, Condor, mientras que HTcondorCE se focaliza en en uso de HTCondor como sistema local de colas.

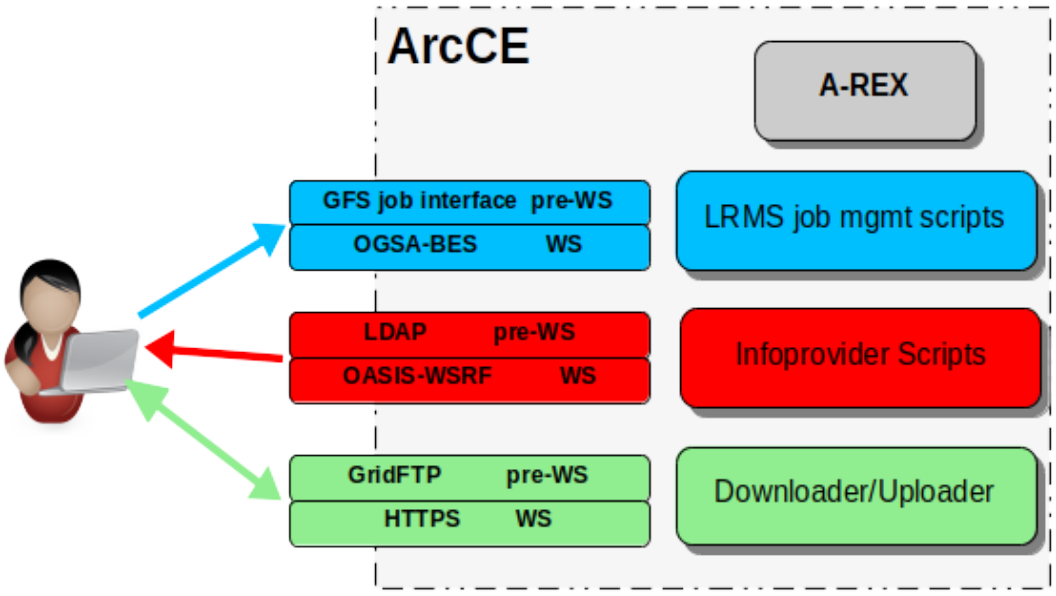


FIGURA 2.16: Descripción del servicio ArcCE.

Está desarrollado por Nordugrid[43], y tiene la ventaja con respecto a su antecesor (CreamCE) de ser capaz, mediante un agente interno al propio "middleware", de realizar tareas de contabilidad y publicación, mientra que el CreamCE necesitaba de un agente externo para este propósito, simplificando de esta forma la infraestructura y la gestión de la misma. Una sola instancia de ArcCE, con las prestaciones de los equipos actualmente en producción (10GB RAM 6 CPUs), es capaz de soportar la ejecuciòn de 10.000 trabajos de forma simultanea. Los agentes de ArcCe (ver fig:2.17) están programados en C++, son mucho más ligeros que los de Cream, por lo que a priori, su estabilidad es mayor, ante un numero de trabajos similar. Instalando dos instancias de ArcCE, nos aseguramos la redundancias del servicio, ante la pérdida de una de ellas, siendo suficiente para el desarrollo normal de la ac-

tividad computacional del IFCA, reduciendo en un equipo la infraestructura actual de tres CreamCEs.

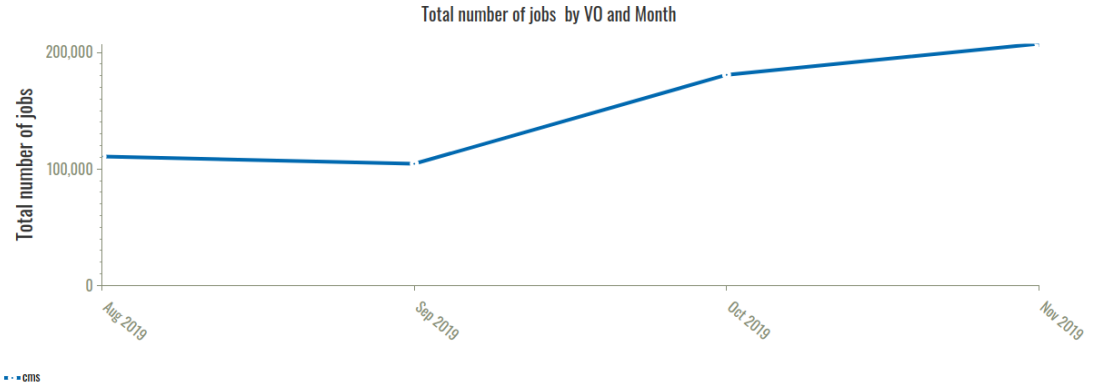


FIGURA 2.17: Trabajos por VO desde Agosto 2019 usando ArcCE. [42]

2.4.1.2. Sistemas de Colas

Durante el desarrollo de ésta tesis, dos sistema de colas ha sido empleados por el IFCA. El primero de ellos SGE, gestionando los equipos de cómputo Grid, y el segundo Slurm, en el equipamiento computación de alto rendimiento (HPC).

- **SGE**: Ha sido el software de gestión de colas local empleado en computación Grid y uso interactivo por IFCA, durante los últimos 10 años (ver fig:2.18). Desarrollado y mantenido por Sun Microsystems, hasta que fue adquirida por Oracle en el año 2010. "Son of Grid Engine" es una continuación comunitaria del antiguo proyecto Sun Grid Engine. Es una bifurcación o "fork" de la distribución de Univa, ante la discontinuidad de Oracle. Tiene como objetivo integrar parches y utilidades existentes. La última actualización de software publicada fue la 8.2 en Marzo del 2016. Este software se ha comportado extremadamente bien, durante el período en el que ha estado en producción, pero su falta de mantenimiento y actualizaciones es un problema. En el año 2019, aprovechando la necesidad del cambio de "middleware" en el CE (CreamCe EoL), y que se ejecutaba de forma satisfactoria otro gestor de colas en el IFCA, que cumplía con nuestras expectativas, nos inclinamos definitivamente por su sustitución.

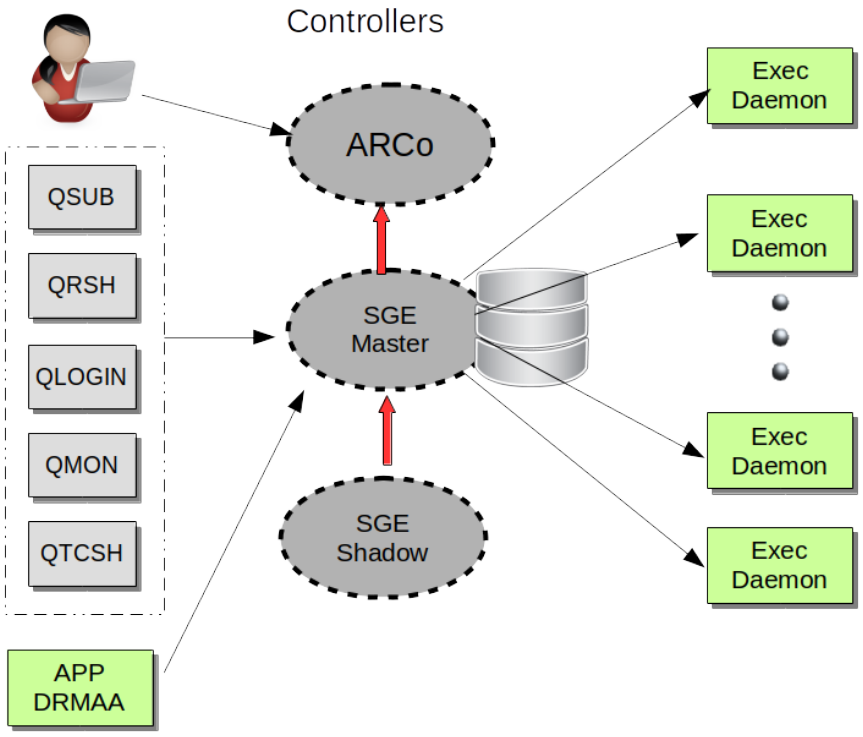


FIGURA 2.18: Descripción del servicio SGE.

- **Slurm**: La elección de Slurm (ver fig:2.19) como el nuevo sistema de colas para computación Grid, fue simple; primero, el sistema HPC del IFCA, con más de 2500 cores, como el resto de nodos de la Red Española de Supercomputación (RES) emplean Slurm; segundo, es totalmente funcional con el nuevo "middleware" ArcCE; tercero, dispone del sistema de actualizaciones de las que SGE carece; y cuarto, el IFCA tiene una amplia experiencia en su implementación. Desde hace tiempo, las VOs demandantes de gran cantidad de recursos, como CMS o ATLAS, han estado buscando la forma de usar de forma oportunística[44] este tipo de recursos HPC. La Integración en Slurm en ambos sistemas posibilita al IFCA el acceso a todos los dispositivos de forma transparente para el usuario final de la VO, accediendo a los recursos HPC a través del "middleware" CE. El uso de éstos recursos será cuantificable dentro del portal de contabilidad Grid, situación que hasta el momento no era posible.

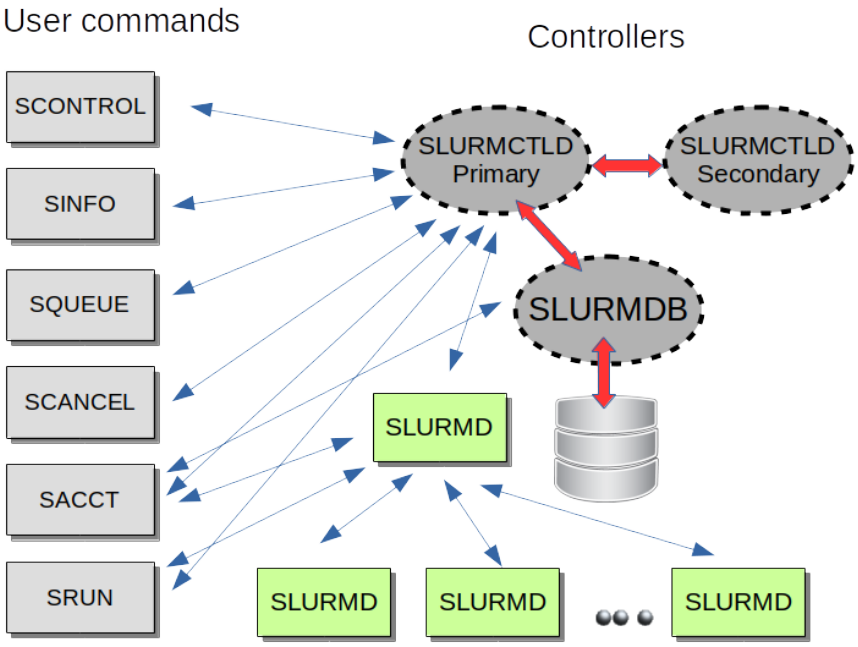


FIGURA 2.19: Descripción del servicio Slurm.

2.4.1.3. Nodos de Cómputo

Llamamos nodo de cómputo o "workernode", al nodo final donde se va a producir la ejecución del trabajo, es decir el nodo que va a utilizar parte de los recursos publicados para cada VO (CPU, RAM, disco, etc). Este tipo de nodos, al igual que el resto de servicios Grid instalan un "middleware" y las variables de entorno necesarias para su correcto funcionamiento. Es el software contextualizado, utilizado por cada una de las VO del LHC, que se instala de manera centralizada mediante un sistema de ficheros caché, accesible mediante dos servidores proxy locales en modo lectura llamado CVMFS[45], evitando accesos continuados al sistema central remoto.

Los nodos de cómputo Grid del IFCA, son gestionados como máquinas virtuales por la suit Cloud, OpenStack. OpenStack es muy versátil, es capaz de

gestionar diferentes proyectos y aplicar sobre ellos diferentes propiedades. Entre sus característica principales, es capaz de añadir o eliminar nodos, distribuir y gestionar imágenes (iniciar, borrar, reiniciar, clonar) sobre diferentes hypervisores como XEN o KVM[46], aplicar reglas del cortafuegos, adjuntar volúmenes, etc.

El uso de OpenStack como gestor de los nodos de cómputo, proporciona una gran versatilidad a la hora de realizar cierto tipo de operaciones rutinarias como, aplicar parches de seguridad, instalación de nuevo software, o simplemente la realización de pruebas de configuración. OpenStack gestiona la red de los diferentes proyectos, creando vlans independientes para cada uno de ellos, minimizando las posibilidades de problemas de seguridad. En definitiva, el IFCA proporciona una IAAS mediante la suit OpenStack al proyecto CMS (ver 2.3.7.2).

La figura 2.20, muestra como el IFCA ha implementado el servicio CE, a partir de los recursos presentados anteriormente.

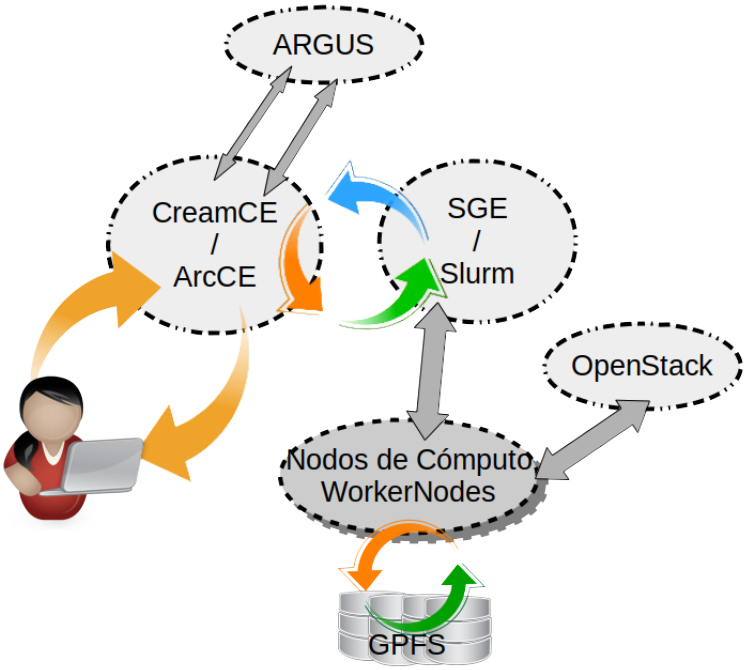


FIGURA 2.20: Conceptualización de la implementación del servicio CE en el IFCA.

2.4.2. Implementación de Storage Element (SE) en el IFCA

La elección llevada a cabo por el IFCA, para los servicios del "Storage Element", tienen gran relevancia en el desarrollo de este trabajo, pues definen la base de estructura sobre la cual se ha implementado el desarrollo técnico de esta tesis. Como vimos en el punto 2.3.5, podemos establecer tres componentes principales para el servicio de SE, el "middleware", el sistema de ficheros a utilizar y los protocolos de acceso a datos soportados.

2.4.2.1. El middleware de Storage Element

Durante la primera década del siglo XX, el almacenamiento del IFCA tenía un tamaño aproximado de 100TB. Este era un volumen considerable en el año 2005, pero todavía asequible mediante la agregación de servidores de disco. Durante el periodo 2005-2008, la elección de Disk Pool Manager (DPM) como SE fue apropiada para unos recursos humanos y materiales limitados. A partir de las primeras colisiones en el LHC y la llegada de datos reales, la aportación de recursos computacionales del IFCA al entorno Grid se incrementó de forma exponencial; con la instalación del Clúster Grid-CSIC y la adquisición de cabina de almacenamiento 1PB por parte del experimento CMS. La mejora de los recursos, unido a problemas en las librerías de acceso a ficheros (rfio) en la versión de DPM en producción, sumado a la entrada del IFCA en la RES, que empleaba General Parallel FileSystem (GPFS) como sistema de ficheros, acotaron la elección del nuevo "middleware" de SE para el IFCA.

La elección de StoRM como "middleware" de SE fue acertada. El resto de las soluciones disponibles, en aquel momento, a parte del ya probado DPM, eran dCache y StoRM. dCache proporcionaba una gran cantidad de funcionalidades, pero su gestión e implementación era muy compleja, en contrapartida, la instalación y gestión de StoRM[47] era relativamente sencilla (ver 2.3.5). Necesita un sistema POSIX como BackEnd (BE) de almacenamiento. Además todas la funcionalidades especiales proporcionadas por dCache eran también proporcionadas por GPFS. Se optó por evitar la duplicidad de servicios, eligiendo StoRM como "middleware" y GPFS como sistema de ficheros POSIX.

Entender la selección e implementación de StoRM como solución se SRM

para el proyecto WLCG, es un hito notable dentro del desarrollo de este trabajo. Condiciona la configuración global del sistema de almacenamiento del IFCA, que define la infraestructura técnica donde se ha desarrollado esta tesis doctoral.

2.4.2.2. El Sistema de Ficheros

Actualmente dos son los recursos de almacenamiento empleados por el IFCA para solucionar sus necesidades, GPFS y Ceph[48]. Ambos encajan dentro de la definición de lo que se denomina "**almacenamiento elástico o de convergencia**", pues los dos son capaces de proporcionar los tres tipos de almacenamiento más habituales hoy en día: Bloque, Objeto y Sistema de Ficheros.

- **Spectrum Scale (anteriormente GPFS)**: Es un sistema de ficheros distribuido de alto rendimiento desarrollado por IBM. GPFS proporciona un acceso concurrente de alta velocidad a aplicaciones que se ejecutan de forma distribuida, en múltiples nodos, mostrando una visión de disco compartido entre todos los equipos. Hay nodos que actúan como clientes y otros como servidores de disco o Network Share Disk (NSD) siendo el software independiente del hardware que proporciona los recursos. A partir de la versión 4.0 (actualmente en la 5.x.x) la denominación de GPFS ha cambiado, pasándose a llamar **Spectrum Scale**, incluyendo nuevas funcionalidades como los "**Nodos Protocolo**", que son nodos que pueden re-exportar el sistema de ficheros de GPFS mediante protocolos como CIFS o NFS. Estas nuevas funcionalidades han permitido integrarlo en entornos **Cloud**, pudiendo emplearse como BE de **Cinder** (gestor de volúmenes de OpenStack) o **Glance** (gestor de imágenes de OpenStack), re-exportando particiones NFS a máquinas en la nube mediante **Manila**[49] o incluso usando GPFS como BackEnd de **Hadoop** para realizar análisis de datos empleando Machine Learning (ML). GPFS ha sido aplicado con éxito en multitud de aplicaciones comerciales incluyendo: servicios digitales, redes de análisis y servicios de archivos escalables. Es utilizado por muchos de los mayores computadores del mundo, que aparecen en el TOP500.

- **Ceph**: Aunque Ceph no es parte de los recursos Grid, al estar integrado con OpenStack y ser el BE de almacenamiento de Cinder del IFCA, es

necesario mencionar este recurso por su utilidad en el transcurso de este trabajo Doctoral. Su funcionalidad como sistema de ficheros es reciente y no se recomienda aún su uso en producción con una elevada carga de trabajo, pero las secciones de Bloque y Objeto, están fuertemente contrastadas desde hace varios años. Ceph es la solución de almacenamiento "opensource" más usada en entornos Cloud.

2.4.2.3. Protolocos de Acceso a Datos

Los protocolos son los componentes que realizan las llamadas al sistema de ficheros para copiar, borrar, listar o modificar. Los protocolos soportados por cada sitio Grid, son definidos en el servicio SE y publicados a través del servicio BDII. Los datos son accesibles realizando llamadas directas a estos protocolos (gsiftp://, xrdcp://, htpp/s://) o enmascarando la petición a través de SRM (srm://). La evolución de los experimentos tiende a la eliminación del enmascaramiento SRM.

En VOs grandes, con cargas de trabajo elevadas, donde la ejecución de trabajos tiene ordenes de magnitud de miles de tareas de forma continuada, es necesario disponer de un sistema robusto de acceso a datos. Debe ser redundado, fiable y bien dimensionado. Para ello, el IFCA, dispone de un sistema de 8 máquinas con esta finalidad, todas ellas con acceso a 10Gbps a la red externa GEANT, siendo capaces de absorber elevados flujos de trabajo de forma continuada. Los protocolos de acceso a datos más comunes son:

- **SRM**: Storage Resource Manager (SRM)[50] es un componente de middleware cuya función es proporcionar una asignación dinámica de espacio y administrar archivos en componentes de almacenamiento compartido en Grid. Más precisamente, el SRM es un servicio Grid con varias implementaciones diferentes (Castor, dCache, StoRM, Bestman, DPM). La interfaz SRM enumera las solicitudes de servicio y estas se agrupan por funcionalidades. Las funciones de gestión de espacio permiten al cliente reservar, liberar y gestionar espacios para las diferentes calidades de espacio de almacenamiento. Las funciones de transferencia de datos tienen el propósito de llevar archivos a espacios SRM, ya sea desde el espacio del cliente o desde otros sistemas de almacenamiento remoto en Grid, y recuperarlos. Otras clases de funciones

son las funciones Directorio, Permiso y Descubrimiento.

- **Gsiftp**: Es una extensión del Protocolo de transferencia de archivos (FTP) para el Grid. El protocolo se definió dentro del grupo de trabajo Open Grid Forum . Hay múltiples implementaciones de este protocolo y el más utilizado es el proporcionado por Globus Toolkit (ver fig:2.21). El objetivo de GridFTP es proporcionar una transferencia de archivos segura y de alto rendimiento. Es el protocolo por defecto usado para copiar datos de gran tamaño. Mueve archivos entre "puntos finales" identificados por una Uniform Resource Identifier (URI)[51]. Usualmente uno de ellos es una máquina local, pero GridFTP[52] también se puede emplear para enviar datos entre dos puntos finales remotos. Los certificados X.509 son usados para gestionar la autenticación. Algunas de sus características son:

  - Usa las especificaciones Grid Security Infrastructure (GSI).
  - Transferencias paralelas.
  - Transferencias parciales de ficheros.
  - Tolera fallos y reinicios en la conexión.
  - Auto optimización del stack TCP.

- **Xroot**: Tiene como objetivo brindar un acceso escalable tolerante a fallos. Está diseñado para proporcionar un alto rendimiento sobre diferentes tipos de SEs. Se basa en una arquitectura escalable, tanto en tamaño como en rendimiento. Permite la implementación de clústeres de acceso a datos de prácticamente cualquier tamaño, que pueden incluir características sofisticadas, como autenticación/autorización (ver 2.3.4.5), integración con otros sistemas, distribución de datos WAN, etc.

  El marco general es proporcionar un acceso a datos rápido, de baja latencia y escalable. Puede servir de forma nativa cualquier tipo de dato organizado como un espacio de nombres jerárquico similar a un sistema de archivos. Se basa en el concepto de directorio. Su principal utilidad y diferencia con el resto de protocolos, es que proporciona tolerancia a fallos durante la ejecución del trabajo ante fallos de acceso local a los datos solicitados. Si por

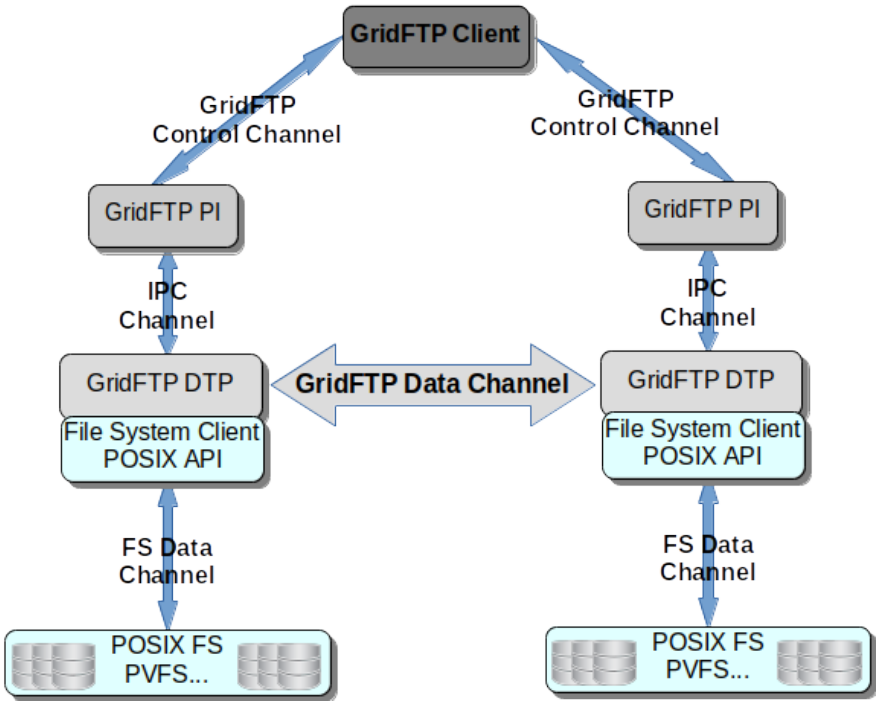


FIGURA 2.21: Diseño del protoloco Gridftp.

cualquier motivo el dato no se encuentra disponible, por estar corrupto, por haber sido eliminado o por estar bloqueado, Xroot solicita la redirección de acceso a otro centro remoto donde se encuentre disponible, realiza una carga del dato remoto por red, evitando el fallo de la tarea, minimizando de esta forma tiempos perdidos en la CPU y en el sistema de colas (ver fig:2.22).

- **Http/s (Webdav)**: El término significa "Web Distributed Authoring and Versioning", y hace referencia a la extensión del protocolo. WebDAV (ver fig:2.23) es un estándar, que describe como a través de la extensión del protocolo HTTP 1.1, pueden realizarse acciones de gestión de archivos, tales como escribir, copiar, eliminar o modificar. Webdav no sólo se trata de escribir ficheros en una ubicación utilizando HTTP, WebDav nos da la posibilidad de actuar moviendo o copiando ficheros en el servidor, modificar sus propiedades, nombre o características de seguridad, niveles de acceso etc. La finalidad de WebDAV es dar un paso maś, transformado la Web en un medio legible y editable en línea, de forma que los datos puedan

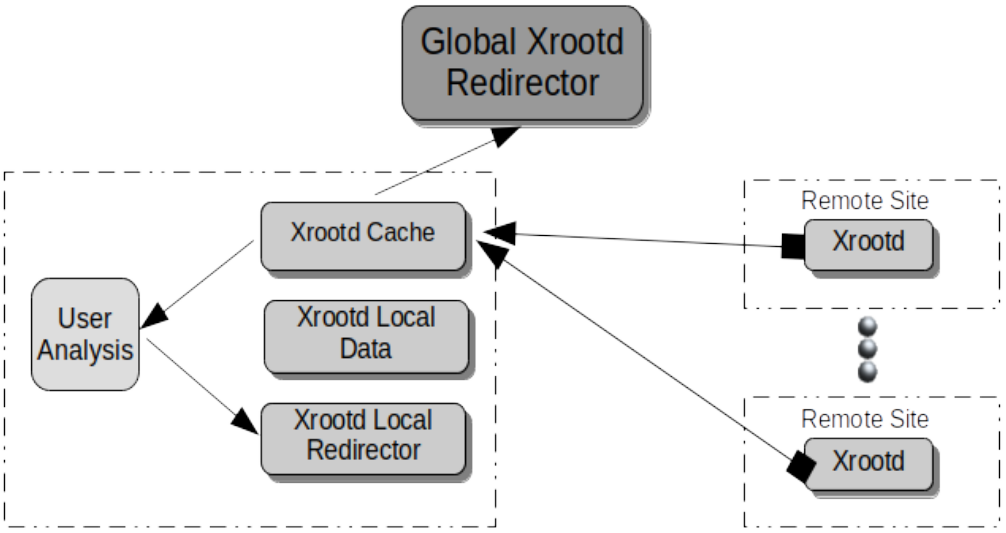


FIGURA 2.22: Diseño del servicio Xrootd

ser accedidos desde cualquier lugar. La mayoría de los sistemas operativos modernos proporcionan soporte para WebDAV, haciendo que los ficheros de un servidor se muestren en una estructura de directorio.

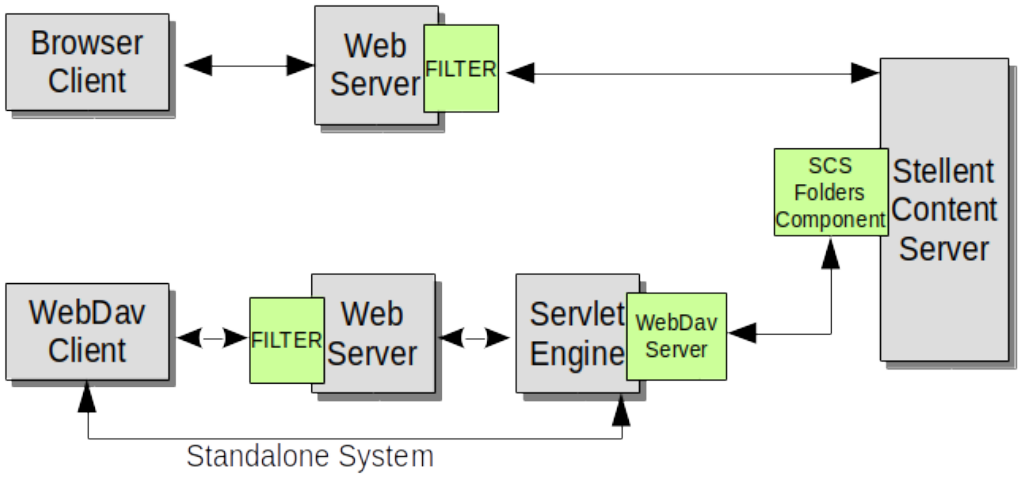


FIGURA 2.23: Diseño del protocolo WebDAV.

En la figura 2.24, podemos observar la implementación del servicio SE, que se ha mostrado anteriormente para el caso del IFCA.

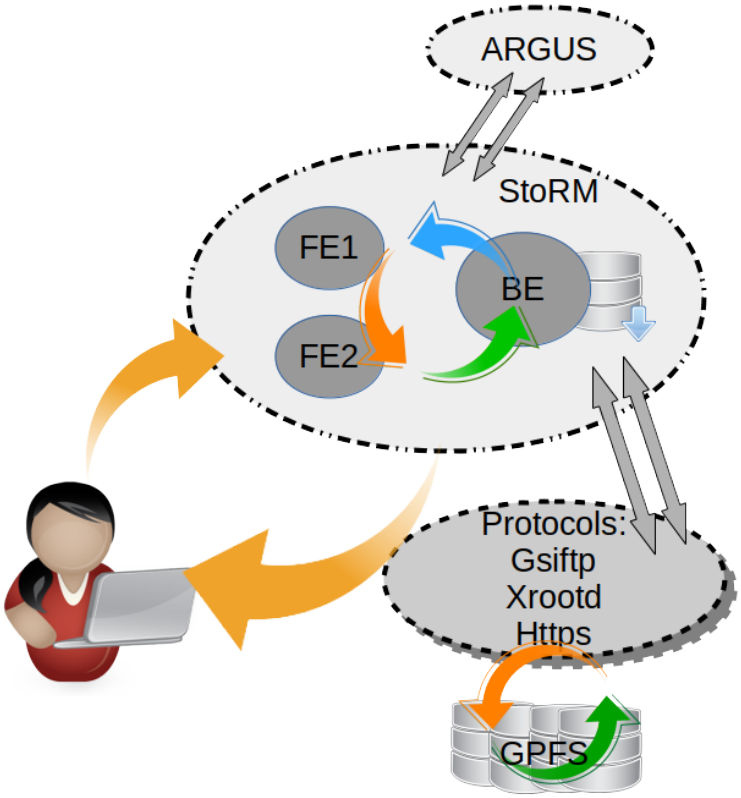


FIGURA 2.24: Conceptualización de la implementación del servicio SE en el IFCA.

Capítulo 3

Gestión y Calidad de los datos en la colaboración CMS

3.1. Introducción

A partir de este capítulo, nos focalizaremos en la colaboración CMS, que es el entorno científico y computacional donde se ha desarrollado este trabajo de tesis. Recordando lo mostrado hasta ahora, podríamos definir a CMS como una gran colaboración científica, muy prolífica científicamente hablando (ver fig:3.1), establecida en forma de VO y que emplea una gran cantidad de recursos tecnológicos para obtener sus fines. En este caso los recursos empleados son, desde un detector de partículas subatómicas de última generación, hasta recursos computacionales basados en el paradigma Grid y más recientemente empleando herramientas Cloud.

El experimento CMS se establece como un modelo de datos en niveles o Tiers; es una arquitectura jerárquica y escalonada, distribuido en 4 niveles, desde el Tier-0 al Tier-3. El modelo quedó definida por el grupo de trabajo MONARC[24]. Esta estructura de niveles, se compone por el Tier-0 situado en CERN, trece Tier-1 globalmente distribuidos y decenas de Tier-2 y Tier-3, situados en universidades e institutos de todo el mundo. Establecer y gestionar con los debidos estándares de calidad este modelo, ha sido una tarea que ha involucrado a cientos de científicos e ingenieros durante casi 15 años.

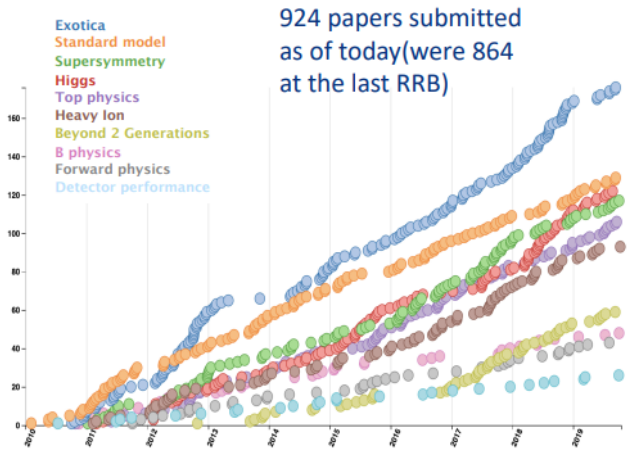


FIGURA 3.1: Publicación de artículos por la colaboración CMS desde el inicio del Experimento.
[53]

3.2. Gestión de datos en el experimento CMS

El modelo computacional de CMS ha recurrido al paradigma Grid para resolver las enormes necesidades que este experimento necesita y que un único centro no era capaz de satisfacer. Las tareas asignadas a cada grupo de centros de diferente nivel (Tier), están definidas en el documento: **CMS Physics, Technical Design Report (TDR)**[54] y los recursos computacionales aportados a la VO por cada nivel, se establecen aproximadamente de la siguiente manera: un 20 % proporcionado por el Tier-0, un 40 % por los Tier-1 y otro 40 % por los Tier-2. Los centros Tier-3 no tienen compromisos en este modelo.

3.2.1. Organización de los datos

La primera clasificación de datos que debemos hacer es la división entre datos generados mediante métodos de Monte Carlo (MC) simulando el paso de partículas a través de la materia mediante softwares como GEANT4[55], y aquellos datos reales de las colisiones p^+ - p^+ que son registrados en el detector. Estos datos son enviados al sistema de procesado del CERN en diferentes flujos de trabajo (ver fig:3.2):

- **Express**: Disponible aproximadamente dos horas después de que los datos han sido tomados; es empleado para tareas de retro-alimentación y calibración.

- **Alineamiento y Calibración (AlCa) streams**: Selección de eventos dedicados a calibración.

- **Physics Streams**: Se dividen en conjuntos de "Primary Datasets" y son reconstruidos para análisis de física. Este proceso puede demorarse hasta 48h. Son sucesos que proceden directamente del detector, agrupados por conjuntos que satisfacen el criterio definido en el "Trigger de alto nivel" (High Level Trigger). El sistema de CMS Online Data Acquisition and Trigger System (TriDAS) selecciona de los millones de eventos registrados en el detector, los 100∼150 más interesantes por segundo y los almacena para su posterior procesamiento. La selección de eventos se realiza usando filtros basados en selecciones simples y de corta duración a nivel de hardware y otros más sofisticados, que requieren mucho más tiempo para ejecutarse implementados mediante software. Al final, el sistema crea eventos en formato tipo RAW que contienen los resultados de la selección final de HLT, y objetos de nivel superior creados durante el procesamiento.

Durante el el segundo periodo de toma de datos "**RUN 2**", CMS, almacenó 74PB de datos en el CERN, de ellos 32PB corresponden a datos simulados y el resto a "Physics Streams".

Comenzando por los datos tipo RAW obtenidos desde el sistema de adquisición en línea, sucesivos grados de procesamiento (reconstrucción de eventos) refinarán estos datos, aplicando calibraciones y creando objetos de física de alto nivel. La información disponible en cada evento después de su paso por las diferentes cadenas de reconstrucción y simulación, se agrupa en diferentes formatos[56]:

- **RAW**: Información completa del evento en el Tier-0, es decir, el CERN. Contiene información del detector sin ningún tipo de procesado, sólo se realiza un filtrado a nivel hardware de los sucesos que a priori tienen visos de ser interesantes. Este formato no se usa en análisis y se almacena en el Tier-0 tal como se genera.

- **RECO**: La primera etapa del procesamiento en el Tier-0. Esta capa contiene objetos físicos reconstruidos en los diversos subdetectores, pero aún es muy detallada, desde "hits" hasta la reconstrucción de objetos más completos como leptones. RECO puede usarse para análisis, pero es demasiado pesado para uso frecuente o intenso, ralentizaría mucho tiempo el análisis, dado el tamaño de este tipo de datos.
- **AOD (Analysis Object Data)**: (∼40 % del tamaño RECO) Contiene información de bajo nivel, pero en menor cantidad que en los eventos RECO. Se utiliza para la mayoría de los análisis. El formato Analysis Object Data (AOD) proporciona una compensación entre el tamaño del evento y la complejidad de la información disponible, optimizando la flexibilidad y la velocidad de los análisis. Existen también versiones reducidas de este formato, que son cada vez más empleadas para los análisis: MiniAOD (∼15 % del tamaño AOD) y NanoAOD (<1 % del tamaño AOD).
- **FEVT**: RAW+RECO
- **GEN**: Evento generado de Monte Carlo.
- **SIM**: Deposiciones de energía de partículas MC en el detector.
- **DIGI**: Hits convertidos en la respuesta del detector. Básicamente equivalente a la salida RAW del detector.

La tabla 3.1, muestra el tamaño aproximado de los diferentes Data Tiers.

Tabla 3.1: Tamaño en MB de los diferentes niveles de datos.

	RAW	**RECO**	**AOD**	**FEVT**	**GEN**	**SIM**	**DIGI**
Size (MB)	0.70∼0.75	1.3∼1.4	0.05	1.75	–	–	1.5

3.2.2. Flujo de datos de los centros distribuidos de CMS

CMS utiliza una infraestructura distribuida de centros informáticos interconectados mediante redes de alta capacidad a través de GÉANT que es la red

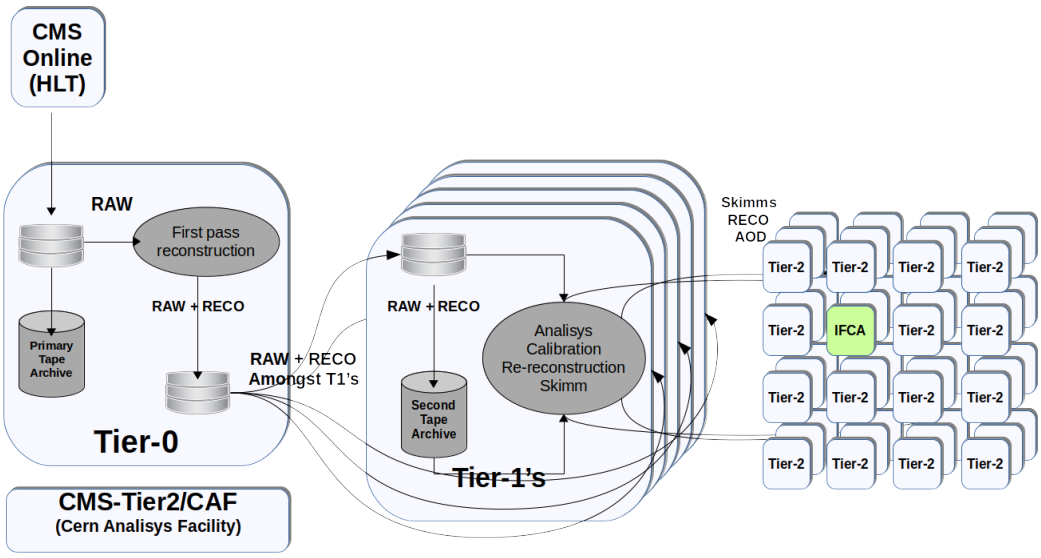


FIGURA 3.2: Flujo de Datos en la infraestructura Tier.

paneuropea que interconecta a todas las redes nacionales europeas de investigación entre sí y con aquellas otras ubicadas en otros países como Estados Unidos, América Latina, Asia,etc. Durante el año 2019 esta infraestructura está siendo mejorada con nuevos enlaces de fibra oscura y lo que es más significativo, con nuevo equipamiento óptico que, entre otras características, permitirá desplegar enlaces a 100Gbps. Estas mejoras supondrán un beneficio para los centros que necesitan redes de alta capacidad, ya que tendrán mejor conectividad para comunicarse con sus centros homólogos en los proyectos de investigación internacionales en los que se participa conjuntamente[57]. La capacidad actual de conectividad que proporciona la red GÉANT se muestra en la figura 3.3.

Como hemos visto en el capítulo 2, el experimento ha ido evolucionando con el paso del tiempo; se han realizado cambios en el "middleware" reemplazando servicios antiguos por otros más sofisticados, que implementan mejoras significativas. Lo mismo ha sucedido con el esquema de interconectividad de modelo de Tiers dentro del WLCG que se definió inicialmente.

El sistema inicial se diseñó en forma de cascada, de modo que los datos fluían

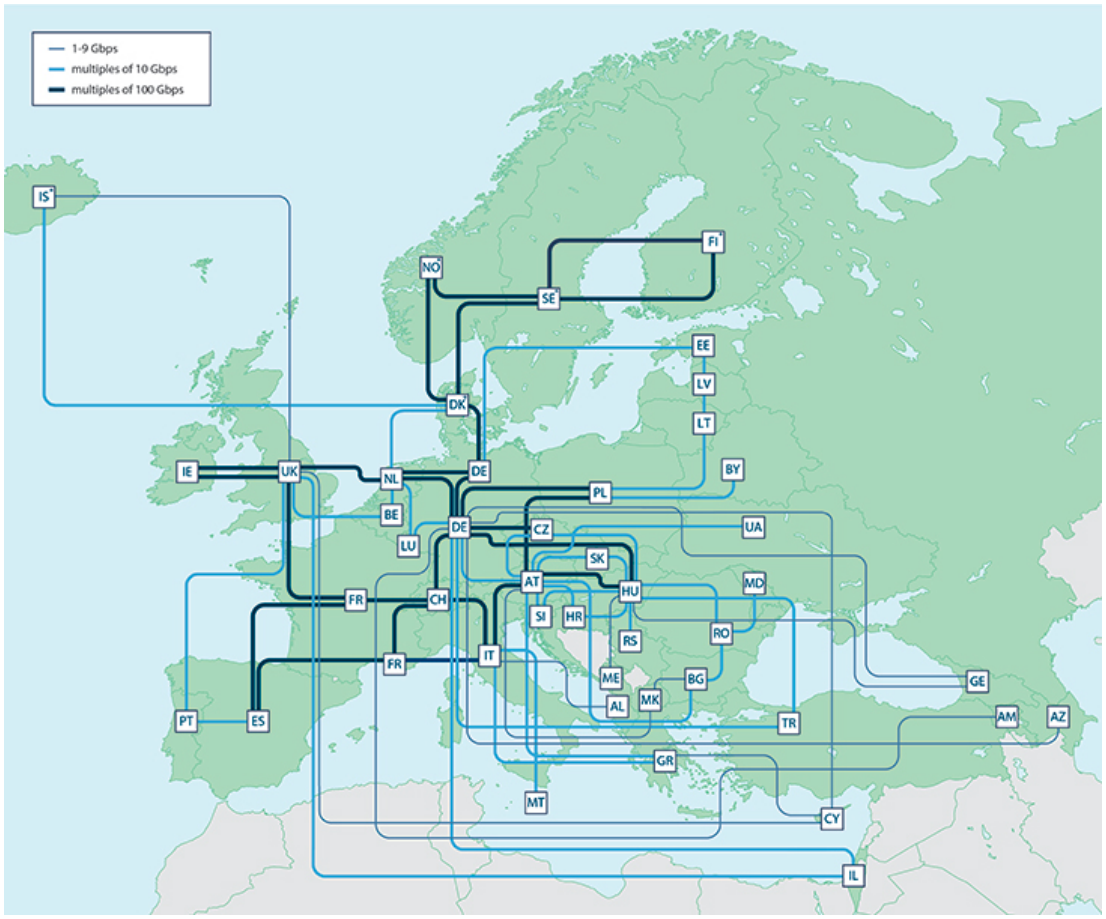


FIGURA 3.3: Mapa de red GÉANT.

desde el Tier-0 a los centros Tier-1 y de cada Tier-1 a sus Tier-2 asociados, es decir que cada Tier-2 tenía un Tier-1 de referencia y el flujo de datos era en exclusiva con este Tier-1. Era una estructura totalmente jerarquizada (ver fig:3.4).

Con la mejora de las capacidades de acceso a GÉANT por parte de los Tier-2, muchos de ellos, entre los que se encontraba el IFCA, actualizaron sus conexiones de red, pasando de 1∼2.5Gbps a 10Gbps. Se comprobó que el modelo inicial de red jerarquizado estaba obsoleto, ya que aprovechaba un mínimo de las capacidades de ancho de banda de los Tier-2. El modelo de transferencia de datos se ajustó a los nuevos tiempos quedando en un modelo donde los datos fluyen libremente entre los centros Tier-1 y los centros Tier-2 (ver fig:3.5). Desde el IFCA, hubo que implementar nuevos equipamientos de electrónica de Red, estableciendo canales

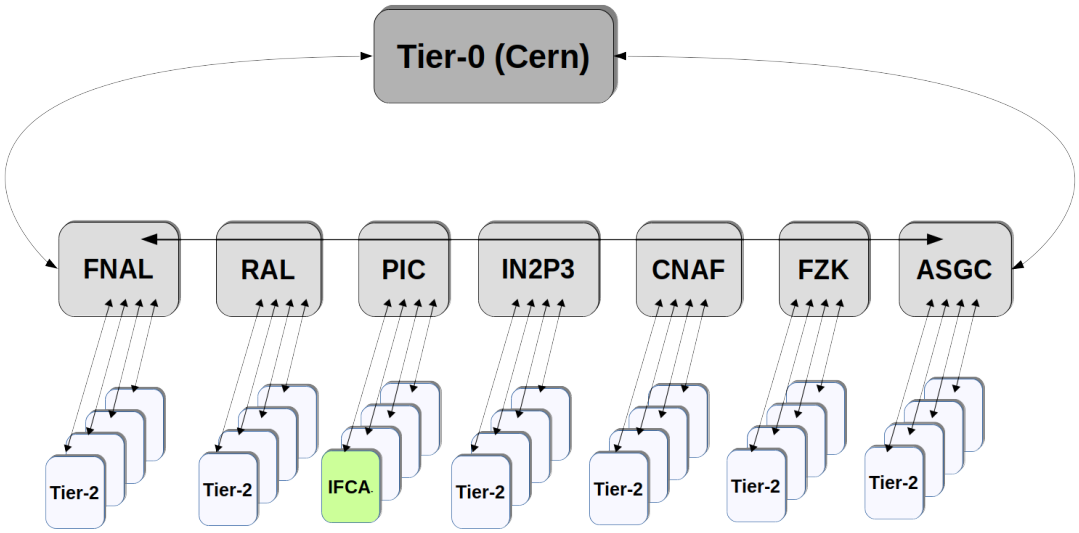


FIGURA 3.4: Diseño del flujo de Datos de CMS en la estructura de Tiers.

de transferencia de datos de 10Gbps, entre el IFCA, la Universidad de Cantabria y RedIris adecuándose así a los nuevos anchos de banda proporcionados por GÉANT.

El flujo de datos de CMS, parte del Tier-0 que es el generador y dependiendo de su formato y utilidad son repartidos entre los diferentes centros de los niveles del modelo (Tiers). La descripción del flujo de datos[58] sería la siguiente:

- Los datos producidos en el detector son seleccionados y se re-formatean en archivos RAW, utilizando los recursos computacionales del Tier-0 en el CERN. Una copia de estos datos es archivada localmente y una segunda copia se envía a los servicios de archivado fuera del CERN. Una tercera, es transferida al almacenamiento en disco que proporcionan los centros Tier-1. La transferencia de los datos, puede realizarse desde el propio CERN o desde el archivado remoto, esta decisión es tomada de forma autónoma, por el sistema de encaminamiento de réplicas de CMS.

- Los eventos en formato Event Summary Data (ESD) o RECO, son almacenados en disco y se mantienen durante varios meses esperando la verificación

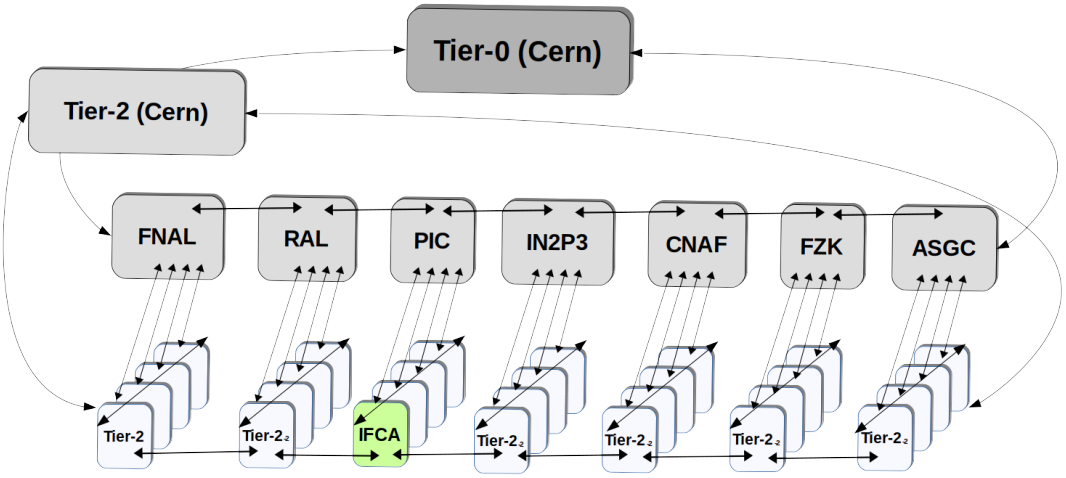


FIGURA 3.5: Diseño a 2019 de la estructura de flujo de Datos de CMS.

del detector y análisis específicos, mientras que una de las copias AOD es almacenada en disco y otra es enviada al servicio de archivado para ser almacenada de forma permanente.

- AOD y RECO son repartidos entre centros Tier-2, proporcionando al usuario final un mejor acceso a los datos de análisis. La replicación y la retención en los Tier-2, queda establecida por la popularidad del conjunto de datos.
- Las muestras de simulación de Tier-2 son transferidas continuamente al almacenamiento en disco de los Tier-1. Una vez que se completan los pasos finales de procesamiento y se validan los datos, son replicados enviándolos al servicio de archivado.
- Las n-tuplas de grupo, normalmente no son replicadas por los servicios provistos por CMS a menos que las muestras sean elevadas a conjuntos de datos oficiales a través de los pertinentes procesos de validación y publicación.

Para gestionar el movimiento de datos entre los diferentes niveles y centros de computación, en CMS se desarrolló un sistema de gestión de transferencia de datos llamado Physics Experiment Data Export (PhEDEx). PhEDEx ofrece a los

administradores y usuarios de los Tier una vista en tiempo real del estado global de transferencia de datos. Está provisto de un sistema centralizado para tomar decisiones de movimiento de datos, seleccionando de forma autónoma los canales más óptimos, de manera que realiza muchas de las operaciones de datos de bajo nivel de forma automatizada: replicación de datos a gran escala, encaminamiento, verificación de migraciones a cinta, consistencia, integridad de datos, etc[59].

3.2.2.1. El centro de datos Tier-0

Todos los datos del LHC pasan a través de este "HUB" central. Aunque el Tier-0, solo proporciona alrededor del 20 % de los recursos computacionales disponibles por la VO CMS, es el responsable de la custodia de los datos, originales, sin procesar. El detector realiza millones de lecturas digitales (ver 2.1.2), seleccionando aproximadamente entre 100∼150 eventos cada segundo; estos datos sin procesar filtrados por el sistema TriDAS (ver fig:3.6) son enviados al Tier-0.

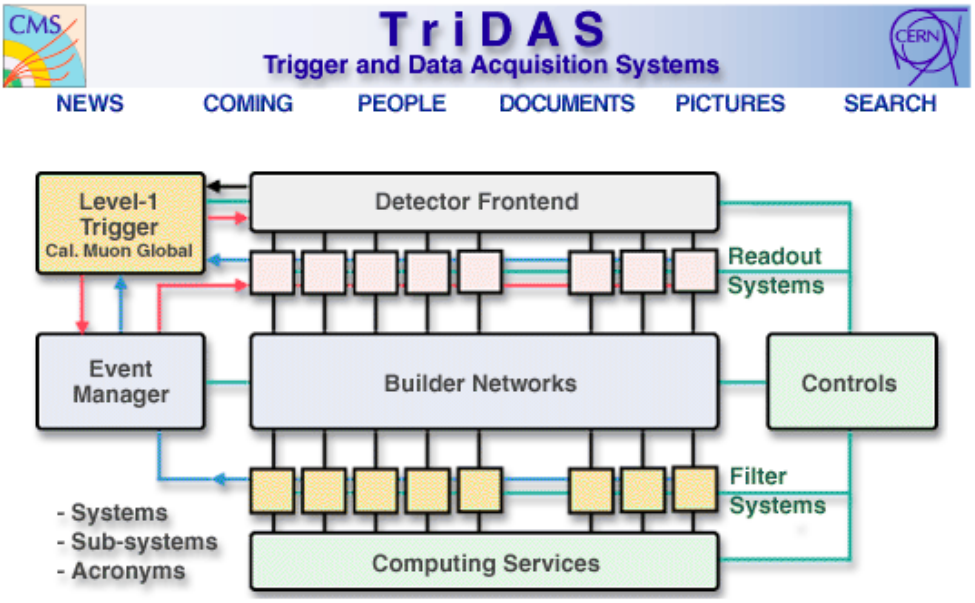


FIGURA 3.6: Sistema de adquisición de datos TriDAS.

El Tier-0, realiza una primer procesado de los datos en bruto en información significativa. Estos datos se vuelven a empaquetar en conjuntos conocidos como "Primary Dataset" y son enviados al sistema de archivado en cinta (normalmente en

tamaños de 2-3 GB). Una copia de estos datos se distribuyen entre la siguiente capa de modelo, los Tier-1, por lo que ahora existen dos copias de cada pieza de datos. Se realiza una primera reconstrucción, generando los ficheros RECOnstructed Data (RECO) y AOD que son distribuidos también entre los Tier-1.

Los recursos computaciones comprometidos por el Tier-0 para el año 2019 con la VO CMS ascienden a unos valores aproximados de 26PB de almacenamiento de datos en disco, 99TB de almacenamiento en cinta y aproximadamente 423.000 HEP-SPEC06 de CPU. El Tier-0 también actúa como centro de reprocesado de datos cuando en el LHC no se toma datos, pero no proporciona recursos de análisis.

3.2.2.2. Los Tier-1

los Tier-1[60], son grandes centros computacionales en países que participan en la colaboración de CMS. Pueden ser grandes laboratorios como Fermilab National Laboratory (FNAL) y Rutherford Appleton Laboratory (RAL) o instituciones más modestas como el Port dÍnformació Científica (PIC). Aunque existen trece Tier-1 dentro del WLCG, solamente siete de ellos aportan recursos a la colaboración CMS. Los Tier-1 en general se utilizan para actividades organizadas centralmente a gran escala, pueden proporcionar datos y recibir las producciones de todos los Tier-2. Proporcionan aproximadamente el 40 % restante de los recursos computacionales distribuidos.

Los recursos computaciones comprometidos por los Tier-1 para el año 2019 con la VO CMS ascienden a unos valores aproximados de 68PB de almacenamiento de datos en disco, 220TB de almacenamiento en cinta y aproximadamente 650.000 unidades HEP-SPEC06[61] de computación.

Son responsabilidad de los centros Tier-1:

- Recibir un subconjunto de los datos del Tier-0 relacionado con el tamaño de los recursos comprometidos en el Memorandum of Understanding (MoU) del WLCG.

- Almacenar parte de los datos RAW, segunda copia segura.

- Proporcionar potencia de procesamiento para tareas programadas centralmente en la colaboración. Alguna de sus tareas de procesamiento son:
  - Reconstrucción.
  - Skimming.
  - Calibración.
  - No proporciona recursos de análisis a los usuarios finales.
- Almacenar una copia completa de la AOD.
- Distribuir RECO, skims y AOD a los otros centros Tier-1 y CERN, así como al grupo asociado de centros Tier-2.
- Proporcionar almacenamiento seguro y redistribución para los eventos de MC generados por los Tier-2 y los generados centralmente.

Los centros Tier-1 proporcionan un entorno de ejecución muy controlado donde el equipo asignado a la producción gestiona los recursos. Los únicos otros usuarios autorizados para procesar datos en los Tier-1, son miembros de grupos de física que realizan análisis de alta prioridad.

La conectividad se realiza mediante fibras ópticas de alta capacidad, 10-100Gbps. Esta red de banda ancha dedicada recibe el nombre de "LCH Optical private Network" (LHCOPN) (ver fig:3.7) y establece la topología de red, para establecer las conexiones y la transferencia de datos entre centros Tier-1 y el Tier-0 dentro del WLCG[62]. LHCOPN establece las configuraciones de IP, direccionamiento y encaminamiento sobre segmentos de GÉANT para el proyecto WLCG.

3.2.2.3. Los Tier-2

Los centros Tier-2 en CMS son la única ubicación, además de la instalación de análisis especializado en CERN (Tier-2 CERN, antigua CAF), donde los usuarios pueden obtener acceso garantizado a las muestras de datos de CMS y recursos computacionales de análisis, siempre de acuerdo con las credenciales proporcionadas por los diferentes roles de la VO de CMS. Al igual que los

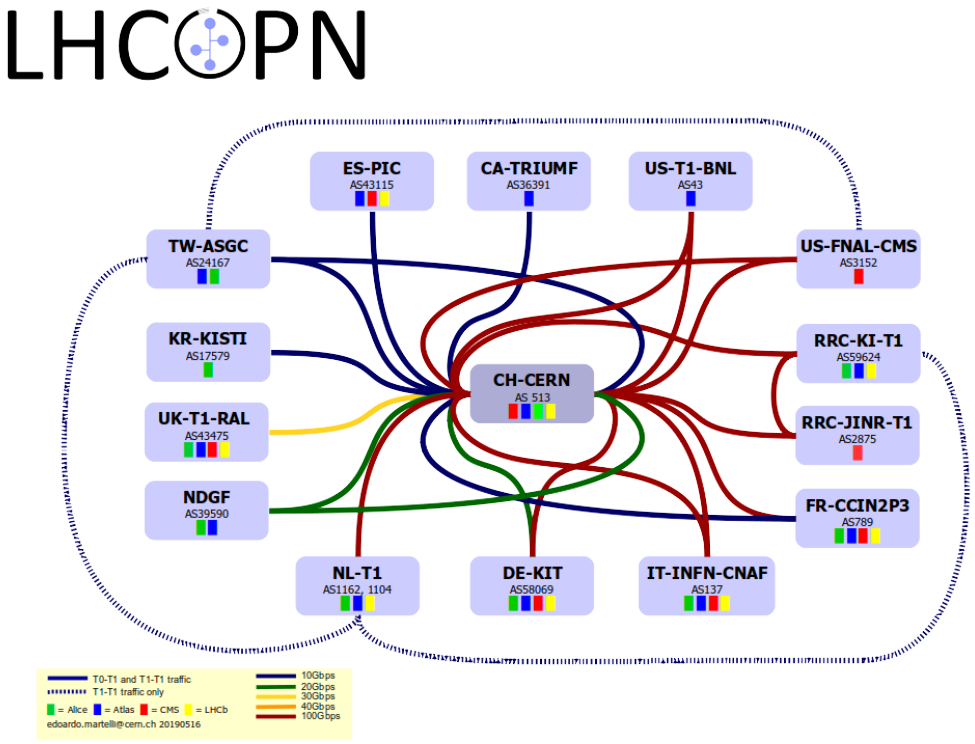


FIGURA 3.7: Modelo de red LHCOPN.

Tier-1, proporcionan aproximadamente el 40 % de los recursos computacionales distribuidos, aproximadamente ∼80.000 núcleos de procesamiento y ∼150PB de almacenamiento en disco.

Los recursos computaciones comprometidos por los Tier-2 para el año 2019 con la VO CMS ascienden a unos valores aproximados de 78PB de almacenamiento de datos en disco y aproximadamente 1.000.000 unidades HEP-SPEC06 de computación.

Los Tier-2 están también especializados en la importación/exportación de datos. Deben estar provistos de una capacidad de red suficiente para permitir que los centros actualicen los datos de forma regular y los usuarios de la VO puedan realizar sus análisis. La conectividad se realiza mediante fibras ópticas de alta capacidad, en su mayoría de 10Gbps. Esta red de banda ancha dedicada recibe el nombre de "LHC Open Network Environment (LHCONE)" y establece

las configuraciones de IP, direccionamiento y encaminamiento sobre segmentos de GÉANT enlazando los centros Tier-2, como el IFCA, con los centros Tier-1 y el Tier-0.

Con el fin de gestionar esta cantidad enorme de recursos geográficamente distribuidos, CMS ha intentado introducir una política y estructura en el almacenamiento y procesamiento de los Tier-2. Por lo general son centros medianos o pequeñas instituciones, departamentos universitarios, pero a menudo tienen altos recursos informáticos que proporcionan una gran capacidad de cómputo para generación de MC, análisis de usuarios o estudios de calibración; además proporciona el área final de almacenamiento para los usuarios, según lo acordado en los "Compromisos". Estas áreas de usuario, pueden ser temporales, con caducidad aproximada de un mes, o perennes si hablamos de usuarios locales, es decir aquellos que pertenecen a la institución que aporta los recursos a la VO. La asignación de almacenamiento en el Tier-2 del IFCA se distribuye de la siguiente forma:

- 30TB Servicios centrales, como producción de Monte Carlo y servicio de "buffer".
- 200TB de espacio central para alojar los "paquetes de datos"de mayor interés para la colaboración 200TB. Este espacio es gestionado centralmente por el grupo/rol "AnalysisOperations" y es usado para almacenar skims primarios o ejemplos de Monte Carlo interesantes.
- 125TB Para el grupo/rol "Analysgroup". Es un espacio para datos relacionados con el detector y los grupos de física, aquí se almacenan "paquetes de datos" interesantes para los grupos de física asociados al Tier-2, y por norma general también para sus usuarios locales de análisis (Forward physics, QCD, Higgs, Electroweak, Top, Exotica...).
- 100GB por usuario de espacio de almacenamiento usado a través de Grid, para usuarios con el rol local o nacional. Esta cuota asignada depende de los recursos disponibles en cada Tier-2 y puede ser ajustada dependiendo de las circunstancias. Es el principal destino de herramientas de envío de trabajos como CMS Remote Analysis Builder (CRAB) (ver 2.3.4.4).

- 170TB como espacio local. Se almacenan los datos de especial interés para la comunidad local o nacional. La gestión de este área, movimiento, borrado, archivado, etc, está bajo responsabilidad total de los administradores del Tier.

Aquellos Tier-2 con más recursos que los comprometidos con la VO, pueden pueden dedicar este exceso como recursos adicional para el espacio central, el de grupo, o el espacio dedicado a los usuarios locales, mientras que aquellos Tiers con déficit de recursos de forma puntual; pueden asignar recursos a un solo grupo de física, solo a espacio central, o si es lo suficientemente pequeño, unicamente a la producción simulada de eventos.

El IFCA Tier-2 desplegó 1,7PB de almacenamiento y aproximadamente 27.249 HEP-SPECS06 (2.8 % del total de los Tier-2) para la VO de CMS en 2018. Proporciona un canal de conexión de 10Gbps (ver fig 3.8), no solo al experimento CMS, sino a todos los experimentos en los que el IFCA participa.

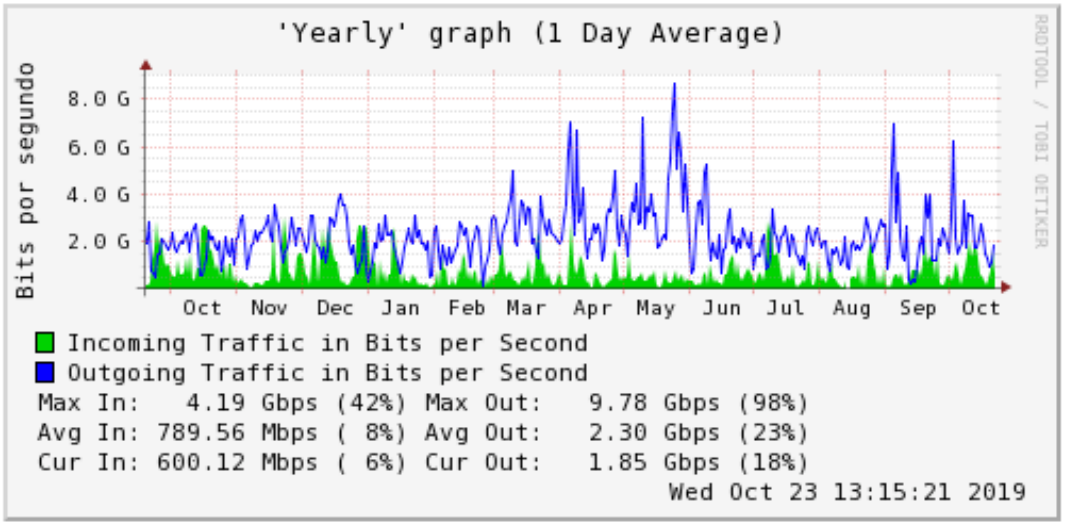


FIGURA 3.8: Ancho de banda utilizado por canal GÉANT del IFCA durante el año 2019.

La expectativa CMS para la capacidad global en los Tier-2 para 2019 es de 72PB de espacio en disco utilizable y 961.493 HEP-SPECS06[63].

3.2.2.4. El Tier-3

Los centros Tier-3 son infraestructuras de computación, por norma general de tamaño pequeño. Satisfacen las necesidades computacionales de los usuarios locales. Proporcionan recursos y servicios a CMS de forma oportunista. Dependiendo de su tamaño, los centros Tier-3 suponen una componente importante en la capacidad de análisis de CMS. Participan en las actividades de computación bajo la coordinación de un centro Tier-2 específico y pueden proporcionar servicios como el desarrollo de software, análisis interactivos finales o producciones de MC.

Parte de la infraestructura computacional del IFCA actúa como Tier-3, nutriéndose de las ventajas en el acceso a datos locales del experimento, al estar alojado en las mismas instalaciones que el Tier-2. Durante algún tiempo el IFCA actuó como Tier-2 de referencia para el Tier-3 localizado en la universidad de Oviedo, y operado por el grupo de física de partículas de esta Universidad, de esta forma aquellos datos de interés para este grupo eran subscritos al Tier-2 IFCA y desde ahí descargados por el Tier-3 mediante un canal de transferencia de datos, establecido entre ambos centros y gestionado por PhEDEx de forma automática,

3.2.2.5. Central Analysis Facility (CAF)

El acceso a la CERN Analysis Facility (CAF), garantiza a CMS disponer de una instalación de procesado de datos que permite flujos de trabajo de alta prioridad y baja latencia, con altas velocidades de acceso a datos, mientras se aplican políticas de acceso controlado y priorizado a cientos de usuarios. Estas actividades son necesarias para asegurar que el detector funciona de forma eficiente, correcta y estable. Actualmente integrada dentro del Tier-2 del CERN. Las principales actividades desarrolladas por este Tier-2/CAF son:

- Datos de calibración y alineamiento para entrenar algoritmos de HLT o la primera reconstrucción.
- Diagnóstico de problemas en el detector.
- Actividades relacionadas con el rendimiento: optimización,reconfiguración

Como actividades secundarias siempre que no interfieran con las señaladas anteriormente, ésta infraestructura también se emplea para:

- Proporcionar servicios de análisis similares a los de un Tier-2.
- Acceso interactivo.
- Reprocesamiento de datos y producción de muestras MC.

3.2.3. El Sistema de transferencia de Ficheros (FTS)

Es el servicio de movimiento de datos de nivel más bajo definido en la arquitectura gLite. Es responsable de mover conjuntos de archivos de un sitio a otro, lo que permite a los sitios participantes controlar el uso de los recursos de la red. Está diseñado para el movimiento punto a punto de archivos físicos. El FTS tiene interfaces dedicadas para administrar los recursos de la red y mostrar estadísticas de transferencias en curso. Admite "Logical Files Names (LFN)", es decir, puede proporcionar la búsqueda y el registro del catálogo.

Es el servicio responsable de distribuir globalmente la mayoría de los datos del LHC a través de la infraestructura de WLCG. Actualmente la versión 3 se encuentra en producción. Ha sido diseñado de forma modular, para ser versátil, permitiendo una escalabilidad eficiente. El servicio se escala horizontalmente muy bien al agregar más recursos con una configuración idéntica en forma de clúster FTS3, ya que la configuración se almacena en la base de datos y se lee durante el inicio del servicio. Sólo se usa un archivo de configuración para almacenar las credenciales de la base de datos.

FTS3[64] se basa en GFAL2, una nueva biblioteca de abstracción que oculta, todo lo que es posible, los detalles subyacentes del protocolo. Esto significa que FTS3 es capaz de ejecutar una transferencia siempre que haya un complemento GFAL2 disponible para el protocolo. A FTS3 no le importa el tipo de protocolo que se está utilizando GridFTP, XRoot, SRM, S3, GCLOUD o HTTP para realizar la transferencia. Los componentes del FTS, se muestran en la figura 3.9.

El "Scheduler" o planificador prioriza las transferencias de un enlace de acuerdo a los siguientes parámetros:

- La prioridad de la transferencia.
- El peso asignado a la Actividad.

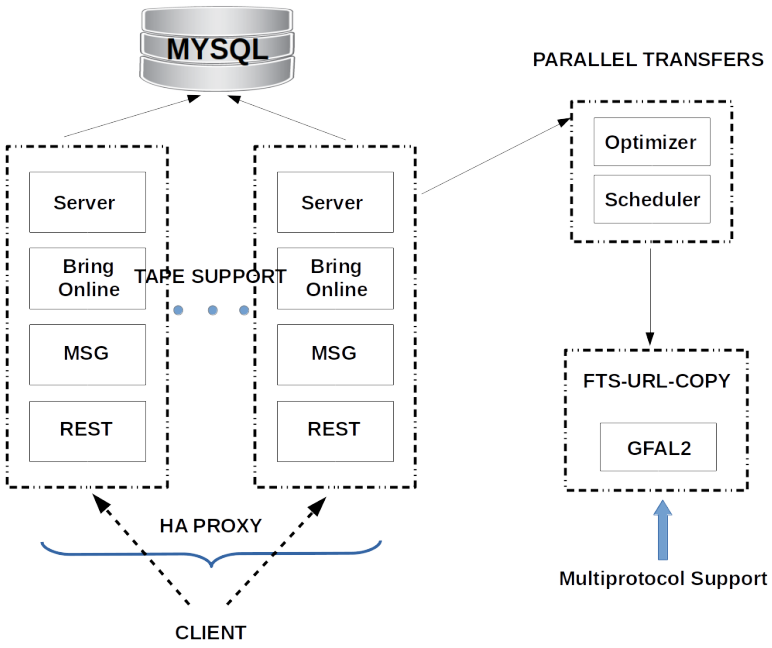


FIGURA 3.9: Arquitectura del servicio FTS3.

- El peso asignado a la VO.

El "Optimizer" asigna los "slot" a los canales dependiendo de la tasa de éxito. La paralelizacíon de la transferencia también es optimizada dependiendo del tamaño del fichero a transferir y de la cola asignada.

A continuación se muestran un par de imágenes (ver fig:3.10 y fig:3.11) de la utilización del servicio FTS3 por parte del IFCA, dentro de la colaboración CMS:

La funcionalidad principal de FTS3 se amplía con varias herramientas orientadas a la web, como la supervisión versátil y la interfaz de usuario de WebFTS, que veremos más adelante, con soporte de identidad federada (IdF).

3.2.3.1. WebFTS

Ya integrada en los marcos experimentales de LHC. La nueva interfaz WebFTS ahora hace que la tecnología de transferencia del FTS3 esté directamente disponible para los usuarios finales. WebFTS [65],es una interfaz intuitiva, que permite a

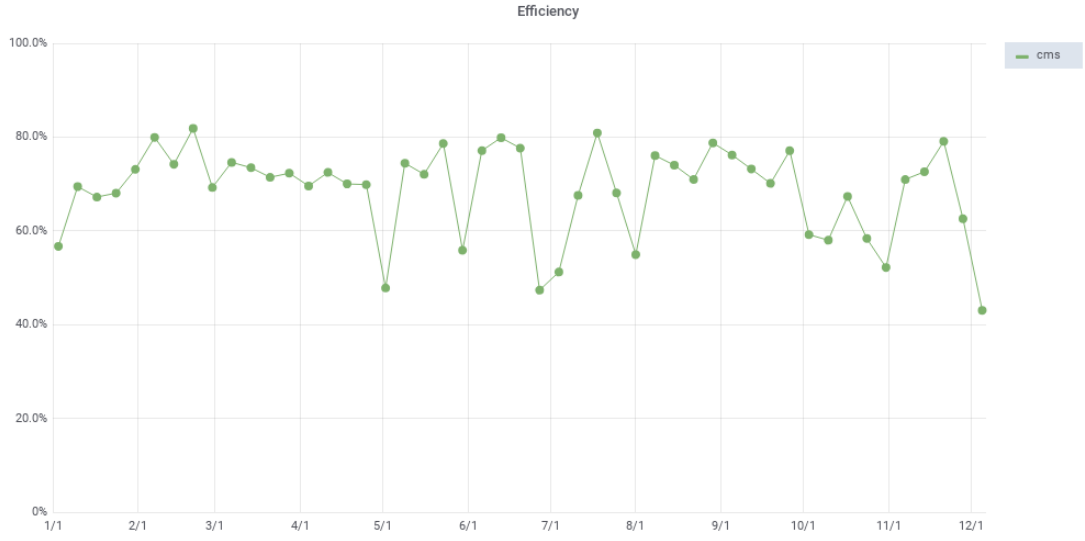


FIGURA 3.10: Eficiencia del servicio FTS3 para la VO CMS y el Tier-2 del IFCA.

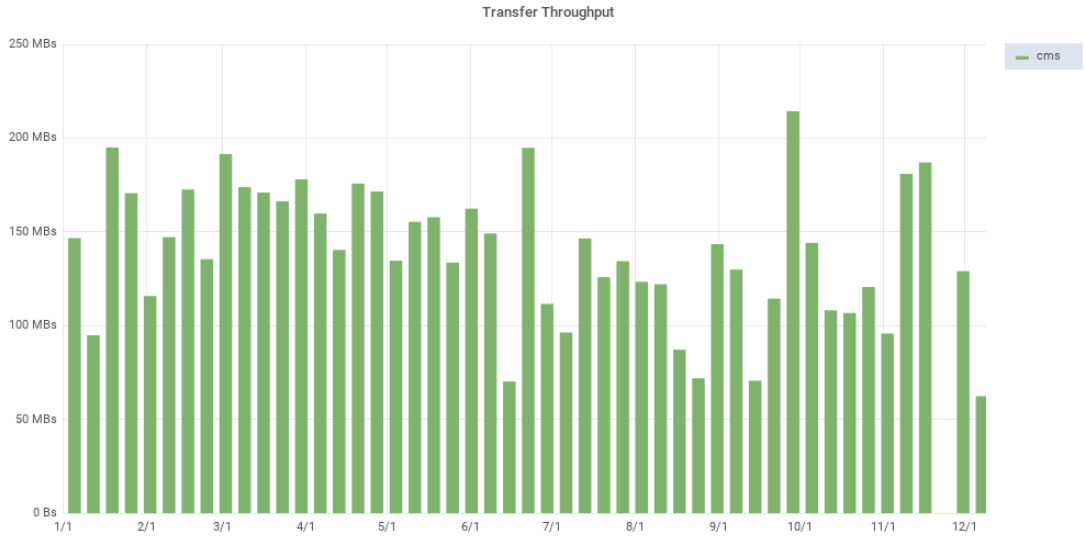


FIGURA 3.11: Ancho de banda en MB/s del servicio FTS3 para la VO cms y el Tier-2 del IFCA.

los usuarios programar y administrar fácilmente grandes transferencias de datos directamente desde el navegador, aprovechando un servicio que ha sido probado en la escala de Petabytes por mes. Los límites de transferencias FTS3, se establecen fuera de la tecnología Grid soportando puntos finales como Dropbox y S3 (ver

fig:3.12).

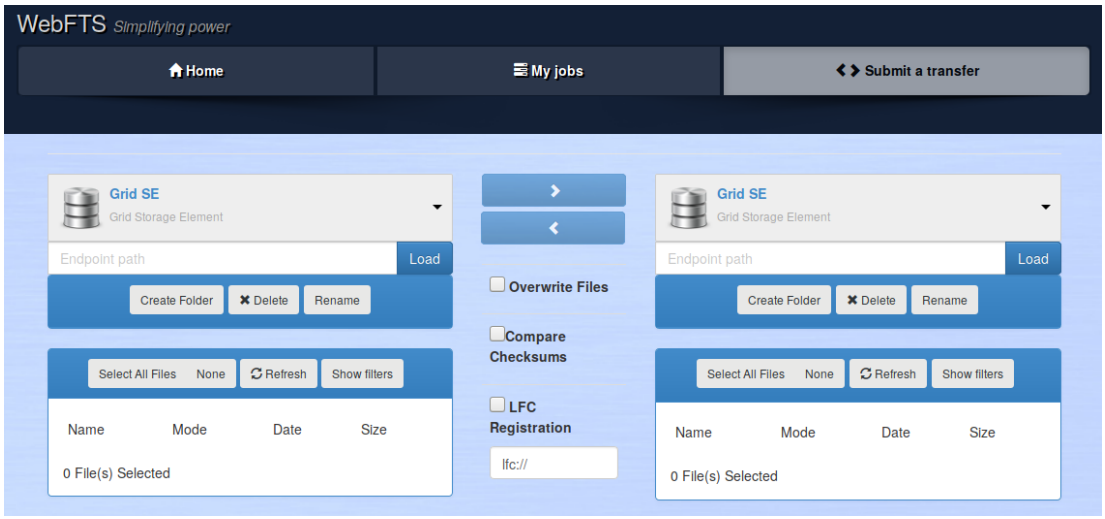


FIGURA 3.12: Interfaz Web de la utilidad WebFTS alojada en el CERN. [66]

Sus principales características son:

- La escalabilidad horizontal del servicio.
- La supervisión avanzada.
- La "configuración cero" para la implementación gracias a la lógica de optimización de transferencia especializada.

El administrador de datos dispone de herramientas para la gestión de los parámetros de transferencia FTS3, como los límites de ancho de banda y las transferencias máximas de archivos activos por punto final y VO, prohibición de usuarios y puntos finales, así como potentes herramientas de línea de comandos.

WebFTS proporciona soporte de tecnologías de identidad federada, demostrando así el uso de los recursos Grid sin la carga de la administración de certificados X.509. De esta manera, FTS3 puede cubrir las necesidades de una amplia gama de estamentos, desde usuarios ocasionales hasta servicios dedicados de alta carga.

En un intento de hacer que todas las funciones de FTS3 sean fácilmente accesibles para los usuarios finales. Estos pueden enviar y monitorizar sus propios trabajos de transferencia. WebFTS proporciona el mismo soporte multi-protocolo

que FTS3, lo que lo convierte en una herramienta muy útil para transferir archivos entre recursos Grid y no Grid. Se decidió agregar soporte de identidad federada (IdF) a WebFTS, lo que permite una transición transparente de credenciales basadas en X.509 a las basadas en web (SSO) con la ayuda de dos servicios adicionales:

- STS (Security Token Service) es un servicio que consume la aserción SAML2 producida por SSO habilitado para IdF y lo transforma en un certificado X.509 de corta duración.

- IOTA CA (Autoridad de Certificación de Garantía de Confianza de Solo Identificador) es una CA con perfil específico que es elegible para emitir certificados X.509 de corta duración que son necesarios para STS.

3.2.4. PhEDEx

PhEDEx es el sistema de gestión de transferencia de datos de CMS[67][68]. Este software, está presente en cada uno de los centros Tier con responsabilidad dentro de la estructura de de flujo de datos de CMS, es decir en el Tier-0, los Tier-1 y los Tier-2 como es el caso IFCA, desde el año 2005. PhEDEx es responsable de transportar datos entre los centros de CMS, así como de realizar un seguimiento de los datos existentes en cada sitio. Está diseñado para manejar las tareas asignadas con un esfuerzo mínimo del operador, automatizando la mayor parte de los flujos de trabajo, desde la distribución a gran escala de los conjuntos de datos de experimentos HEP, hasta las órdenes de borrado programadas centralmente, la verificación de los bloques transferidos o la migración a cinta.

La unidad más pequeña en el espacio de cómputo es el bloque de archivos, que corresponde a un grupo de archivos ROOT a los que se puede acceder juntos. Los bloques de agrupan a su vez en “Datasets”. El bloque es la unidad mínima disponible para ser transferida a través del sistema de transferencia de réplicas PhEDEx.

Esto requiere una asignación de la abstracción entre la colección de eventos y la ubicación de los archivo. Para ellos CMS dispone de un catálogo de datos global llamado CMS Data Aggregation System (DAS), que proporciona la asignación

entre el conjunto de datos o colección de eventos y la lista de bloques de archivos correspondientes a esta abstracción, dando al usuario una visión general de lo que está disponible para el análisis, ya que dispone del catálogo completo de la ubicación de estos bloques de archivos dentro de CMS. Varios centros pueden proporcionar acceso al mismo bloque de archivos, y/o "Dataset" y es PhEDEx el encargado de establecer las políticas de encaminamiento de datos entre Tiers.

PhEDEx tiene conocimiento de los ficheros, mediante el proceso conocido como "Data Injection". En este proceso, los atributos de los ficheros son insertados en la base de datos central de PhEDEx la Transfer Management Data Base (TMDB) inyectando el LFN, que es la ruta de acceso común a los datos para todos los Tier de CMS independientemente de la tecnología de almacenamiento subyacente que sea empleda por cada centro.

La gestión de datos distribuida en el marco de LHC es un desafío complejo, con dos problemáticas diferentes. La primera de ellas es administrar, asignar y gestionar las miles de peticiones, cruzadas entre los múltiples centros, usando en cada caso los canales y recursos apropiados (ver fig:3.13). La segunda es la implementación de sistemas de almacenamiento de alto rendimiento, capaces de mantener una entrada y salida de datos de varios Gbps sostenidos de forma indefinida en el tiempo (ver fig:3.14). PhEDEx y FTS son la solución elegida por CMS para la primera de ellas.

PhEDEx ha sido diseñado y probado a escala más allá de las necesidades actuales de CMS. La robustez se aplica tanto en la detección, como a la recuperación de errores locales en un entorno distribuido, proporcionando a los administradores del Tier y a los usuarios de la colaboración, una vista en tiempo real del estado global de transferencia de datos CMS mediante una interfaz web[69]. Se encuentra operativo desde 2004, y actualmente se ejecuta la versión 4.2.1. Establece un control total sobre el estado de la réplica en transferencia. Es capaz de comunicarse con FTS (ver 3.2.3) y acceder a la información sobre el estado de los canales de tráfico de datos, asignado pesos a cada canal según su estado y su fiabilidad (ver tab:3.2).

PhEDEx se compone de una serie de procesos o agentes autónomos, robustos, persistentes y sin estado. Estos agentes comparten el estado de transferencia de la réplica con los agentes centrales (ver fig:3.15), a través de la TMDB central

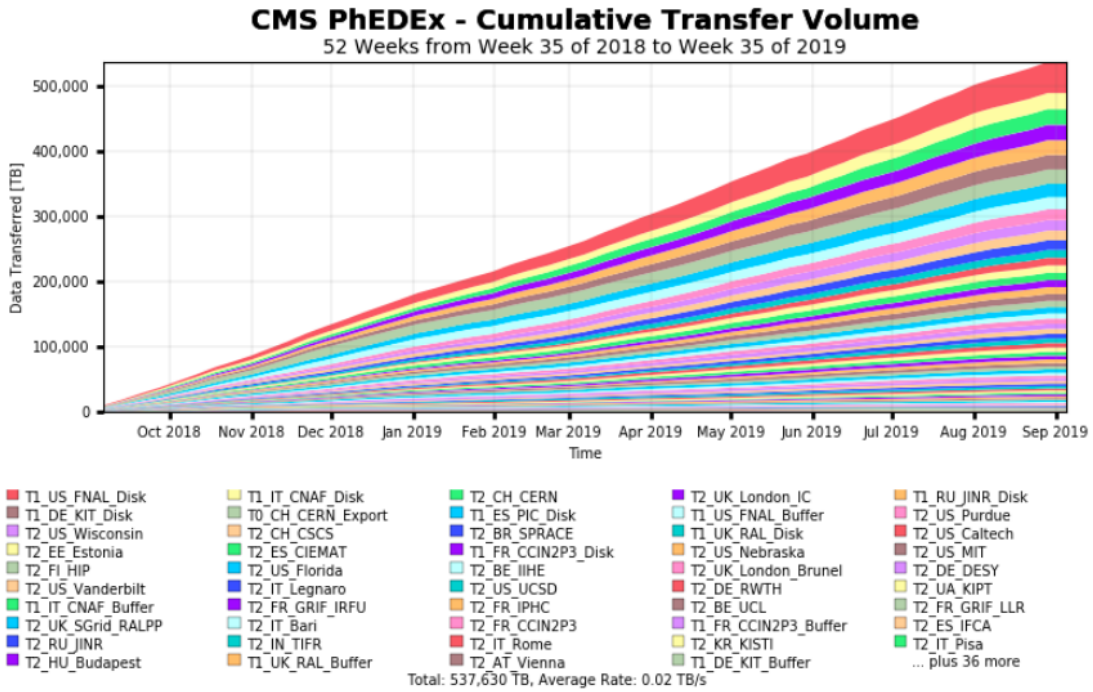


FIGURA 3.13: Velocidad media de Transferencias entre los centros de CMS en el último año.

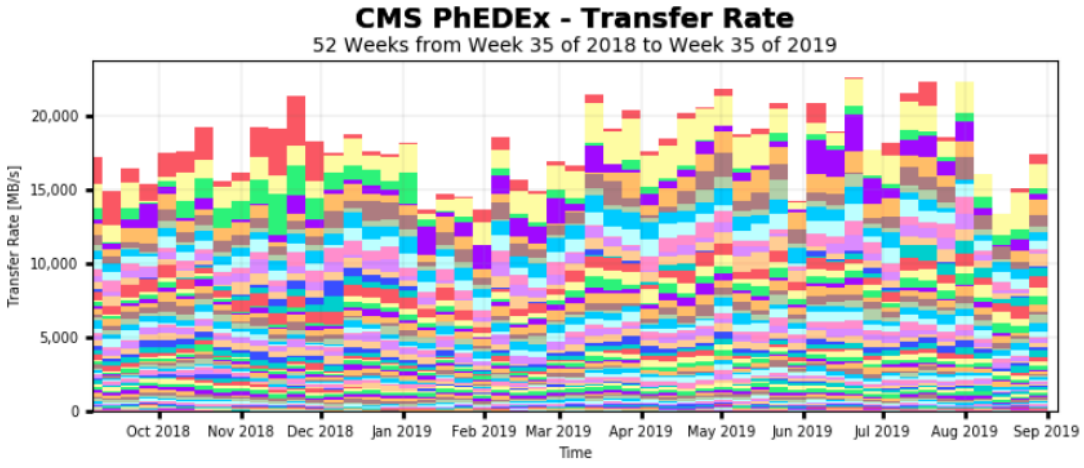


FIGURA 3.14: Volumen acumulado de las transferencias entre los centros de CMS en el último año.

localizada en el Cern. Los agentes no mantienen el estado localmente y gracias a ello pueden detenerse o iniciarse sin consecuencias negativas para las réplicas en curso, incluso después de un bloqueo o reinicio del sistema, gracias a que el

TABLA 3.2: Pesos asignados a los diferentes canales en el Tier-2 IFCA para establecer transferencias.

```
set status to inxfer for task=368315475
set status to inxfer for task=368315630
set status to inxfer for task=368315632
fetched 1 new tasks for link T2_EE_Estonia -> T2_ES_IFCA
fetched 1 new tasks for link T2_ES_CIEMAT -> T2_ES_IFCA
fetched 1 new tasks for link T2_KR_KISTI -> T2_ES_IFCA
fetched 1 new tasks for link T2_PL_Swierk -> T2_ES_IFCA
fetched 1 new tasks for link T2_UA_KIPT -> T2_ES_IFCA
fetched 3 new tasks for link T2_US_Vanderbilt -> T2_ES_IFCA
balancing transfers on 6 links
Transfer::FTS3::isBusy Link Stats T2_US_Vanderbilt -> T2_ES_IFCA:
Transfer::FTS3::isBusy Link Stats T2_UA_KIPT -> T2_ES_IFCA:
Transfer::FTS3::isBusy Link Stats T2_KR_KISTI -> T2_ES_IFCA:
Transfer::FTS3::isBusy Link Stats T2_EE_Estonia -> T2_ES_IFCA:
Transfer::FTS3::isBusy Link Stats T2_PL_Swierk -> T2_ES_IFCA:
Transfer::FTS3::isBusy Link Stats T2_ES_CIEMAT -> T2_ES_IFCA:
T2_EE_Estonia->T2_ES_IFCA:P=[0.000,0.195),W=1.000,USED=0,DONE=0,ERRORS=0
T2_ES_CIEMAT->T2_ES_IFCA:P=[0.195,0.391),W=1.000,USED=0,DONE=0,ERRORS=0
T2_KR_KISTI->T2_ES_IFCA:P=[0.391,0.586),W=1.000,USED=0,DONE=0,ERRORS=0
T2_PL_Swierk->T2_ES_IFCA:P=[0.586,0.781),W=1.000,USED=0,DONE=0,ERRORS=0
T2_UA_KIPT->T2_ES_IFCA:P=[0.781,0.977),W=1.000,USED=0, DONE=0,ERRORS=0
T2_US_Vanderbilt ->T2_ES_IFCA:P=[0.977,1.000),W=0.120,USED=78,DONE=0,ERRORS=3
```

estado de flujo del trabajo se almacena en la TMDB.

Cada agente, encuentra trabajo pendiente en la TMDB, selecciona y prioriza tareas y las ejecuta; finalmente marca las tareas exitosas completadas, asignando tareas indirectamente a otros agentes en el proceso. Las operaciones SQL, son incrustadas literalmente en el código del agente, que son utilizadas para toda la comunicación con la TMDB. Las ganancias en robustez, disponibilidad y flexibilidad del sistema han superado en gran medida las desventajas.

El servicio PhEDEx ejecuta varios agentes:

- **FileDownload**: Es el agente principal. Inicia transferencias desde un sitio remoto al elemento de almacenamiento local.
- **FileExport**: Inicia transferencias desde un almacenamiento local a un sitio remoto.
- **FileRemove**: Gestiona las órdenes de borrados de datos.
- **BlockDownloadVerify**: Gestiona la verificación de la transferencia correcta de un bloque de datos.
- **FileStager**: Controla el acceso a cinta. Solo requerido en Tier-1 y Tier-0.

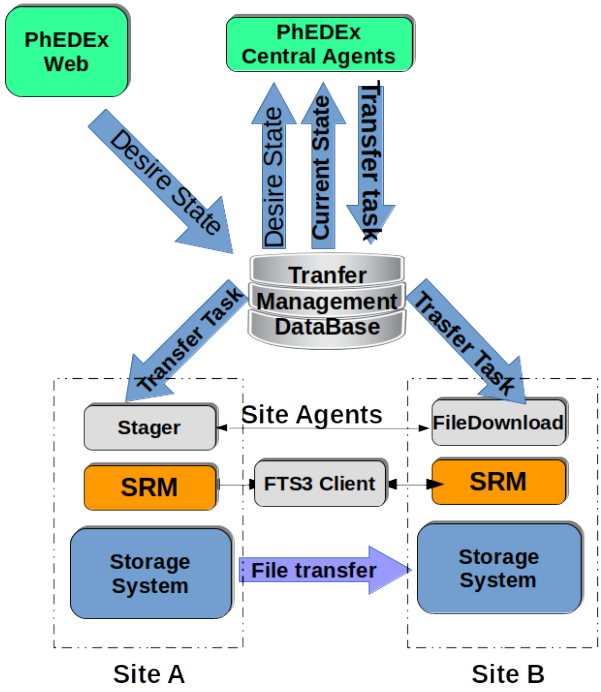


FIGURA 3.15: Componentes de PhEDx.

- **Watchdog**: Supervisa que el resto de agentes se estén ejecutando.

No todos los centros Tier requieren la instalación de todos los agentes. El "FileStager" es solo requerido en los Tier-1 y Tier-0, que son aquellos a los que se les requiere la disponibilidad de un sistema de cintas como almacenamiento secundario "nearline".

Actualmente los agentes de PhEDEx, pueden ejecutarse contra tres instancias de manera simultanea e independiente. Gracias a la asignación de roles, de la que hemos hablado anteriormente (ver 1.2.1), todas las instancias son independientes entre si, y no realizan ningún cambio de información entre ellas. La instancias que emplea PhEDEx son:

- **Production**: Los datos producidos y archivados por el experimento, utilizan esta instancia para su distribución.

- **Debug**: Es la instancia crítica utilizada para la puesta en marcha y operación del Tier. Los enlaces de transferencia de sitios de Tier-1 y Tier-2 se pueden

deshabilitar si los problemas de transferencia en esta instancia no son resueltos. El enfoque principal de esta instancia es depurar los agentes y enlaces de transferencia entre los diferentes Tiers.

- **Development**: Es empleada por norma general durante lanzamientos de una versión de PhEDEx con cambios significativos, "mayor release change". Con ella se testean los posibles errores de código o de funcionalidad de la nueva versión, sin interferir en las instancias operativas.

PhEDEx ha evolucionados durante los últimos 15 años, proporcionando a los administradores nuevas funcionalidades, como la consistencia de datos, la monitorización de espacio[59], y facilitando de esta forma la gestión de la transferencia de datos.

Durante estos años, el Tier-2 del IFCA, ha participado activamente en los procesos de verificación de las nuevas versiones como "Early Adopters", aportando código para las utilidades de consistencia de datos en sistemas de almacenamiento para sistemas de ficheros posix (ver apendice 01), que actualmente sigue en producción.

3.3. Calidad de Datos

CMS debe garantizar que el hardware y el software funcionan correctamente, para ello es necesario comprobar la calidad de los datos que se han adquirido desde el detector. Dentro de CMS este proceso lo lleva a cabo el grupo responsable de analizar la calidad de los datos, llamado Physics Performance & Dataset (PPD).

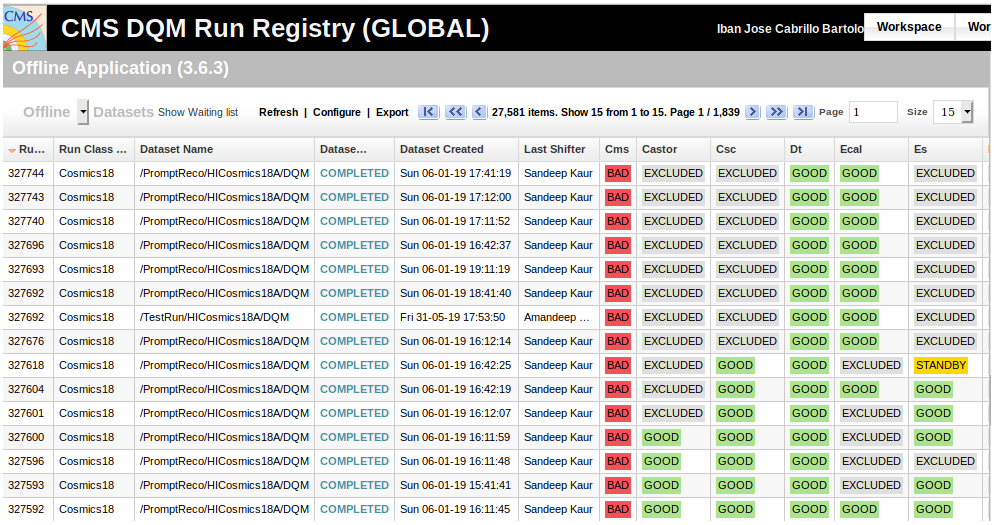
El grupo PPD trabaja con el grupo del detector y el Physics Object Group (POG) en las siguientes áreas:

- Calidad de Datos y certificación.
- Alineamiento y validación.
- Software y validación.
- Gestión y producción de los eventos de MC para validación.

- Organización y configuración de los Datasets y del procesado de datos.

El PPD tiene que garantizar la calidad de los datos que se proporcionan a los grupos de análisis de física. El análisis de la calidad de los datos se realiza en dos fases: al registrarse los datos durante las colisiones y una vez que los datos se han procesado y almacenado.

De los datos que el sistema TriDAS envía al Tier-0, un subgrupo de los datos RAW es reconstruido rápidamente y enviado al sistema Data Quality Monitoring (DQM) "Online" en la sala de control de CMS donde son analizados por el grupo PPD. Si se detecta un problema, los datos se marcan como "malos para el análisis", quedando este estado almacenado en la herramienta de Registro de ejecución (RR) (ver fig:3.16). El equipo intentará comprender las posibles causas de problema y dependiendo del origen del fallo, intentará recuperarlos.



Ru...	Run Class ...	Dataset Name	Datase...	Dataset Created	Last Shifter	Cms	Castor	Csc	Dt	Ecal	Es
327744	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 17:41:19	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327743	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 17:12:00	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327740	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 17:11:52	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327696	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:42:37	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327693	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 19:11:19	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327692	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 18:41:40	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327692	Cosmics18	/TestRun/HICosmics18A/DQM	COMPLETED	Fri 31-05-19 17:53:50	Amandeep ...	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327676	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:12:14	Sandeep Kaur	BAD	EXCLUDED	EXCLUDED	GOOD	GOOD	EXCLUDED
327618	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:42:25	Sandeep Kaur	BAD	EXCLUDED	GOOD	GOOD	EXCLUDED	STANDBY
327604	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:42:19	Sandeep Kaur	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD
327601	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:12:07	Sandeep Kaur	BAD	EXCLUDED	GOOD	GOOD	EXCLUDED	GOOD
327600	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:11:59	Sandeep Kaur	BAD	GOOD	GOOD	GOOD	EXCLUDED	GOOD
327596	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:11:48	Sandeep Kaur	BAD	GOOD	GOOD	GOOD	EXCLUDED	EXCLUDED
327593	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 15:41:41	Sandeep Kaur	BAD	GOOD	GOOD	GOOD	EXCLUDED	GOOD
327592	Cosmics18	/PromptReco/HICosmics18A/DQM	COMPLETED	Sun 06-01-19 16:11:45	Sandeep Kaur	BAD	GOOD	GOOD	GOOD	GOOD	GOOD

FIGURA 3.16: Interfaz Gráfica de la herramienta de Registro de ejecuciones (RR).

3.3.1. DQM

Todos estos flujos de datos son monitorizados por DQM, que es un sistema empleado para producir los "Plots" durante la ejecución de reconstrucciones (RECO),o cualquier otro flujo dentro del CMS Software (CMSSW). Sus funciones principales son:

- **Online**: Toma una muestra "plots" de eventos después del HLT con muy baja latencia. De esta forma es capaz de monitorizar el rendimiento del detector durante la toma de datos. Disparan histogramas a una velocidad de aproximadamente 1015Hz y ejecuta su elección de algoritmos y módulos de análisis que generan resultados en forma de Elementos de Monitorización (EM), incluidos histogramas de referencia y resultados de pruebas de calidad.

- **Offline**: Lee todos los eventos mientras son reconstruidos. Estos eventos son usados para certificar los datos y validar la versión de software.

El DQMGUI (ver fig:3.17) es una interfaz web donde los "shifters" acceden a los histogramas de las ejecuciones en curso.

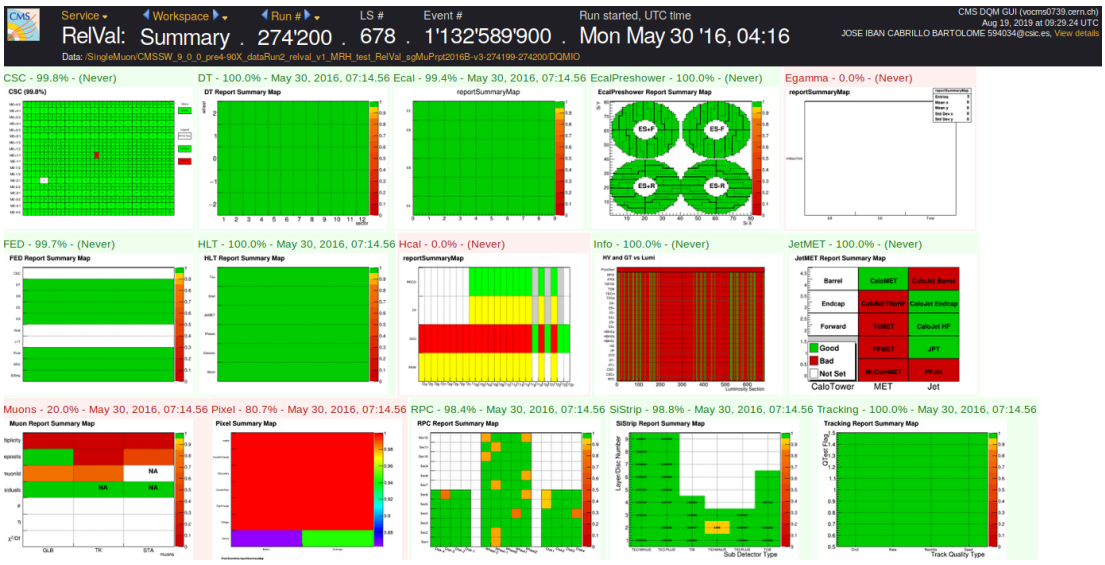


FIGURA 3.17: Interfaz Gráfica de usuarios para la monitorización de DQM.

Dentro de DQM (ver fig:3.18), están involucrados diferentes grupos:

- El Detector Performance Groups (DPG), que supervisa el estado y el comportamiento de cada sub-sistema hasta la reconstrucción local. DPG DQM es ejecutado en las etapas "Online" y "Offline".

- El POG, que supervisa la calidad de los objetos físicos reconstruidos (muones, "jets", electrones, ...), POG DQM se ejecuta solo en la etapa "Offline".

- El Physics Analisys Group (PAG), que monitoriza niveles de objetos más altos (Picos de masa de dileptones), correlacionando diferentes objetos físicos o distribuciones cinemáticas simples con cortes más orientados al análisis, PAG DQM se ejecuta solo en la etapa "Offline".

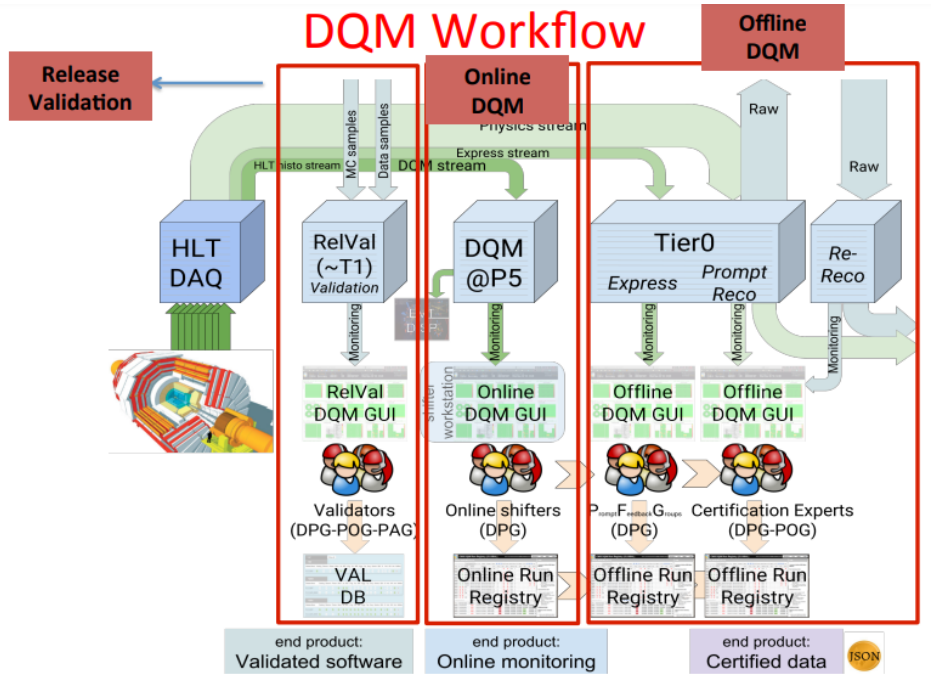


FIGURA 3.18: Flujo de trabajo del DQM.
[70]

3.3.2. Certificación de Datos

El sistema de adquisición de datos es complejo, y aunque hasta ahora ha sido tremendamente estable, es necesario ejecutar diferentes tareas de monitorización continua de los diferentes sistemas que lo componen para asegurar la calidad de los datos. Para ello, el equipo que gestiona el detector y expertos del POG, comprueban los "plots" generados por DQM (ver 3.3.1) de cada ejecución, seleccionando aquellos que son aptos para el análisis. A través de DQM, buscan distorsiones, picos de ruido, regiones sin datos del detector o calibraciones erróneas, que puedan afectar a la calidad del análisis.

El grupo de certificación, es el encargado de generar el "Golden JSON", seleccionando aquellos sucesos donde el estado del detector se califica como "Golden", es decir, aquellas que son válidas para el análisis. Este fichero puede ser usado mediante CRAB[71], de manera que solo tendrá en cuenta aquellos sucesos que hayan satisfecho los criterios de calidad.

Las actividades del equipo de certificación son:

- Proporcionar la lista de ejecuciones que deben certificarse.
- Brindar ayuda a los expertos en certificación DPG/POG.
- Producir los archivos JSON que incluyen los sucesos y las secciones de luminosidad para el análisis de la física.
- Actualizar los requerimientos que definen la calidad.
- Informar a la colaboración de CMS sobre los archivos "Golden JSON" oficiales.
- Mantener la información relevante en el Registro de ejecución (RR).

3.3.3. Validación de las versiones de Software

CMSSW es el software empleado por CMS, para realizar todo tipo de tareas; generación, simulación, reconstrucción y flujos de análisis. Está desarrollado en C++.

Un trabajo se compone de una serie de algoritmos que son procesados en un determinado orden. Los algoritmos se comunican entre si a través de los datos almacenados en el suceso. Sólo existe un ejecutable y varios módulos "plug-in" que ejecutan los algoritmos. El archivo binario de CMSSW se configura en el momento de la ejecución mediante un fichero de configuración, que es proporcionado por el usuario. Este fichero usa sintaxis de un lenguaje de alto nivel como python, y proporciona al ejecutable los datos que debe usar, los módulos a ejecutar y establecerá el orden determinado de procesamiento de cada uno de ellos. Estos módulos, son cargados dinámicamente al comienzo de la ejecución del trabajo. CMSSW es actualizado de forma regular, cada 6 meses aproximadamente, una vez

validada la nueva versión. CMSSW es un software crítico, ya que es el responsable de la reconstrucción de los datos. Un funcionamiento incorrecto de CMSSW implicaría una reconstrucción incorrecta de los datos.

La integridad de la nueva versión se vincula mediante pruebas de calidad supervisadas por el DQM. Pequeñas pruebas de producción (muestras RelVal), son realizadas en cada nueva "pre-release". En cada una de ellas los "plots" son generados y evaluados, identificando problemas por parte del software, de las calibraciones empleadas o de los propios algoritmos de reconstrucción, pudiendo volver a versiones anteriores si la evaluación no es correcta. Este ciclo continúa indefinidamente hasta que se da luz verde a la "Major Release" (ver fig:3.19). Todas las iteraciones quedan almacenadas en una DB de validación. Una vez que la nueva versión esta lista, se prepara campaña de re-reconstrucción o producción de MC, mediante la cual se comprueban y validan las condiciones de alineación y calibración.

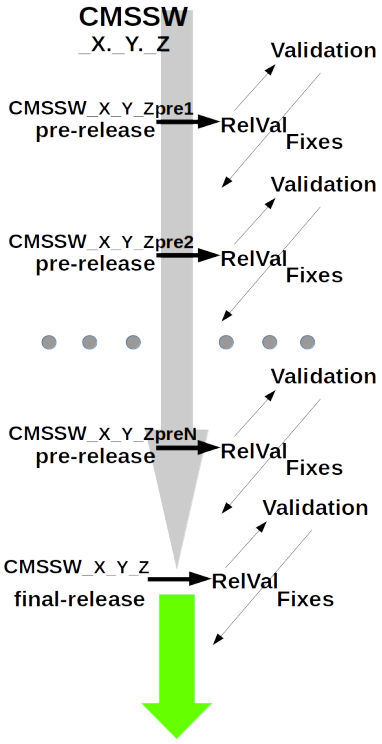


FIGURA 3.19: Proceso de validación de las nuevas versiones de CMSSW.

3.4. Integración de una infraestructura HPC en el entorno de Datos de CMS

En su búsqueda incesante de recursos, las grandes colaboraciones siempre han deseado acceder a los grandes centros de computación HPC, de los diferentes países que participan en los experimentos. Si nos situamos desde el punto de vista técnico se nos presentan diversas dificultades, entre ellas:

- Forma de acceso: Grid vs Local Clúster.
- Seguridad: Acceso a los diferentes servicios distribuidos de la colaboración.
- Acceso a Datos: Datos distribuidos frente a Datos locales.

Durante el año 2012, los investigadores de CMS del IFCA y la Universidad de Oviedo estaban utilizando los recursos Grid del Tier-2 del IFCA para trabajos de skimming y análisis de datos finales en canales leptónicos, incluyendo WW, Higgs y estudios superiores. A pesar del importante volumen de recursos disponibles en el Tier-2 local, la ejecución de múltiples trabajos de skimming requería acceso simultáneo a los datos. Esta ejecución masiva de trabajos generaba alta latencias en el acceso al sistema de ficheros, debido a la alta demandas de I/O entre el Tier-2 y el sistema de archivos GPFS. Las tareas eran más lentas de lo esperado y los investigadores tenían necesidades de entregar sus resultados a tiempo a la colaboración CMS, y ultimar las ponencias a presentar en las conferencias de verano/invierno. No podían permitirse un retraso de un mes o más en la producción de n-tuplas utilizadas en la etapa final del análisis de física.

Al mismo tiempo, la Universidad de Cantabria (UC) adquirió una nueva infraestructura HPC, que se denominó Altamira, orientada a la investigación básica y aplicada, diseñada para soportar eficientemente grandes procesamientos de datos mediante una red InfiniBand con soporte Remote Direct Memory Access (RDMA).

Debido a la experiencia del grupo de computación del IFCA en este tipo de tecnologías, la UC decidió que el nodo fuese instalado en las instalaciones del IFCA y operado por su grupo de computación.

A continuación, describiremos como se realizó una instalación eficiente de esta gran infraestructura informática de alto rendimiento y su integración como recurso oportunístico para la colaboración CMS, tratando de minimizar la gestión de ambos sistemas (Tier-2 y nodo HPC), por parte de administradores y con una curva de aprendizaje insignificante para estos usuarios finales.

3.4.1. Descripción del nodo HPC Altamira

Altamira es un supercomputador de propósito general. Entre sus usuarios se incluyen investigadores de proyectos nacionales e internacionales de la UC, investigadores que solicitan acceso a recursos computacionales a través de la RES y empresas que desarrollan proyectos de interés científico y que necesitan recurrir a instalaciones Instalación Científico Técnica Singular (ICTS) para llevar a cabo sus objetivos.

Está formado por un clúster basado en servidores Intel utilizando red InfiniBand FDR [72] para su interconexión. Altamira fue diseñado por el IFCA en colaboración con IBM y el Barcelona Supercomputer Center (BSC). Incluye 240 nodos iDataplex dx360m4, cada uno con dos procesadores Intel SandyBridge E5-2670 2.6 Hz/1600 20MB de caché Intel, 64GB de memoria RAM y Disco duro SATA II de 500GB. Integra además siete nodos iDataplex dx360m3 con dos GPU nVidia Tesla M2090 cada uno, y once servidores ps702 IBM Power7, necesarios para mantener la compatibilidad con el nodo de supercomputación anterior, basado en servidores JS20. El sistema completo instalado en el IFCA, se muestra en la figura 3.20.

El despliegue de InfiniBand FDR Mellanox, permite una latencia muy baja entre nodos, menos de 1 microsegundo, y un ancho de banda muy elevado, 40Gbps. Está configurado en una topología denominada “FAT Tree”, que proporciona una arquitectura de red sin bloqueo. InfiniBand utiliza el mecanismo RDMA[73], lo que permite que el adaptador de red transfiera datos directamente desde o hacia la memoria de la aplicación, eliminando así la necesidad de copiar datos entre aplicaciones, usando memorias intermedias ó la memoria del sistema operativo. Proporciona un ancho de banda muy elevado en operaciones de acceso a datos. Gracias a esta configuración, Altamira alcanzó más de 80 TFlops, entrando en la lista del Top500 en junio de 2012 en la posición 358. Más tarde, en 2012, el clúster

FIGURA 3.20: Imagen del computador HPC Altamira.

de Altamira se reconfiguró y el equipo actualmente alojado en IFCA incluye 158 nodos dx360m4 y 5 dx360m3, con una potencia superior a 50 TFlops. Este sistema era accedido desde el exterior mediante el protocolo ssh, conectando a un nodo de inicio de sesión, proporcionando usuario/contraseña y enviando trabajos directamente al planificador de recursos, Slurm (ver 2.4.1.2).

3.4.2. Integración HPC-Grid

Como mencionamos anteriormente, el CPD del IFCA aloja los sistemas Grid del Tier-2 y el supercomputador Altamira, pero su gestión era independiente. Por un lado, la información requerida por los investigadores de CMS como entrada para el análisis se almacena bajo el sistema de ficheros GPFS en Tier-2 de CMS. Por otro, los nodos de Altamira son capaces de manejar en paralelo una gran cantidad de trabajos (hasta 2.500) con necesidades muy exigentes en el acceso a datos, gracias a la disponibilidad de acceso a red InfiniBand en cada uno de los nodo. La integración requería realizar algunos cambios en el diseño original planificado para el clúster HPC. Estos cambios no habían sido probados en otra infraestructura de estas características, pero en teoría eran viables.

Integrar el sistema de autenticación para ambos sistemas utilizando el mismo servicio LDAP[74], proporciona un área de inicio común tanto para los usuarios

locales del Tier-2, como para los del clúster HPC, posibilitando la gestión de los datos, accesibles bajo una única ruta al sistema de ficheros. GPFS necesita siempre conectividad IP para establecer la integridad del clúster.

Usualmente en este tipo de instalaciones HPC, se establece una red IP, sobre la capa de red InfiniBand, dejando el clúster HPC aislado. Para llevar acabo esta integración, se requirió el despliegue de red Ethernet en los nodos de Altamira y establecer un único clúster GPFS extendido con el Tier-2, para poder acceder así a los datos, mientras que la red InfiniBand de alto rendimiento seguiría usandose en los nodos Altamira y la red Ethernet en los nodos del Tier-2. Esta discriminación en el uso de red es gestionada por los controladores de GPFS. Una descripción general de este hardware y la integración se describe en el esquema que se muestra en la figura 3.21.

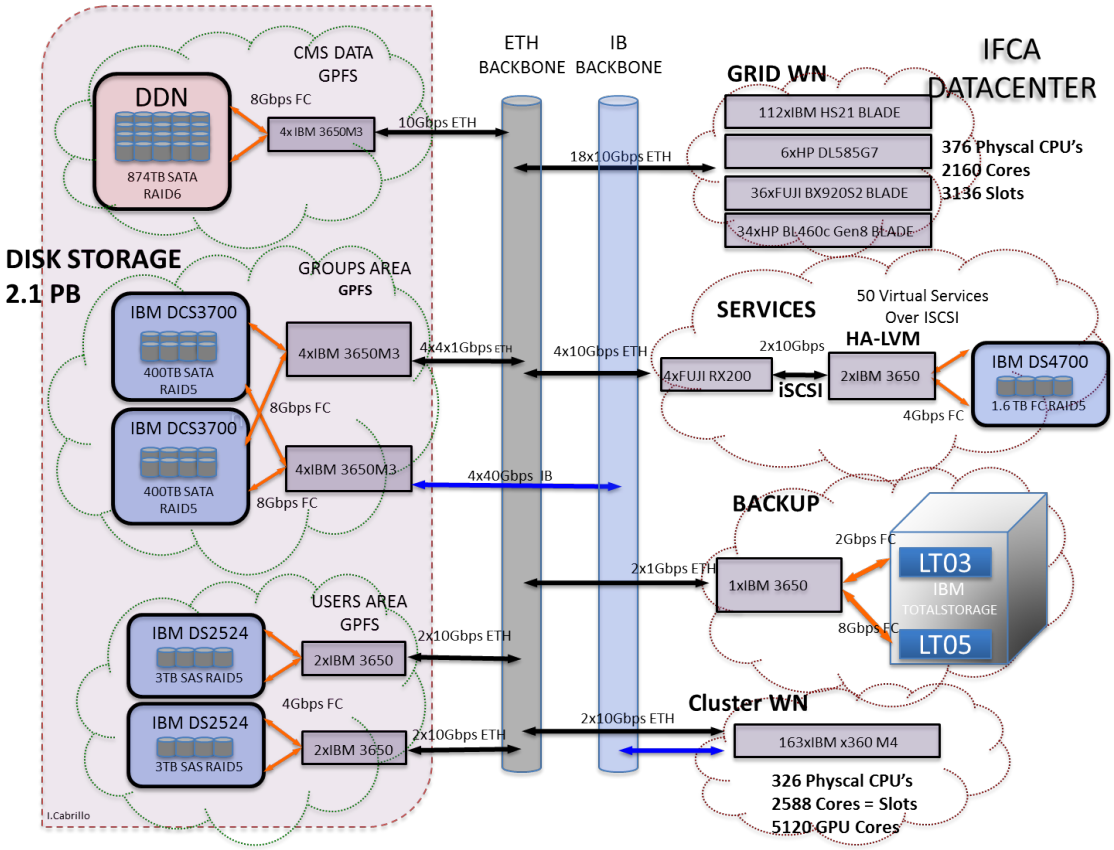


FIGURA 3.21: Esquema de integración del Hardware del CPD del IFCA año 2013.

Todos los nodos de Altamira se conectaron a la capa Ethernet del IFCA utilizando una interfaz de red de 1Gbps, desde cada nodo a la parte superior del "switch" Ethernet instalado en cada armario. Este ancho de banda, es suficiente para establecer la coherencia del clúster. El acceso a la red troncal del IFCA se realiza mediante un enlace redundado de 10Gbps. Esta conexión de red se utiliza principalmente para mantener la integridad en el clúster GPFS, pero descubrimos una segunda utilidad: funciona como red secundaria, en forma de red redundada. Si la red InfiniBand no esta disponible, los nodos de Altamira usarán Ethernet para la transferencia de datos, disminuyendo sus prestaciones, pero manteniendo la integridad. El esquema de conectividad se representa en la figura 3.22.

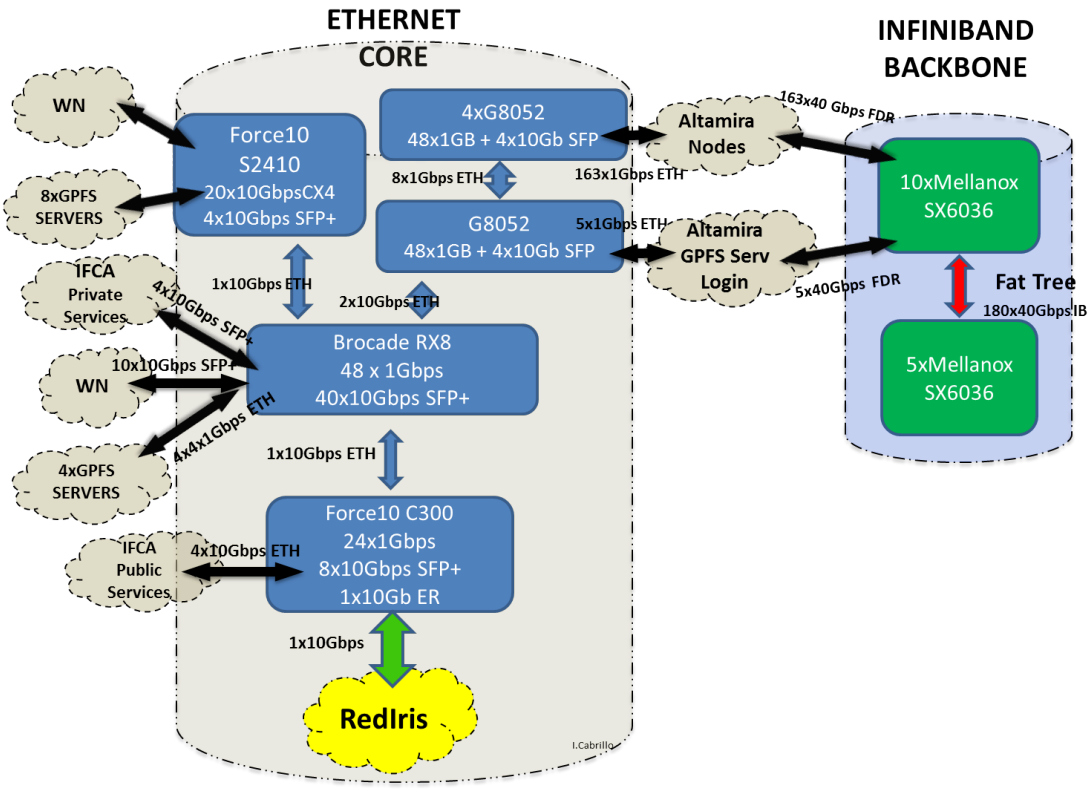


FIGURA 3.22: Esquema de conectividad de red del CPD del IFCA año 2013.

Cuatro servidores GPFS gestionaban el almacenamiento de Altamira, conectados directamente por fibra redundada de 8Gbps a cabinas de almacenamiento IBM DCS3700 con una capacidad bruta de 1 Petabyte, con conexiones Infiniband

TABLA 3.3: Activación de la interfaz RDMA en el módulo GPFS.

```
#mmchconfig verbsrdma=enable,verbsPort="mlx4_0" –N "node1,node2,...,nodeN"
 ..................................................................
Loading modules from /lib/modules/2.6.32-358.14.1.el6.x86_64/extra
Module                  Size  Used by
mmfs26                1762439  0
mmfslinux              310536  1 mmfs26
tracedev                29456  2 mmfs26,mmfslinux
Wed Sep 25 13:11:26.505 2013: GPFS: 6027-310 mmfsd initializing.
Wed Sep 25 13:11:28.437 2013: VERBS RDMA starting.
Wed Sep 25 13:11:28.438 2013: VERBS RDMA library libibverbs.so initialized.
Wed Sep 25 13:11:28.811 2013: VERBS RDMA device mlx4_0 port 1 opened.
Wed Sep 25 13:11:28.812 2013: VERBS RDMA started.
```

FDR 40Gbps a todos los nodos de Altamira y conexiones 1Gbps a la red Ethernet, suficiente si tenemos en cuenta que en el esquema no se esperaban lecturas desde los nodos Ethernet al almacenamiento del clúster HPC.

Las mejoras que implica el uso de RDMA, solo son accesibles en versiones superiores a GPFS 3.4. Todos los nodos del CPD se actualizaron a la versión 3.5.0.2, para llevar a cabo este diseño.

Una vez que la red Infiniband está desplegada en los nodos (debe verificarse mediante los comandos "ibstat" e "ibstatus"), el dispositivo abierto debe capturarse para que el módulo de GPFS establezca la interfaz que debe utilizar para la transferencia de datos.

Para utilizar el soporte nativo de InfiniBand con GPFS en lugar de IP, el módulo verbsRdma debe estar activado, indicando el dispositivo Infiniband "verbsPort" que será empleado para la transferencia de datos. Es necesario un reinicio del demonio GPFS en los nodos Infiniband *(node1,..., nodeN)* para que activen el módulo "VERBS RDMA" (ver tab:3.3).

El uso de las mejoras RDMA para GPFS sobre Infiniband, da como resultado un acceso de datos de mayor rendimiento en comparación con el acceso sobre Ethernet. Los resultados obtenidos mediante la herramienta GpfsPerf, se muestran en la tabla 3.4. Como puede verse, indican un factor 3 aproximado de mejora en los accesos desde nodos cliente InfiniBand, con respecto a los nodos cliente que usan Ethernet.

Finalmente, para permitir que el clúster Altamira ejecute trabajos CMS, era necesario que estos nodos tuvieran acceso al entorno de software CMS. Esto fue posible usando el sistema de archivos CVMFS [45] accesible en los nodos mediante

Tabla 3.4: Resultados de los test Gpfsperf para Servidores y Clientes con diferentes Interfaces de red.

	CREATE	READ	WRITE	READ 8TH	READ 16TH	READ 32TH
`Eth Server`	1200MBps	920MBps	1200MBps	925MBps	930MBps	925MBps
`Eth Client`	460MBps	300MBps	455MBps	308MBps	315MBps	305MBps
`IB Server`	1700MBps	1500MBps	1850MBps	1170MBps	1420MBps	1370MBps
`IB Client`	1600MBps	2290MBps	1600MBps	1132MBps	1135MBps	1135MBps

dos servidores proxy, que generan un sistema de ficheros cacheado, al igual que en los nodos de computo del Tier-2. Este sistema de ficheros es un sistema de solo lectura. La instalación del software se realiza de forma centralizada por el experimento (CMS en este caso particular), evitando modificaciones no deseadas y accesos recurrentes a su infraestructura.

Las colas de Altamira están limitadas a 72h de ejecución para un trabajo, y estos trabajos generalmente usan entre 32 y 512 núcleos. Los trabajos cortos (menos de 6h) tienen prioridad para optimizar el llenado de nodos. La mejor manera de explotar esta configuración para los trabajos de skimming de CMS, era enviar trabajos desde un solo script, solicitando más de 150 núcleos en modo paralelo, aprovechando el acceso a datos de Altamira a través de Infiniband.

Los cambios introducidos tanto software como hardware se muestran en la figura 3.23). Gracias a ellos, conseguimos solventar los problemas de integración que habíamos planteado al inicio de esta sección.

- Unificar el sistema de acceso externo para ambos sistemas mediante LDAP.
- Establecer un misma red IP para todos los nodos, permite el acceso a todo el sistema de ficheros distribuido Tier-2 $\Longleftrightarrow$ Altamira, manteniendo las ventajas I/O e IOPS de la red InfiniBand en los nodos HPC.
- La instalación de CVMFS en los nodos Altamira nos permite acceder al área de software de CMS.

El acceso local al clúster, a los datos y al área de software de la colaboración nos permite acceder al sistema de colas local, Slurm y el envío de trabajos con garantía para su correcta ejecución (wrapping 150 jobs).

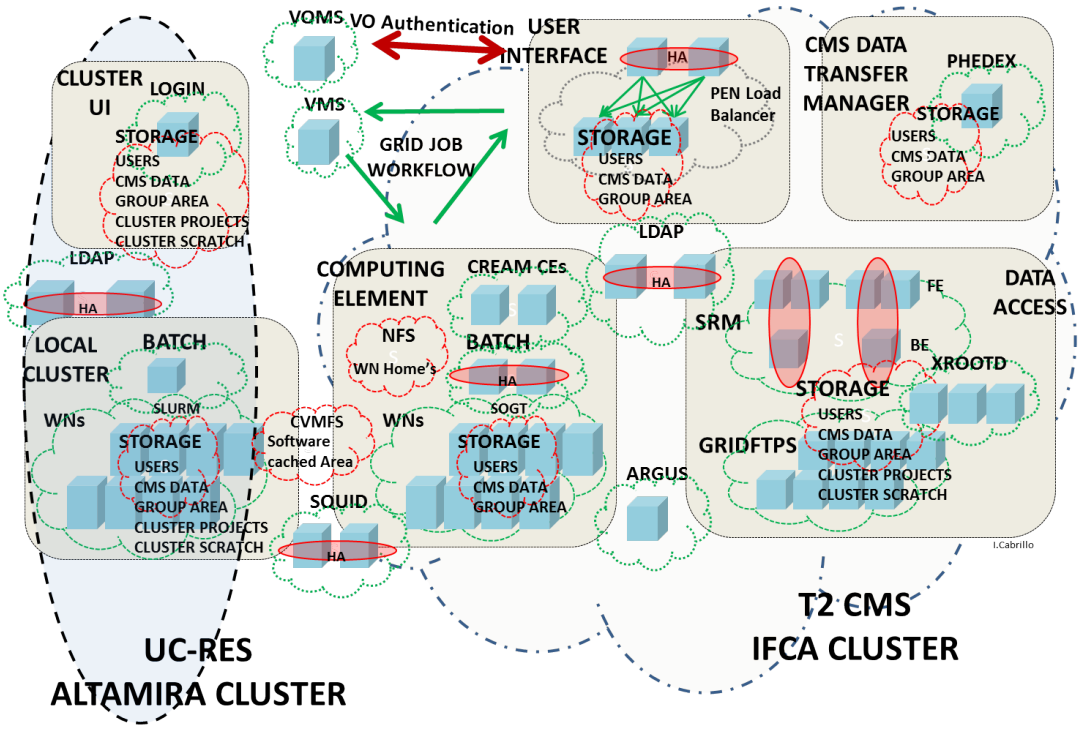


FIGURA 3.23: Esquema de servicios del CPD del IFCA después de la integración.

3.4.3. Resultados cuantificables de la Integración

En abril de 2013 se realizó una prueba de aplicación real de la nueva infraestructura, aprovechando los recursos proporcionados por Altamira, para una búsqueda dileptónica de SUSY mediante un análisis desarrollado por el grupo de física IFCA-Oviedo. Comprendía tareas de skimming sobre el conjunto completo de datos de CMS 2012, muestras de MC, y la producción de "Root Tree", tomando como entrada la salida del mencionado skimming (ver fig:3.24).

La producción requirió un total de 250.000 horas de CPU, y se procesaron más de 2.000 millones de eventos, con un tiempo promedio de CPU por evento de 6.3 milisegundos y un "walltime" adicional por debajo de 0.4 milisegundos. En total, se leyeron más de 23TB de muestras de datos de entrada: 18TB a través de la red Ethernet y aproximadamente 5,6TB con la red Infiniband. En términos de producción, más de 8TB de datos se escribieron en el disco usando la red Infiniband: 5.6TB en el paso de skimming (formato de datos CMSSW) y 2.5TB

en Formato de "Root Tree" para el análisis del usuario final.

El tiempo de procesamiento estimado para este análisis, con los recursos del Tier-2 fue de alrededor de dos meses; la producción con Altamira se realizó en una semana. Además, la saturación en el sistema StoRM (ver 2.3.5) del Tier-2, que podría esperarse en miles de trabajos de Grid, con un flujo aproximado de 8TB de datos de salida, fue evitado.

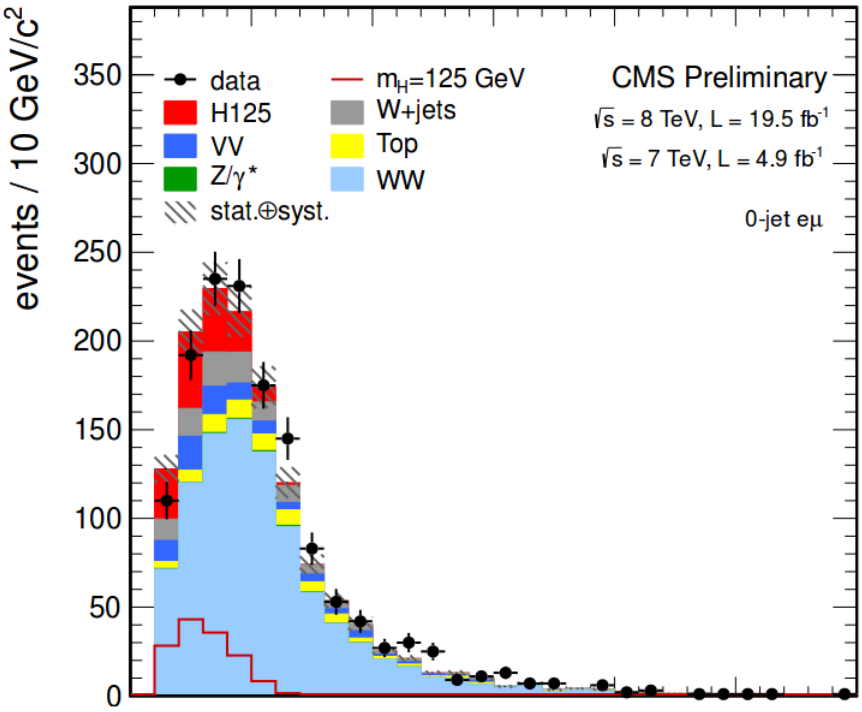


FIGURA 3.24: Distribución de la búsqueda del Bosón de Higgs en el canal a dos bosones WW.

La computación HPC demostró ser un sistema ideal para el procesamiento de grandes datos. Permite una potencia extra eficiente para análisis en períodos de grandes cargas de trabajo. La integración entre los recursos IFCA Tier-2 y el supercomputador Altamira fue posible gracias a servicios como LDAP y GPFS, que posibilitan la integración de ambos sistemas. Más de 500.000 horas se utilizaron durante 2013 por investigadores que trabajan en análisis de CMS, relacionados tanto con Higgs como con búsquedas de SUSY, y medidas de sección transversal de modelo estándar en canales de dileptón. El rendimiento medido para GPFS

sobre Infiniband mostró muy buena velocidad de transferencia de datos a cualquier nodo, evitando tiempo muertos en la CPU, aumentando la eficiencia del trabajo y minimizando la tasa de fallos, pasando de un 20 % en el modelo GRID a un 2 % o menos en HPC. El software específico de la comunidad de investigación (para la colaboración de CMS en este caso) se instaló con un bajo esfuerzo, usando CVMFS. Esta experiencia se puede extrapolar a otras áreas. El impacto en los investigadores fue mínimo, solo tuvieron que introducir modificaciones menores en sus scripts de envío de trabajos. El tiempo de ejecución se reduce en un orden de magnitud, de meses a semanas, o de semanas a días. Este desarrollo fue presentado en CHEP2013: 20. International Conference on Computing in High Energy and Nuclear Physics; Amsterdam (Holanda); titulado "**Direct exploitation of a top 500 Supercomputer for Analysis of CMS Data**" que dio lugar a un artículo en la revista "Journal of Physics. Conference Series (Online)"; ISSN 1742-6596; v. 513(3); [7 p.] por Cabrillo et al.[44].

3.4.4. Consolidación de la Integración

Si bien es verdad que la implementación anterior ha tenido innegables beneficios de cara a los usuarios, reduciendo la tasa de errores frente al modelo Grid y el tiempo de ejecución en un orden de magnitud como hemos visto anteriormente, no es posible contabilizar esta cantidad ingente de recursos (500.000 horas de CPU), que se proporcionan de forma local, frente a la colaboración. Valores importantes a la hora de realizar justificaciones de uso, necesarias ante la petición de nuevos proyectos.

La infraestructura ha evolucionado en estos años, pero se mantiene el diseño de integración anteriormente expuesto (ver fig:3.25).

Para solucionar el problema de la contabilidad anteriormente mencionado, hemos propuesto la integración de dos nuevos componentes en este esquema. Mediante la implementación de ArcCe y Slurm (ver 2.4) en el CPD del IFCA, conseguimos una total integración del clúster HPC dentro del Tier-2 de CMS, de forma transparente, sin ser invasiva y manteniendo la seguridad. Ahora tenemos un sistemas de colas unificado para ambos sistemas, Slurm. Definimos diferentes "Particiones" o "Colas" en los nuevos componentes, cada una de ellas apunta a los

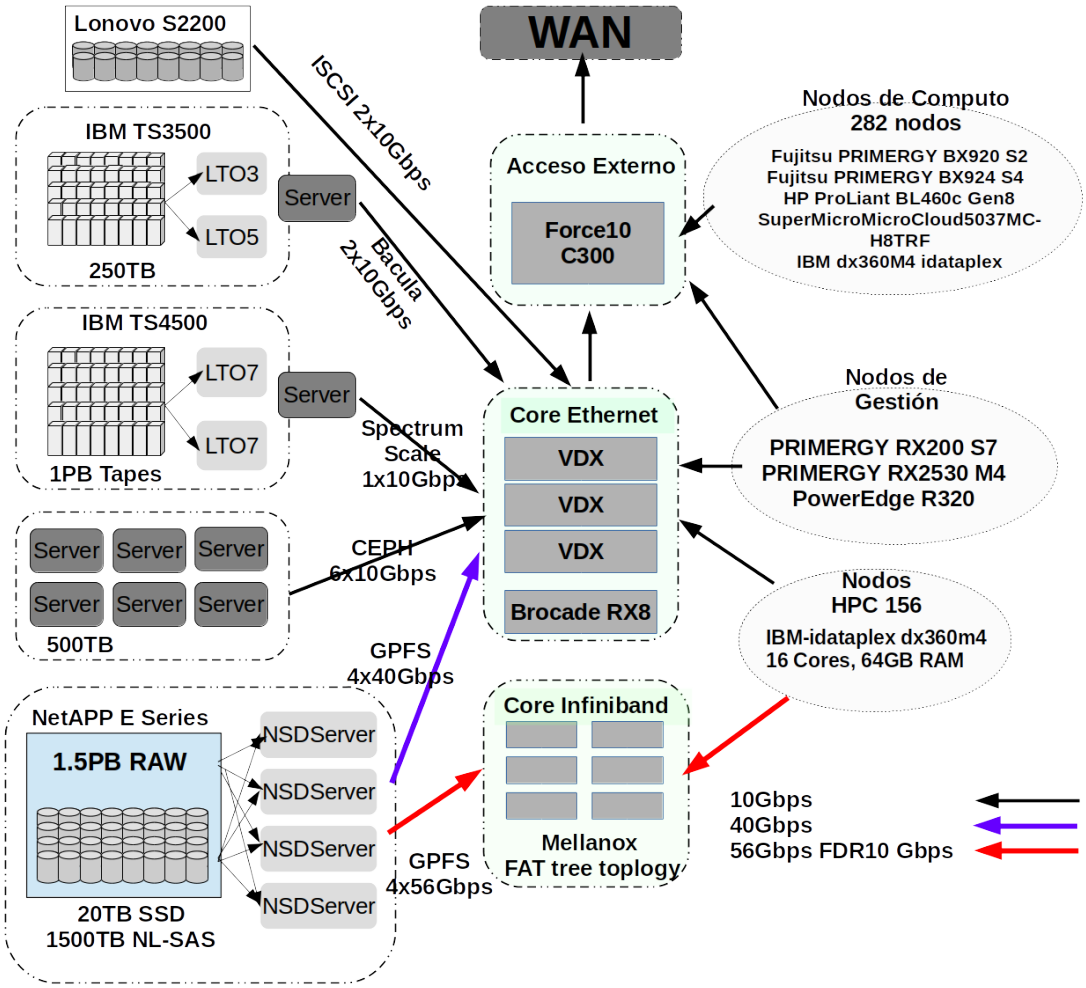


FIGURA 3.25: Descripción de la infraestructura hardware del CPD del IFCA año 2019.

diferentes grupos de nodos de cálculo, tres para el Tier-2 de CMS accesibles de forma genérica desde entornos Grid y una cuarta que hace referencia al hardware HPC, solo accesible para usuarios con un determinado "subject" (ver tab:3.5)

Para que los trabajos Grid se ejecuten de forma satisfactoria en los nodos HPC, ha sido necesario que los nodos tengan acceso tanto al "sandbox" del trabajo que se ejecuta, como área local al que el "subject" de usuario es mapeado por el servicio local de autorización (ver 2.3.4.5). En el caso del IFCA este acceso es simple ya que ambos directorios son exportados mediante NFS a los recursos del CPD que así lo soliciten.

Tabla 3.5: Configuración de la cola HPC en el ArcCE.

```
[authgroup: hpc]
subject = /DC=org/DC=terena/DC=tcs/C=ES/O=Consejo Superior de
    Investigaciones Cientificas/CN=JOSE IBAN CABRILLO BARTOLOME XXXXXX@csic.
    es
....
[mapping]
#map_with_file=any /etc/grid-security/grid-mapfile
map_to_user = hpc ghpcuser:ghpcgroup
map_with_plugin=all 30 /usr/libexec/arc/arc-lcmaps %D %P liblcmaps.so /usr/
    lib64 /etc/lcmaps/lcmaps.db arc
policy_on_map=stop
....
[queue: compute]
homogeneity=True
comment=Queue for HPC Jobs
totalcpus=16
nodecpu=Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
nodememory=64237
defaultmemory=2048
architecture=x86_64
opsys=Scientific Linux 7.5
osname=Scientific Linux
osversion=7.5
osfamily=linux
allowaccess=hpc

[queue: cloudcms]
homogeneity=True
comment=Queue for CMS Jobs
totalcpus=24
nodecpu=Intel(R) Xeon(R) CPU X5670 @ 2.93GHz
nodememory=45032
defaultmemory=2048
architecture=x86_64
opsys=Scientific Linux 7.5
osname=Scientific Linux
osversion=7.5
osfamily=linux
advertisedvo=dteam
advertisedvo=ops
advertisedvo=cms
advertisedvo=fusion
advertisedvo=enmr.eu
advertisedvo=ilc
advertisedvo=opencoast.eosc-hub.eu
advertisedvo=iber.vo.ibergrid.eu
advertisedvo=ops.vo.ibergrid.eu
denyaccess=hpc
```

Esta implementación, nos permite que a un usuario específico, o un grupo, con un rol determinado, acceda de forma exclusiva a los recursos HPC desde el entorno local del IFCA, asignado al DN del usuario Grid (ver 3.5) un usuario local con acceso a los recursos HPC (ghpcuser:ghpcgroup) y denegando el acceso a este usuario en el resto de colas. Es una configuración "ad hoc". Estos trabajos quedarán perfectamente contabilizados por el CE Arc y su sistema integrado de contabilidad Jura[75]. El usuario se beneficiará del acceso a recursos de cálculo paralelo, baja latencia y elevado ancho de banda en el acceso a disco.

Aunque la cola HPC no es accesible a las VOs Grid, se ha establecido una rutina (ver apéndice 02) que comprueba la existencia de trabajos pendientes en las particiones Grid, y la disponibilidad de recursos en la partición HPC, siendo capaz de mover los trabajos entre ellas de forma transparente tanto para el usuario, como para el sistema Grid (ver fig:3.26).

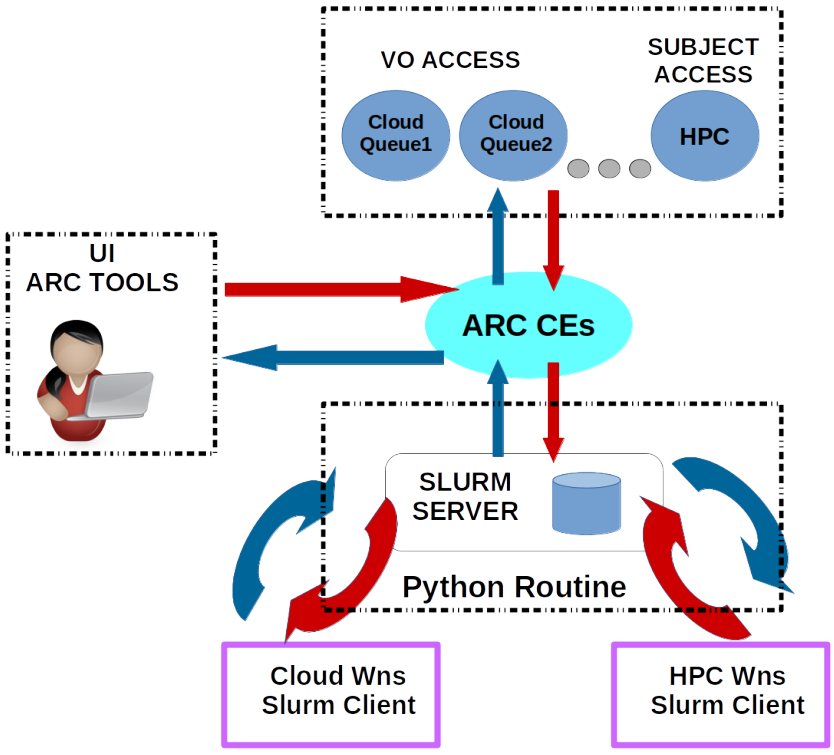


FIGURA 3.26: Esquema de acceso a recursos HPC desde el entorno Grid.

Tabla 3.6: Salida del comando sinfo para el controlador Slurm.

```
[root@ojancano02 ~]# sinfo
PARTITION   AVAIL  TIMELIMIT  NODES  STATE NODELIST
compute*       up   infinite      2 drain* node[5,17]
compute*       up   infinite     12  down* node[4,13,19,31-36,43,48,115]
compute*       up   infinite     18   resv node[2-3,6-12,14-16,18,72,90-92]
compute*       up   infinite      3   down node[1,64,93]
compute*       up   infinite      1    mix node28
compute*       up   infinite    122  alloc node[20-27,29-30,37-42,44-47,49]
login          up   infinite      2  drain altamira[1-2]
gpus           up   infinite      3 drain* gpu[1,3-4]
gpus           up   infinite      1  down* gpu2
cloudcms       up   infinite     12  down* cloudprv-01-[1-2,6,9]
cloudcms       up   infinite      6   idle cloudprv-01-[3-5,7-8]
cloudcmshp     up   infinite      6  down* cloudprv-02-7,cloudprv-02-E
cloudcmshp     up   infinite     23   idle cloudprv-02-[3-6,8-9]
cloudcmshp     up   infinite      2   down cloudprv-03-[6-7]
cloudcmsfj     up   infinite      6  down* cloudprv-04-3,cloudprv-04-C
cloudcmsfj     up   infinite     12   idle cloudprv-04-[4-9],cloudprv-04-A
```

Durante la pruebas realizadas con trabajos de muestra, se observamos que aproximadamente el uso medio de la infraestructura HPC por trabajos enviados vía Grid, es de ∼100 cores, que son los que quedan sin uso, el la partición con estado "**mix**" (ver tab:3.6), que es donde Slurm situa los nodos de una partición en uso, pero con cores libres. Esto supone un incremento aproximado del 8 % en los recursos computacionales que el Tier-2 proporciona a la CMS. La rutina, actualmente limita el acceso a nodos que estén totalmente libres, pero podría extenderse de forma simple aun número de cores limitados o incluso al llenado de los nodos sin uso (estado "**idle**").

Capítulo 4

Preservación de Datos

La preservación de los datos es un problema recurrente en las grandes colaboraciones científicas. Este problema, se vuelve más predominante en colaboraciones que gestionan datos de gran tamaño (∼GB) y en gran cantidad (∼PB) como vimos en la sección 1.1. Se presenta en aquellas colaboraciones con grandes infraestructuras de base tecnológica, debido al aumento en la precisión y capacidad de los dispositivos tecnológicos de captura de datos, en los modelos predictivos, relevantes en las ciencias experimentales o debido a la ingesta de datos y/o catalogación en colaboraciones dedicadas a las ciencias sociales, archivística, documentación. Si a esto le añadimos la necesidad de preservar los entornos de acceso, desarrollo y/o análisis, nos encontramos frente a un verdadero desafío tecnológico y conceptual.

4.1. La iniciativa DPHEP

La principal motivación, aunque no la única, detrás de cualquier iniciativa de preservación de datos dentro de HEP, es la posibilidad de obtener nuevos resultados físicos. En el caso de los experimentos LHC, existe un retraso considerable en la tasa de publicación. Es típico que los análisis de precisión continúen mucho tiempo después del final de la toma de datos, y así, poder utilizar todo el poder estadístico y el mejor conocimiento de los errores sistemáticos. Si tomamos como referencia los experimentos Large Electron-Positron collider (LEP), es posible que hasta el 10 % de los trabajos se finalizen en un período posterior a la toma de

datos. Prolongar la disponibilidad de los datos, puede generar una ganancia en la producción científica.

No debemos olvidar que tanto la divulgación como la formación son objeto de una política de preservación bien definida. Se ha comprobado que fomentar las actividades de divulgación científica repercute directamente en la percepción de la sociedad, generando prestigio y valor añadido, mientras que las actividades de formación acercan a los estudiantes los métodos y las herramientas que usarán en el desarrollo de su actividad profesional.

Las disciplinas HEP, no han tenido sin embargo tradición o modelo claro de preservación a largo plazo. La mayoría de los datos de los experimentos más antiguos se han perdido o no son accesibles fácilmente. El diseño del software, el presupuesto requerido y en general, las iniciativas de preservación, no han sido emprendidas por la colaboración en su conjunto, sino por un pocos individuos después del final de la toma de datos, con diferentes grados de éxito. En cambio, ahora es un campo que emerge rápidamente, donde el grupo de estudio Preservación de Datos para Física de Altas Energías (DPHEP)[76] se establece como el cuerpo coherente de múltiples laboratorios y experimentos, que situa la atención sobre la preservación.

4.1.1. Introducción

La iniciativa DPHEP, se creó a finales de 2008 para abordar la falta de soluciones concretas o pautas, en la problemática de la preservación de datos en las disciplinas HEP. La composición del grupo fue inicialmente impulsada por el experimento **BaBar** y el acelerador **Hadron Electron Ring Accelerator (HERA)** y sus experimentos **H1**, **ZEUS** y **HERMES**, a los que pronto se unieron **Belle**, **BES-III** y el Tevatron, con sus experimentos **CDF** y **DØ**. LEP y LHC también están representados dentro de la iniciativa DPHEP; **ALICE**, **ATLAS**, **CMS** y **LHCb** se unieron al grupo de estudio en 2011. Actualmente la iniciativa está constituida por diferentes laboratorios y centros de computación asociados como el **BNL**, el **CERN**, el **CSC**, **Deutsches Elektronen-Synchrotron (DESY)**, **Fermilab**, el **IHEP**, el **IN2P3**, el **INFN**, **IPP**, **KEK** y **SLAC**, **STFC** además de varias agencias de financiación[76].

DPHEP cuenta con el respaldo oficial del Intenational Comitee For Future Acelerator (ICFA). Los hallazgos iniciales del grupo de estudio se resumieron en un breve informe provisional de diciembre de 2009 y en el informe de estado completo, publicado en mayo de 2012, donde se abordan los siguiente puntos:

- Recorrido por las actividades de preservación de datos en otros campos.
- Descripción minuciosa del caso de Física.
- Guía para definir y establecer los principios de la preservación de datos.
- Actualizaciones de los experimentos y proyectos.
- Estimaciones de personal.
- Pasos futuros propuestos para establecer plenamente DPHEP.

Su objetivo principal, es crear un foro donde fomentar la discusión y la transferencia de conocimiento sobre soluciones tecnológicas aplicadas a la preservación de datos, software y conocimientos en la comunidad HEP. Coordina proyectos comunes de investigación y desarrollo, y establece herramientas comunes para toda la disciplina. Las áreas potenciales de colaboración se identifican en el documento del plan DPHEP como:

- Mejoras en la herramientas y en el proceso de ingesta de datos.
- Datos reconocibles para comunidades claramente identificadas bajo políticas de acceso bien definidas.
- Mejoras en la herramientas y en el proceso de gestión de archivos.
- Marco de validación.
- Entornos offline.

La información digital, es decir, los datos mismos, son cruciales, pero anteriores casos han confirmado que la conservación de cintas magnéticas de archivado, no es equivalente a la preservación, aunque es una parte crucial del proceso, es sólo la parte básica del mismo. Tratar con conjuntos de datos de diferente tamaño,

tipología y finalidad, incrementa La complejidad de las tareas de preservación; desde datos sin procesar, a datos reconstruidos o n-tuplas de análisis. Un modelo de preservación bien establecido debe incluir el hardware que permita el acceso a los datos, el software y el entorno comprenderlos. Estos aspectos implican un aumento de la dificultad a la hora de definir el modelo de preservación. Si el software experimental no está disponible, la posibilidad de incorporar nuevos algoritmos de reconstrucción, simulaciones de detectores o generadores de eventos se pierden. Sin un entorno de software bien definido y comprendido, el potencial científico de los datos puede estar limitado. Igual de importantes son los diversos tipos de documentación, que cubren todas las facetas de un experimento. Esto incluye las publicaciones científicas en revistas y bases de datos online, pero también las tesis publicadas, documentación interna, manuales, wikis, grupos de noticias, etc. Teniendo en cuenta esta definición inclusiva, el grupo de estudio DPHEP ha establecido una serie de niveles de preservación que quedan resumidos en la tabla 4.1.

Tabla 4.1: Modelo DPHEP.

	Modelo de Preservación	**Casos de Uso**
1	Proporcionar documentación adicional	Publicaciones relacionadas, wikis
2	Preservación de datos en un formato simplificado	Formación, Análisis de entrenamiento
3	Preservar los análisis, las versiones de Software y el formato de datos empleado en el análisis	Análisis científico completo, basado en la reconstrucción existente
4	Preservar el software de reconstrucción y simulación, así como los datos básicos necesarios	Salvaguardar toda la capacidad futura de los datos del experimento

Los niveles están organizados en orden de beneficio creciente, pero este orden implica un aumento en la complejidad de implementación y en el coste asociado. Cada nivel define unos casos de uso, y el modelo adoptado por cada experimento debe reflejar cual será el nivel de análisis que estará disponible en el futuro.

Los cuatro niveles representan diferentes áreas, que requieren iniciativas com-

plementarias:

- **Nivel 1**: Documentación.
- **Nivel 2**: Divulgación y simplificación; formatos para el intercambio de datos.
- **Nivel 3**: Preservación técnica. Acceso a datos utilizable a nivel de Análisis.
- **Nivel 4**: Software de reconstrucción y simulación.

Mientras que la mayoría de las colaboraciones involucradas en DPHEP persiguen algún tipo de estrategias de nivel 1 y 2, los niveles 3 y 4 son realmente el foco principal del esfuerzo de preservación de datos. La implementación de estos niveles, puede ejecutarse utilizando dos concepciones distintas:

- Manteniendo vivo el entorno actual el mayor tiempo posible; adaptando el código a cambios futuros a medida que suceden.
- Empleando técnicas de virtualización, que permitan seguir usando el entorno de forma aislada.

4.1.2. Casos de uso

A priori, se supone que los datos más antiguos son reemplazados por los datos del experimento de la próxima generación. Sin embargo, hay conjuntos de datos únicos disponibles en términos de partículas de estado inicial o centro de energía de masa o ambas, así como datos de una variedad de experimentos con blanco fijos. En estos casos, puede ser deseable revisar las mediciones antiguas o realizar otras nuevas con dichos datos, para lograr una mayor precisión a través de nuevos cálculos teóricos mejorados o para explorar nuevos análisis. Por ejemplo, un nuevo análisis de los datos del experimento JADE[77] tomados en Positron-Elektron-Tandem-Ring-Anlage (PETRA) llevó a una mejora significativa en la determinación de la constante del acoplamiento fuerte, algo que sin embargo, no era posible en el momento del análisis original.

Comenzaremos presentado un par de ejemplos de casos de preservación llevados a cabo por los aceleradores Hera y Tevatron.

4.1.2.1. Hera

HERA (ver fig:4.1) fue el primer y hasta ahora el único acelerador en el que colisionaron electrones(e^+) y protones(p^+). Alemania y once países más lideraron su desarrollo. Las instalaciones del acelerador se localizaba en DESY, Hamburgo, entre 15∼30m bajo tierra.

Los leptones y los protones se colisionaban en dos anillos de 6.3km de longitud, independientes, uno encima del otro, dentro del túnel. Tenía cuatro regiones de interacción, utilizadas por los experimentos H1, ZEUS, HERMES y HERA-B, todos ellos a su vez contaban con detectores específicos de partículas. Su energía en el centro de masas era de aproximadamente **320GeV** y alcanzó la máxima luminosidad posible de la época $7{,}5 \cdot 10^{31} cm^{-2} s^{-1}$. Los primeros datos de HERA se tomaron en el verano de 1992 y sus operaciones cesaron en junio de 2007, después de un período de toma de datos de 16 años.

El grupo de preservación de datos (DESY-DPHEP) se creó en 2009, poco después de la iniciativa DPHEP(ver 4.1). DESY-DPHEP está formado por miembros de las colaboraciones de HERA (H1, HERMES y ZEUS) y representantes de DESY-TI, DESY "Library" y de INSPIRE.

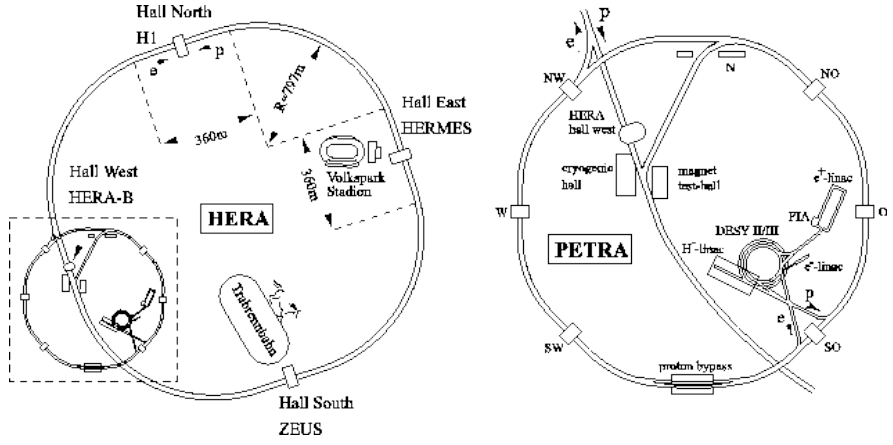


FIGURA 4.1: Hera.
[78]

A pesar de las diferencias en los modelos informáticos, los experimentos de HERA tenían muchos retos comunes que fueron identificados durante el proceso de planificación. Entre ellos, la migración a sistemas operativos más nuevos, acceso

a datos, ejecuciones de simulación, reconstrucciones y análisis, recompilación de software experimental en compiladores más nuevos, dependencias de software externo y validación de nuevas versiones de software. Era entonces sencillo, pensar en la necesidad de un marco de validación unificado.

DESY, permitía una prueba rigurosa del software del experimento contra cambios en las versiones del sistema operativo o dependencias de software externo. La preservación del entorno de ejecución, se realizó usando técnicas de virtualización. Una máquina virtual ejecutaba la instalación limpia del sistema operativo con requisitos de dependencia definidos por el software experimental. Las aplicaciones eran descargadas, ejecutadas y probadas en la Virtual Machine (VM) y finalmente, se comprobaba la salida de datos y el análisis de errores, estableciendo así el marco de validación.

Cada uno de los experimentos proporcionaba los paquetes de aplicaciones y las pruebas de validación, mientras que la división DESY-IT, era la encargada de mantener las imágenes virtuales con versiones de los diferentes sistemas operativos, ejecutar las VM y gestionar la salida de datos resultante.

El software de análisis utilizaba n-tuplas de uso común (datos reales y MC), contenían la información necesaria para realizar los análisis futuros y la mayoría de los análisis en curso, seleccionándose el formato n-tuples planas, sin objetos ni histogramas. El tamaño total de n-tuplas resultante a preservar, fue aproximadamente del 10-20 % del tamaño de los datos originales en Mini Data Summary Tape (MDST).

Una vez definidos los modelos de datos y estudiados los diferentes modelos de preservación de datos establecidos por el grupo DPHEP, organizados por niveles de beneficios versus complejidad y costo crecientes, HERA decidió establecer su modelo de Preservación en los niveles 3 y 4.

El método de almacenamiento y el acceso, seguiría siendo el predefinido por el experimento, basado en cintas magnéticas y en dCache (ver 2.3.5) como BackEnd de almacenamiento. También se incluiría un paquete independiente de MC, que utilizaría los ejecutables congelados existentes, necesarios para generar pequeños conjuntos de MC adicionales en el futuro, previendo posibles desarrollos teóricos.

El paquete de simulación de MC fue diseñado específicamente para no tener dependencias externas. Contenía todas las dependencias necesarias, como cali-

bración, alineación, geometría y ejecutables. Durante la prueba de validación, los archivos de registro eran comprobados después de cada paso de simulación en busca de algún tipo de error; si los resultados eran satisfactorios, se ejecutaba el siguiente paso y los resultados de las comprobaciones se almacenaban en archivos de registro de prueba, estableciendo un proceso de documentación. Al final de la prueba, el contenido de las n-tuplas se validaba por comparación con las n-tuplas de referencia que se produjeron con las mismas condiciones de simulación del detector, pero para diferentes condiciones ambientales. Utilizando el test de Kolmogorov se asegura de que los cambios del sistema operativo o la versión del software no cambiasen el contenido físico de la salida de simulación[79]. A continuación, se muestra el tipo de preservación final seleccionada por cada uno de los experimentos de HERA, así como una estimación de los datos preservados:

- **Zeus**: El modelo de análisis ZEUS estaba basado en MDST y contenía una gran serie de dependencias externas, que no era posible mantener una vez finalizada la toma de datos. Los datos RAW, MDST y las cintas con la producción de MC se eliminaron de los robots y se almacenaron en un lugar seguro. La cantidad total de espacio necesario para preservar los datos de ZEUS se redujo de 1PB a aproximadamente 100TB.

- **Hermes**: En el caso del experimento HERMES, se planificó la preservación de la cadena completa de análisis de software, basada en MDST. El software experimental se preservó usando máquinas virtuales. Los datos se almacenaron en dispositivos de cinta y la cantidad total de espacio necesario para preservar los datos de HERMES fue aproximadamente de 150-200TB.

- **H1**: Planeó preservar datos tipo RAW, así como al menos un DST y varias versiones de análisis de datos y producción de MC. La cantidad total de datos conservados del experimento H1 es de unos 200-500TB.

- **Hera-B**: La cantidad total de los datos que se conservaron en las cintas se estima en 250TB.

4.1.2.2. Tevatron

Continuando con otro experimento de HEP, el Tevatron se encontraba ubicado en Batavia, Illinois (Estados Unidos) en el Fermi National Laboratory (Fermilab). Aceleraba haces de protones (p^+) y antiprotones hasta energías de casi 1TeV por cada haz (de ahí su nombre), alrededor de una circunferencia de 6.3km. El Tevatron era el segundo acelerador de partículas más poderoso del mundo antes de finalizar la toma de datos el 29 de septiembre de 2011. El túnel de Tevatron está enterrado unos 10m bajo tierra, donde dos haces colisionaban en dos detectores de 5.000 toneladas cada uno de ellos, colocados alrededor del tubo del haz, en dos ubicaciones diferentes **Dzero** y **CDF**[5] (ver fig:4.2).

Los datos de los experimentos CDF y Dzero, conservan su valor para realizar mediciones de precisión a medida que aparecen nuevos cálculos teóricos, y poder validar cualquier nuevo descubrimiento que se observe en el LHC.

El "Run II Data preservation de Fermilab (R2DP)", tiene el objetivo de garantizar que ambas colaboraciones experimentales, Dzero y CDF, tengan la capacidad de realizar física completa al menos hasta el año 2020. Como hemos visto anteriormente en el caso de Hera (ver 4.1.2.1), para mantener completa la capacidad de análisis, el proyecto debe preservar no solo los datos experimentales, sino también sus entornos informáticos y de software. Esto requiere asegurar que los datos permanecen totalmente accesibles y que el software experimental y los entornos informáticos sean compatibles con los hardware modernos. Además, los trabajos de los usuarios deben poder ejecutarse sin recursos informáticos dedicados y el movimiento de datos a estas nuevas instalaciones debe realizarse dentro de un entorno de software familiar, con un mínimo esfuerzo por parte del usuario final y de los administradores de la infraestructura. La documentación crítica de R2DP, incluye las páginas web existentes, bases de datos, documentos internos e instrucciones claras y concisas que detallen cómo los usuarios deben modificar su entorno de trabajo habitual para trabajar en este entorno R2DP[80].

- **Preservación de la Documentación**: Preservar el conocimiento es una parte de vital importancia en los experimentos. En ellos se define el conocimiento, documentación interna, notas, presentaciones, wikis, tutoriales o incluso archivos de listas de correo. CDF como D0 se asociaron con INSPIRE,

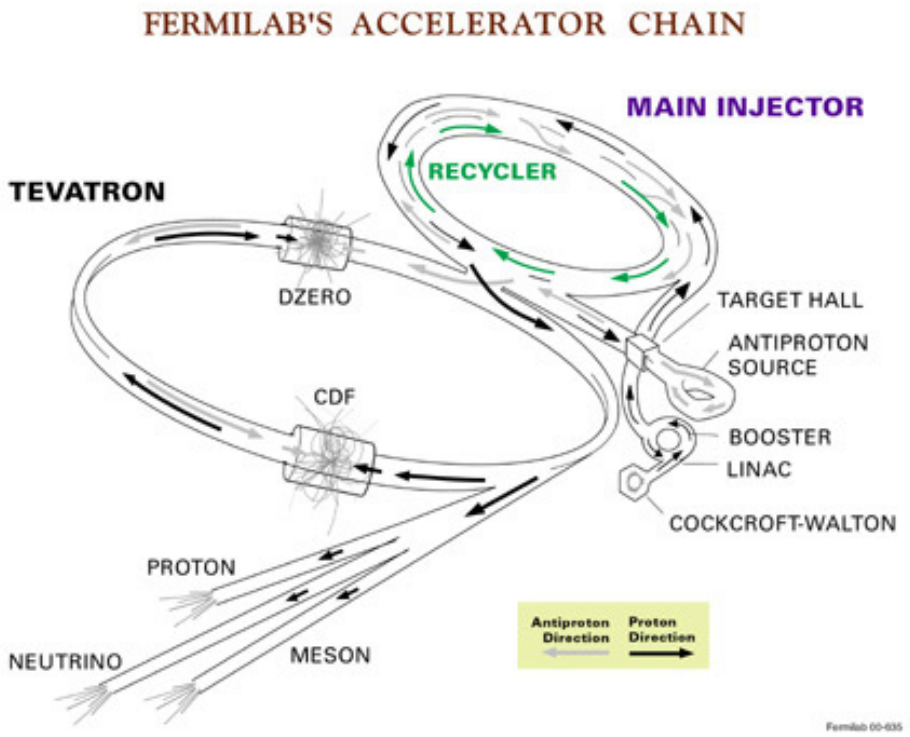


FIGURA 4.2: Tevatron.

para transferir la documentación interna del proyecto a repositorios a largo plazo. D0 mudó sus notas internas a una instancia de Indico[81] alojada por Fermilab. Fermilab aseguraría que los archivos y las listas de correo de cada experimento estén accesibles durante la vida del proyecto; las instancias de Wiki/Twiki se movieron a páginas web estáticas para facilitar su migración a nuevos servidores si fuera necesario.

- **Preservación de Datos**: En el momento del cierre de Tevatron, los datos para CDF y D0 se almacenaron en cintas magnéticas LTO4, que tienen una capacidad por cinta de 800GB. Un análisis de las tecnologías de cinta disponibles en ese momento concluyó que las cintas T10K, con una capacidad de hasta 5TB por cinta, sería la mejor opción a corto plazo para almacenamiento de archivos en Fermilab. Si bien era teóricamente posible dejar los datos CDF y D0 en el almacenamiento LTO4, se tomó la decisión de migrar estos datos al almacenamiento T10K por dos razones:

- **Disponibilidad**: El almacenamiento en LTO4 podía ser un configuración no soportable, a medida que el uso de las cintas LTO4 disminuye en la industria, podía no ser un almacenamiento de reemplazo fácilmente disponible.
- **Precio**: La escasez de los medios de almacenamiento para una tecnología antigua, aumenta los costos generales a largo plazo de mantener un almacenamiento LTO4.

En previsión de falta de disponibilidad y aumento del precio de los dipositivos LTO4, se decidió adquirir cintas T10K y todos los datos de CDF y D0 fueron migrados al nuevo formato. La migración duró aproximadamente **dos años** en completarse, utilizando la tecnología del momento. Los datos CDF, $\sim$ 8.8PB y los de D0 $\sim$ 8.8PB, comparten ahora recursos de acceso de cinta con experimentos activos en Fermilab. Tanto CDF como D0 migraron todos los datos almacenados en cinta, incluidos datos de detector sin procesar (RAW), datos de detector reconstruidos (RECO) y conjuntos de datos derivados y simulación. CDF también realizó una copia externa adicional de sus datos sin procesar $\sim$ 4PB4 en el Centro Nacional de Investigación y Desarrollo en Tecnología de la Información (CNAF) en Italia. Tanto CDF como D0 utilizaron la base de datos de Oracle soft-software para datos no estadísticos, como calibraciones de detectores. El costo de mantener licencias de Oracle, presentaba un desafío a largo plazo. En las bases de datos, el esquema estaba fuertemente entretejido con el software de análisis. Convertir a una solución de base de datos de código abierto más económica, implicaba una inversión prohibitiva en recursos humanos. Ambos experimentos decidieron mantener los sistemas de bases de datos Oracle durante toda la vida del período de preservación de datos. La DB de Oracle se migró a la versión mas actualizada en el momento del cierre de Tevatron. Futuras actualizaciones de la DB de Oracle podría potencialmente alterar el esquema existente, y por lo tanto, el software de análisis. Se elaboró un plan de contingencia donde la versión actual y el esquema podrían congelarse y ejecutarse, en aislamiento de red, si fuera necesario en el futuro, incluso aunque el soporte para esa versión hubiera cesado. Para evitar la obsolescencia del hardware, los experimentos

movieron sus bases de datos a entornos virtuales.

- **Preservación del Software y el Entorno**: CDF y D0 migraron su software al repositorio CernVM File System (CVMFS) del CERN. CVMFS ha sido ampliamente adoptado por muchos experimentos, entre ellos los actuales de Fermilab. Esta migración, permite mantener versiones de software CDF y D0 para el futuro sin una inversión significativa en recursos dedicados. En el momento del cierre de Tevatron, las versiones de software de CDF y D0, se ejecutaban bajo Scientific Linux 5 (SL5), y dependían de bibliotecas de compatibilidad de Scientific Linux 3 (SL3). Los dos experimentos eligieron diferentes estrategias para garantizar funcionalidad de sus ejecuciones durante todo el periodo de preservación.

  - **CDF**: Empleaba dos versiones estables de software, una para reconstrucción de datos y análisis de colisión y otra para generación MC y simulación. CDF preparó versiones heredadas de ambos softwares que fueron despojadas de cualquier paquete obsoleto y dependencia anterior a SL5. El código CDF para cualquier análisis se verificó y validó. Se garantizó que los recursos disponibles en el centro serían totalmente edificables y ejecutables en SL5, además, se estableció un proceso relativamente simple para garantizar que la versión heredada se pudiera construir y ejecutar en Scientific Linux 6 (SL6), sistema operativo de destino R2DP. El tamaño total de la base del código CDF en las versiones anteriores era de 326GB, incluyendo el código compilado y la mayoría de las dependencias externas.
  - **D0**: Decidió quedarse con las versiones actualizadas de software que estaban vigentes en el momento del cierre de Tevatron, y también se aseguró de que la compatibilidad de las bibliotecas 32bits del sistema se instalara en nodos de trabajo en Fermilab, donde D0 planeaba ejecutar trabajos durante toda la vida útil del proyecto R2DP, asegurando la compatibilidad pre-SL6 requerida. Las bibliotecas se agregarían al repositorio CVMFS si fuera necesario. El tamaño total del repositorio de software D0 en CVMFS, incluida la base de código, los ejecutables y las dependencias externas, era de 227GB.

- **Gestión de Acceso a Datos**: Tanto CDF como D0 utilizaban el servicio de Acceso Secuencial a Metadatos (SAM) para la gestión de los datos. Para D0, parte de la infraestructura de SAM incluía discos de caché dedicados de 1PB agregado en los nodos de cálculo del clúster D0, que permitían la organización rápida de los archivos de entrada y los trabajos . Este método no era soportable a largo plazo, por lo que, D0 implementó una instancia dCache (ver 2.3.5) de 100TB para almacenar archivos de entrada a los nodos de trabajo. Los resultados de la prueba no mostraron degradación en el rendimiento en relación con la caché dedicada a SAM. CDF ya usaba dCache para el almacenamiento en caché con respaldo en cinta magnética, por lo que una vez que se realizaron los cambios necesarios en el código para usar versiones actualizadas de SAM, el acceso a datos continuó, sin necesidad de aplicar cambios en la infraestructura de hardware.

4.2. Política de Preservación de Datos para CMS

Los datos de CMS son únicos y son el resultado de una gran inversión humana y financiera, llevada a cabo por parte de la comunidad internacional participante en el experimento. Ninguna muestra de datos de esta complejidad y valor, ha sido preservada o está disponible hoy en día para su posterior reutilización.

Mediante la política de preservación de datos, la colaboración de CMS se compromete a preservar sus datos, en diferentes niveles de complejidad, y así permitir su reutilización por una comunidad amplia donde se incluyen:

- Miembros de la colaboración mucho después de que se tomen los datos.
- Científicos HEP experimentales y teóricos que no sean miembros de la colaboración
- Iniciativas educativas y de divulgación, ciudadanos científicos y el público en general.

CMS proporcionará mediante lo establecido en su política (2012)[82], acceso abierto a sus datos después de un período de embargo adecuado pero relativamente corto, lo que permitirá a los colaboradores de CMS explotar plenamente su

potencial científico. La política de preservación de datos de CMS describe los principios de reutilización y acceso abierto, así como los actores relevantes en estas tareas, sus roles y responsabilidades. Para explotar estas oportunidades de reutilización, se necesitan recursos. Así el nivel de soporte proporcionado a los usuarios externos, dependerá de la financiación disponible.

Los datos de CMS son adquiridos de muchas formas, desde datos en bruto, experimentales o simulados, datos reconstruidos, conjuntos de datos generados por flujos de trabajo de análisis o datos representados en publicaciones científicas, como ya hemos visto en el Capítulo 3. Cada una de estas capas tiene el potencial de ofrecer diferentes oportunidades para la reutilización a largo plazo y plantea diferentes desafíos a la hora de la preservación como hemos visto en los puntos anteriores (ver 4.1.2.1 y 4.1.2.2).

Los datos representados en las publicaciones, pueden conservarse aprovechando prácticas ya existentes en la colaboración, como la publicación de acceso abierto y las plataformas de terceros como INSPIRE (ver 1.1.4.2). Si nos situamos más cerca de los datos sin procesar, aparecen diferentes desafíos que implican un cambio de paradigma en documentación y el archivado en profundidad de los análisis, durante el proceso de publicación: la preservación de los paquetes de software de reconstrucción y simulación con todas sus dependencias.

Se han identificado cuatro niveles de preservación para datos HEP (ver 4.1). Si tratamos de establecer una relación entre estos niveles y la política de preservación y reutilización de datos de CMS, podemos definir 4 diferentes niveles:

- **Política de nivel 1 de CMS**: Publicación de resultados científicos en revistas de acceso abierto. CMS se esforzará por proporcionar información numérica adicional para facilitar la reutilización inmediata y la combinación de estos resultados. Esta información es proporcionada y archivada a largo plazo por terceros confiables como INSPIRE, Rivet[83] y HEPData[84].

- **Política de nivel 2 de CMS**: Acceso a formatos de datos simplificados para varios niveles de reutilización inmediata: interpretaciones teóricas, análisis limitados, educación, divulgación a través de CERN Open Data y HEPData.

- **Política de nivel 3 de CMS**: CMS conservará los datos reconstruidos y

las simulaciones. Mantendrá disponible una copia de los datos reconstruidos y las condiciones del detector para cada período de toma de datos. Se proporcionará y mantendrá un entorno informático virtualizado, compatible con la versión de software con las que se puedan analizar los datos originales. Formatos de datos simplificados, comunes a todos los conjuntos de datos, e independientes de versiones de software específicas, serán adaptados para el análisis de la física a largo plazo. Los procedimientos de análisis, los flujos de trabajo y el código se conservan como parte del repositorio de código de CMS bajo la responsabilidad del proyecto fuera de línea. La responsabilidad del archivo de los datos recae en la actual infraestructura jerarquía de datos de CMS (ver 3.2.2).

- **Política de nivel 4 de CMS**: Incluye los datos sin procesar, el software y la documentación necesaria para acceder, reconstruir y analizar. Cualquier versión del software de reconstrucción y análisis de CMS debe ser compatible con los datos sin procesar desde el inicio de la toma de datos. Se almacenará una copia de custodia de los datos en el CERN, correspondiente para ese conjunto de datos. El software CMS se lanza bajo una licencia de código abierto y CMS asegurará de que las copias de custodia del software también se conserven y estén accesibles gratuitamente.

CMS proporcionará acceso abierto a sus datos en diferentes momentos con demoras apropiadas, lo que permitirá a los miembros de la colaboración explotar por completo el potencial científico de los datos antes del acceso abierto:

- **Para el nivel 1**, los datos adicionales están disponibles en el momento de la publicación.

- **Para el nivel 2**, las muestras de formato de datos simplificados se distribuyen rápidamente según lo determine el "Collaboration Board" de CMS.

- **En el nivel 3**, la publicación estará acompañada de software estable, de código abierto adecuado y bien documentado. Se llevará a cabo anualmente entre los periodos de tomas de datos del LHC , "Long Shutdowns (LS)", y en la medida de lo posible durante los períodos de toma de datos. Durante

la vida útil de CMS, se define el límite máximo de datos de acceso público, comparado con los disponibles para la colaboración, en un 50 % de la luminosidad integrada recogida por CMS. Los datos serán publicados 3 años después de la toma, aunque la Junta de Colaboración podría, en circunstancias excepcionales, decidir liberar algunos conjuntos de datos particulares antes o después.

- **Para el nivel 4**, los datos sin procesar, RAW, no son útiles para el análisis y no se incluirán en los datos públicos.

La política de preservación de datos, reutilización y acceso abierto, motiva y define el enfoque de CMS en la preservación y el acceso a los mismos, con varios niveles de complejidad. La implementación de la política ha derivado en un proyecto dedicado dentro de la colaboración, cubriendo dos áreas:

- Análisis y preservación de datos para uso interno de la colaboración.
- Política de datos abiertos, difusión (Outreach).

4.2.1. Niveles de Datos a Preservar

Como hemos visto anteriormente, CMS archiva los datos en niveles diferentes (ver 3.2.1) dependiendo de su uso final. Existen dos copias distintas de los datos sin procesar, RAW, y la responsabilidad de su custodia es del Tier-0 y del Tier-1 que almacena los datos. Nunca se eliminan datos sin procesar. Los datos reconstruidos (RECO), se almacenan en formato AOD, que es un subconjunto de los objetos de datos reconstruidos (RECO). AOD (MiniAOD, NanoAOD) por si solo, es suficiente para la mayoría de los tipos de análisis de física y es el formato de datos elegido por CMS para fines de preservación a largo plazo (Política de CMS nivel 3).

Todos los datos de colisiones tomadas durante 2011-2012 y las correspondientes simulaciones, fueron reprocesados en un conjunto de datos en formato AOD, con una única versión de software CMS. Los duplicados de estos conjuntos de datos, pueden ser eliminados después de una consulta dentro de los grupos de análisis de física de CMS, una vez que los análisis activos han concluido.

4.2.2. Estado de la iniciativa de preservación de datos y acceso abierto de CMS

Alrededor del 50 % de los datos de la primera toma de datos, "**RUN 1**" del LHC ya habían sido publicados en años anteriores:

- Los datos de 2010 liberados en 2014
- Los datos de 2011 liberados en 2016
- Los datos de 2012 liberados en 2017

En julio del 2019, CMS ha liberado los datos restantes correspondientes a la primera toma de datos. El volumen de datos en abierto se eleva a más de 2PB. El lanzamiento incluye conjuntos de datos preparados específicamente para su uso en aplicaciones de ML o en ciencia de datos, que por primera vez incluyen muestras de simulación en el formato "MiniAODSIM". Se incluyen además una pequeña muestra de datos sin procesar del periodo 2010 a 2012 y se proporcionan instrucciones y ejemplos sobre generación de eventos simulados y análisis de datos en contenedores aislados, todo ello bajo la exención Creative Commons CC0[85], y accesibles a través del portal CERN Open Data.

4.3. Casos de uso en CMS

La colaboración de CMS acordó la publicación en abierto de varios conjuntos de datos tomados en el año 2010, con dos objetivos específicos: divulgación y educación. Se publicaron más de 300.000 eventos entre los que se incluían electrones, muones, J/psis, Upsilons, bosones W/Z y eventos candidatos de Higgs. La publicación de datos de 2010 se basó en la última versión del reprocesamiento del conjunto de datos completo, que tuvo lugar en la primavera de 2011.

4.3.1. Implementación de la política de Preservación de CMS

Como hemos visto, la política de datos de CMS, define el enfoque de reutilización y acceso abierto a datos de la colaboración. La implementación de la política

motiva la preservación de datos y evoluciona en una iniciativa dedicada dentro de la colaboración. La búsqueda de soluciones de CMS para implantar esta política, nos lleva a promover infraestructuras comunes allá donde sea posible, intentando que sean extrapolables para el resto de experimentos del LHC.

Siguiendo la conclusiones extraídas del modelo DPHEP (ver tabla 4.1) se intentan cubrir diferentes áreas: publicación, validación, preservación de datos y análisis y acceso abierto.

El contexto de ejemplo de esta aplicación de la política de preservación de CMS, es parte de un proyecto más amplio del Ministerio de Educación de Finlandia[86] llevado a cabo durante el año 2014. Un público de importancia para el uso de estos datos abiertos es la enseñanza de nivel secundaria y dado que varios grupos de secundaria de Finlandia visitan el CERN cada año, los ejemplos basados en datos de colisión reales son a nuestro entender, un material gratificante para cualquier clase y grupo interesado en el tema, pueda o no asistir al CERN. El contenido de estos ejemplos no necesita limitarse a la física de partículas, sino que abarca otros campos como son:

- Conceptos básicos de física.
- Análisis estadístico de datos.
- Métodos numéricos
- Gestión de datos.
- Visualización.

4.3.1.1. Publicación y Datos de Contexto

Siguiendo la política de datos abiertos de CMS de nivel 3, se enfatiza la importancia de la posibilidad de reutilizar los datos de la colaboración. Paralelamente a los documentos de acceso abierto, la publicación de datos numéricos adicionales debe proporcionarse en un formato que se pueda extraer fácilmente para su uso posterior. El formato de archivo que la colaboración designó como apropiado para el análisis de física fue AOD y su publicación, fue acompañada de un software estable y de código abierto necesario para una serie de análisis de

ejemplo, acompañado de la documentación adecuada. Sin embargo, los archivos AOD y el software de análisis son intrínsecamente complejos y es necesario un buen conocimiento de ambos para realizar un análisis físico completo.

Como caso de uso para el contexto definido anteriormente, se seleccionó la preparación de aplicaciones para estudiantes de secundaria, con alumnos comprendidos entre los 15 y los 19 años. Fue necesario identificar las herramientas e instrucciones necesarias para presentar los datos a este tipo de audiencia, con limitados conocimientos de la materia.

Se necesitó procesar el formato AOD original en un formato apropiado para el tipo de usuario final, como consecuencia, hubo que proporcionar herramientas de validación para este reprocesamiento intermedio, así como proporcional acceso al entorno de software de la colaboración CMS en este caso. Esto proporcionó una novedad en comparación con la liberación de pequeñas muestras de datos preseleccionados del acceso abierto completo, es decir, posibilitó a los miembros de CMS, tener acceso directo a los datos de investigación originales y a las herramientas para procesarlos y analizarlos más a fondo.

Mediante herramientas como Robust Independent Validation of Experiment and Theory (RIVET), actualmente en su versión 3.0.1, disponemos de un sistema que nos permite la validación mediante eventos de MC. Nos proporciona un conjunto de análisis experimentales útiles para el desarrollo, validación y ajustes de MC. El uso de este tipo de herramientas, nos permite realizar comparaciones entre los datos observados y los eventos de MC generados y sus validaciones. Esta es una de las formas más extendidas de preservar el código de análisis dentro del LHC. CMS emplea servicios de biblioteca digital como CDS/Invenio[13], INSPIRE y HEPData donde los usuarios pueden alojar los resultados.

4.3.1.2. Preservación a nivel de Bit

Si bien es verdad que conceptualmente la fase de preservación de bit, o dato físico es la más básica, el manejo de estos volúmenes de datos, no está exento de desafíos tecnológicos, por ello, aunque el modelo de computación CMS (ver 2.3) ofrece una base sólida para preservación de datos a largo plazo, se establece HEPiX[87] como el grupo de referencia encargado de proporcionar las directrices

de preservación de datos a nivel de bit. Según las estimaciones de HEPiX, los datos a preservar por el LHC son:

- **Run-2** (2015-2018): $\sim$ 50PB/año.
- **Run-3** (- 2022): $\sim$ 150PB/año.
- **Run-4** (2023 -): $\sim$ 600PB/año.

HEPiX nos proporciona recomendaciones para el almacenamiento sostenible de archivos de datos HEP en múltiples sitios y sobre diferentes tecnologías. HEPiX ayuda a dar respuesta a diferentes aspectos de la preservación de datos a nivel de bit, como:

- Tecnología de preservación a largo plazo. La tabla 4.2, muestra los costes asociados al almacenamiento a largo plazo, eligiendo los formatos disponibles en 2019. En el caso del mantenimiento, se supone que el disco se renueva por completo a los 5 años, exigiendo una garantía de 5 años al fabricante. Se ha tenido en cuenta una reducción anual del 30 % del precio en la adquisición de los nuevos sistemas basados en disco, mientra que para el caso de la cintas simplemente se renovarían las cabezas lectoras, manteniendo la compatibilidad de acceso. Con estos datos obtenemos un coste por TB acumulado de 750€ aproximadamente para el disco y 110€ para la cinta, ambos a 20 años.

Tabla 4.2: Costes Disco vs Cinta a 20 años.

	EoL	**Coste Energía 1PB**	**Coste Adquisición 1PB**	**Coste Mantenimiento**	**Coste Renovación**
Disco	5	∼98k€	∼250k€	0	∼400k€
Cinta	20	∼12k€	∼50k€	∼20k€	∼25k€

- Monitorización y verificación del contenido del archivo. “Read-Write-Read”. Valores Checksums (md5sum, adler32).
- Procedimientos para recuperar datos no disponibles y/o perdidos.

- Migración de archivos a tecnología de nueva generación, migración entre medios (Repacking).

4.3.1.3. Preservación del Análisis

Siguiendo el esquema DPHEP, a nivel de los datos reconstruidos y su reutilización, la experiencia de los otros experimentos HEP indica que el mayor desafío en la preservación de datos es la pérdida de conocimiento y experiencia. Si bien los experimentos de CMS y HEP en general, funcionan muy bien en el registro de metadatos "inmediatos", tales como números de eventos, ejecuciones, condiciones de haz, versiones de software utilizadas en el reprocesamiento de datos, etc, en el ámbito de los metadatos contextuales, queda mucho por hacer. Definimos datos contextuales como aquellos datos que nos proporcionan la información práctica necesaria para poner los datos en contexto y analizarlos. Por lo tanto, CMS está buscando activamente soluciones que permitan el registro sencillo de este conocimiento detallado, disponible en el momento de la realización del análisis, pero que es rápidamente olvidado una vez que cesa la actividad.

Para poder realizar una validación a largo plazo, tanto de los datos como del software, se debe extraer y registrar la información y las herramientas necesarias. CMS ya dispone de potentes herramientas para la validación de versiones de software y para comprobar la calidad de los datos (ver 3.3.1 y 3.3.3).

En el caso de los procesos computacionales, los análisis físicos en sí mismos son intrínsecamente complejos debido al gran volumen de datos y algoritmos involucrados. Aunque se mantiene una documentación exhaustiva sobre los métodos de análisis, la complejidad de las implementaciones de software a menudo oculta detalles mínimos pero cruciales. Sin embargo, es necesario garantizar la usabilidad de la investigación a largo plazo. Esto es particularmente desafiante hoy en día, ya que gran parte del código de análisis está disponible principalmente dentro del pequeño equipo que realiza un análisis. La reproducibilidad requiere un nivel de atención y cuidado que no se satisface con la publicación de código indocumentado o datos inaccesibles. Así, el marco de preservación y reutilización de análisis CMS se basa en tres pilares:

- **Descripción**: Describir la estructura adecuadamente, el conocimiento detrás de un análisis de física de cara a su futura reutilización. Describir todos los activos de un análisis y realizar un seguimiento de la procedencia de los datos, asegurando suficiente documentación, así como enlaces asociados.

- **Captura**: Retener la información sobre los datos de entrada de análisis, el código de análisis y sus dependencias, el entorno computacional durante la ejecución, así como los pasos del flujo de trabajo de análisis, y cualquier otra dependencia necesaria en un repositorio digital confiable.

- **Reutilización**: "Instanciar" activos de análisis conservados y flujos de trabajo computacionales en entornos virtuales para permitir su validación o ejecución con nuevos conjuntos de parámetros.

Por ello es necesario el desarrollo de herramientas personalizadas, en lugar de un repositorio de datos abierto. Estas herramientas deben preservar la experiencia de una gran colaboración, proporcionando un lugar centralizado donde los componentes dispares de un análisis se puedan agregar al principio, y luego evolucionar a medida que el análisis se valida y verifica.

Herramientas como CERN Analysis Preservation (CAP)[88] y Reusable Análysis (REANA)[89], dejan la decisión de cuándo un conjunto de datos o un análisis completo se comparte públicamente en manos de los investigadores. Se puede admitir el acceso abierto, pero la arquitectura no depende de que los datos o el código estén disponibles públicamente. Proporcionan a las colaboraciones experimentales un control total sobre el procedimiento de publicación y, por lo tanto, admite totalmente el procesamiento interno, los protocolos de revisión y los posibles períodos de embargo.

El servicio CERN Analysis Preservation (CAP), aún en desarrollo, es una instancia de repositorio digital dedicada a describir y capturar activos de análisis. El servicio utiliza una estructura flexible de metadatos conforme a la notación abierta de JavaScript (JSON). Estos esquemas describen los análisis, ayudando a los investigadores a identificar, preservar y encontrar la información sobre los componentes de los mismos. En los ficheros "JSON" se define una descripción completa, desde configuraciones experimentales, muestras de datos, código de

análisis, incluso enlaces a presentaciones y publicaciones. El servicio CAP crea una forma estándar de describir y documentar un análisis, facilitando la búsqueda de información física de alto nivel asociada con análisis físicos individuales y su reproducibilidad. Permite a los investigadores individuales depositar material por medio de una interfaz o mediante un cliente de línea de comandos . El CAP importa información de las bases de datos internas de las colaboraciones de LHC.

REANA intenta automatizar los análisis físicos, de tal manera que puedan ejecutarse con un solo comando. La automatización de todo el proceso de análisis mientras aún este se encuentra activo, permite ejecutarlo fácilmente, así como preservarlo completamente sin problemas, una vez que termine y los resultados estén listos para su publicación. REANA es un componente independiente dedicado a "instanciar" análisis de datos de investigación conservados en la nube. Si bien nació de la necesidad de volver a ejecutar los análisis conservados en el marco del CERN Analysis Preservation, se puede usar para ejecutar análisis activos antes de que se publiquen y preserven. Usa información sobre los conjuntos de datos de entrada, el entorno computacional, el marco del software ,el código de análisis y los pasos del flujo de trabajo computacional para ejecutar el análisis. REANA incorpora tecnologías modernas de contenedores para encapsular el entorno de ejecución necesario para varios pasos del análisis[90].

Todos estos servicios, desarrollados a través de software gratuito y de código abierto, se esfuerzan por habilitar datos y son extensibles a otras comunidades utilizando modelos de datos flexibles compatibles con FAIR[91] . Para todos estos servicios, la captura y preservación de la provisión de datos ha sido una característica clave del diseño. La procedencia de los datos facilita la reproducibilidad y el intercambio de datos, ya que proporciona un modelo formal para describir los resultados publicados (ver fig:4.3).

4.3.1.4. Acceso Abierto

Poner en práctica la política de acceso a datos abiertos, requiere un entorno informático compatible. El primer lanzamiento público (ver 4.2.1) se basó en las versiones de software que se ejecutaban sobre el mismo sistema operativo SLC5. Para facilitar el acceso a los usuarios no CMS, se preparó una imagen de VM, en

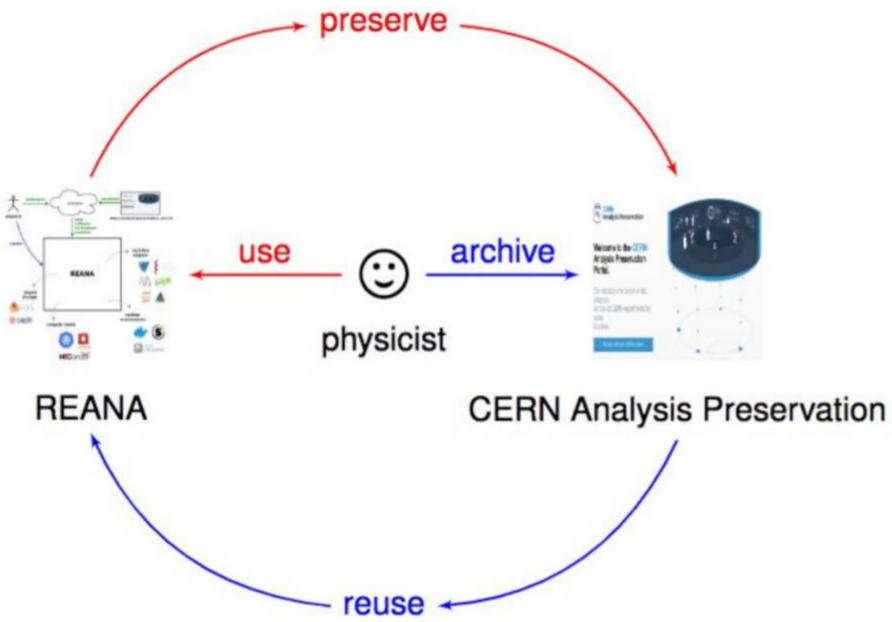


FIGURA 4.3: Flujo de la preservación del Análisis.

un formato utilizable por la aplicación VirtualBox de acceso libre, muy popular y ampliamente extendida. El flujo de trabajo inicial se mantuvo lo más cercano posible al estándar. Se proporcionaba acceso a los datos mediante un servidor Xroot, que había sido comisionado con acceso anónimo en modo solo lectura a las áreas de datos definidas sobre directorios localizados en EOSCMS[92] en el CERN, limitado aún más por el cortafuegos del equipo, evitando accesos no deseados y solo accesible para aquellos centros involucrados en la fase de prueba.

Como los espacios de tiempo y mano de obra son limitados, las posibilidades de reprocesar los datos completos de 2010 se estableció con la misma versión heredada de los datos 2011-2012. El entorno VM preparado para acceso abierto sirvió como punto de entrada a los datos de 2010. Los esfuerzos realizados para documentar los datos y establecer un conjunto conciso de instrucciones para su uso, beneficiará a los usuarios finales, tanto dentro como fuera de la colaboración[93].

4.3.2. International Particle Physics Outreach Group

Cada año, más de 13.000 estudiantes de educación secundaria de 52 países se acercan a una de las 215 universidades o centros de investigación que participan en las "International Masterclasses - hands on particle physics". Su propósito: desvelar los misterios de la física de partículas. Esta iniciativa permite que durante una jornada los alumnos analicen datos de experimentos reales del mayor acelerador de partículas del mundo, el LHC, convirtiéndose así en físicos de partículas por un día.

Las Clases Magistrales son impartidas por científicos en activo de reconocido prestigio, proporcionando la visión necesaria sobre temas fundamentales de materia, fuerzas, y métodos de investigación básicos. Así, se facilitarán las herramientas requeridas para realizar las mediciones con los datos en abierto del LHC. Al final del día, como en las colaboraciones internacionales, los participantes se unirán a una videoconferencia donde se discutirán y combinarán los resultados[94].

Extendiendo la experiencia llevada a cabo en la "implementación de la política de datos abiertos de CMS" a todos los experimentos del LHC, las clases son desarrolladas y organizadas por QuarkNet[95] y el International Particle Physics Outreach Group (IPPOG)[96]. Existen varias herramientas disponibles para examinar los eventos. Desde el año 2015, una interfaz de acceso a los eventos en línea desarrollada por I2U2[97] está disponible. Esta interfaz gráfica, mucho más intuitiva que la forma de acceso anterior, se basa en el aspecto y la funcionalidad del iSpy[98] (ver fig:4.4) muestra el evento y lee el formato de archivo ig. Cada evento está en formato JavaScript Object Notation (JSON) y un archivo ".ig", que es un archivo ".zip".

Los archivos ig se crean utilizando el marco de software del experimento, en este caso CMS, convirtiendo el formato CMS a ig. Este software de visualización en línea solo requiere un navegador actualizado. Un ejemplo de visualización de un evento se muestra en la fig:4.4. Se pueden desarrollar otras herramientas en línea del lado del cliente desde csv y JSON, archivos de resumen generados a partir de los archivos ig.

El IFCA ha participado durante más de 15 años en este evento anual de IPPOG. Durante este tiempo más de 1.000 alumnos de secundaria han participado

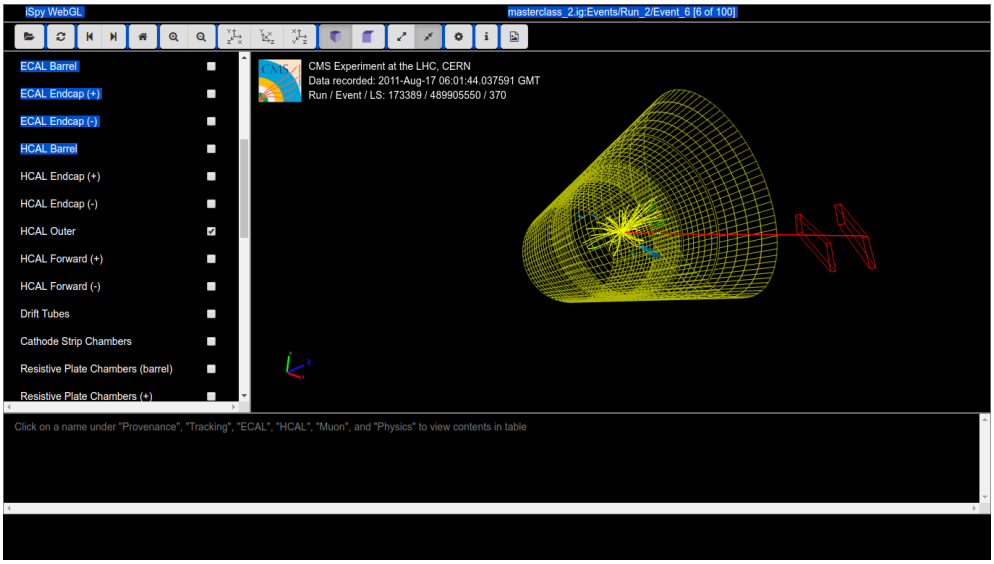


FIGURA 4.4: Evento a través de la interfaz de iSpy.
[99]

en las Clases Magistrales de Física de Partículas, siendo un caso de éxito de la política de acceso a datos abiertos.

4.4. Iniciativa de Preservación desde el IFCA

Gracias a las experiencia adquirida participando en las iniciativas anteriormente expuestas, como el Tier-2 de CMS, EGI, DPHEP o la Política de Datos de CMS (ver 4.3.1), se han llevado a cabo algunas iniciativas de preservación dentro de los diferentes niveles expuestos anteriormente (ver 4.1).

El IFCA es un centro multi-disciplinar, en el que fluyen diferentes lineas de investigación. Los recursos computacionales son transversales a todas las lineas, pudiendo dar soporte a las necesidades de recursos de cada una de ellas. Gracias a ello la infraestructura computacional del IFCA es muy versátil.

A continuación se mostrará una breve descripción de la infraestructura computacional del IFCA y dos casos reales de explotación de iniciativas de preservación en diferentes niveles, desarrolladas gracias a la experiencia adquirida debida a la implicación en los grupos de trabajo de preservación de datos dentro de la colaboración CMS.

4.4.1. La Infraestructura Computacional del IFCA

Describiremos la infraestructura del CPD del IFCA, que nos aportará una visión global de las herramientas computacionales que disponemos, de su magnitud, y de como el empleo de estas herramientas nos capacita para el desarrollo e implementación de diferentes niveles de preservación.

Esta infraestructura podemos dividirla en cuatro bloques; Red, Nodos de Cómputo, Almacenamiento y Gestión (ver fig:3.25).

- **Red**: Es la capa que define cómo las diferentes máquinas y/o servicios se conectan entre si. Según la electrónica que usen, se definen las velocidades de acceso y latencia entre los equipos internos (red privada) y los externos (red pública). Según la capa de red del modelo OSI[100] la red del IFCA se divide en:
  - Ethernet: Constituido por un chasis lógico, integrado por tres switches VDX con 52 interfaces cada módulo (48 interfaces a 1Gbps, 96 interfaces a 10Gbps y doce a 40Gbps), además de un Brocade RX8, con 48 interfaces (24 de 1Gbps y 24 de 10Gbps). Estos equipos forman el core interno de red Ethernet que proporciona conectividad al equipamiento del IFCA. La conexión externa a GÉANT, gestionada por RedIris, se establece mediante un Router Force10 C300 con un enlace de 10Gbps.
  - InfiniBand: 15 equipos Mellanox SX6036 FDR10 40Gbps Mellanox con topología "FAT Tree" sin bloqueo, conectando los nodos HPC y los servidores de almacenamiento. La red Infiniband FDR10 posee unas características específicas, como baja latencia y un ancho de banda muy elevado, útiles en la resolución de problemas que necesitan paralelización. Esta tipo de nodos son capaces de usar RDMA[73] sobre InfiniBand, proporcionando una alta capacidad de I/O sobre determinados sistemas de ficheros.
- **Nodos Computo y/o Servicios**: El CPD del IFCA, dispone actualmente unos 5.000 cores, distribuidos en equipamiento de diferentes fabricantes y pertenecientes a diferentes proyectos. Según su conexión a la capa de red, los diferenciamos en dos:

- Equipos que actúan como hypervisores utilizando Kernel Virtual Machine (KVM) como virtualizador, sobre los cuales ejecutar nodos de computo (nodos Grid y no Grid), servicios Web, Indico, Nextcloud, etc. Actualmente lo forman 300 nodos aproximadamente, conectados mediante red Ethernet de 1-10Gbps. Este equipamiento proporciona aproximadamente unos 2.500 cores de cómputo al IFCA. Todos ellos son gestionados mediante la suit Cloud OpenStack (actualmente en versión Pike).
- Mediante la red Infiniband, otros 156 nodos, con 2.500 cores proporcionan cómputo HPC a la Universidad da Cantabria y a la RES. Estos nodos pertenecen al nodo Altamira (ver 3.4.1), ICTS de la UC.

▪ **Almacenamiento**: El almacenamiento es un servicio crítico en cualquier CPD. Cualquier incidencia, repercute directamente sobre el trabajo de los usuarios. Según su estado, el IFCA gestiona diferentes tipos de almacenamiento:

- **Almacenamiento Online**:

  * Spectrum Scale (GPFS versión 4.2.3): Es un sistema de ficheros posix, de acceso paralelo, distribuido y de alto rendimiento (ver 2.4.2.2). Implementado sobre Cabinas NetApp series E, conectada a cuatro servidores mediante conexiones SAS redundadas de 16Gbps, que reexportan el sistema de ficheros al clúster de cómputo del IFCA, mediante cuatro interfaces Ethernet de 40Gbps y 4 interfaces InfiniBand FDR10 de 40GBps. El sistema de ficheros tiene una capacidad de 1.5PB brutos utilizando 150 discos "Nearline-SAS" de 10TB, además dispone de una cache interna de 20TB sobre discos SSD de alto rendimiento, capaces de absorber más de 400.000 IOPS. Para evitar perdida de datos y de rendimiento ante la reconstrucción de discos de tamaño elevado, un Raid DDP[101] se crea sobre el conjunto de todos los discos. Esto implica una pérdida directa de un 20 % del espacio en bruto, pero minimiza el tiempo de reconstrucción y proporciona redundancia ante la pérdida de discos. Su implementación en el IFCA es singular como vimos en la sección 3.4. El mismo sistema de ficheros esta distribuido

entre todos los nodos cómputo, ya usen la red Infiniband o la Ethernet, siendo todos los equipos capaces de acceder a todo el volumen de almacenamiento, pero cada uno de ellos con rendimientos diferentes. Este es el almacenamiento principal para los datos de las diferentes lineas de investigación del IFCA, así como de los proyectos en los que participa. Es la entrada de datos de los nodos de computo y también el área local de trabajo, "home", de los usuarios locales del CPD del IFCA.

* iSCSI de 10TB: Instalado sobre una cabina Lenovo S2200, con acceso redundado de 10Gbps al núcleo de red del IFCA. Exporta almacenamiento en forma de bloques, donde se instalan las VM de servicios y gestión del CPD del IFCA.

* Ceph (Hammer ver 2.4.2.2): Servidores de disco de diferentes marcas y modelos (IBM, DELL), con una capacidad en bruto de 500TB, configurado en redundancia 1:3. Proporciona almacenamiento elástico, con tres tipos de almacenamiento diferentes dentro de un mismo "middleware", necesarios para satisfacer las necesidades actuales:

  - * Sistema de Ficheros posix: 50TB compartidos entre los hypervisores de gestión del IFCA. Proporcionan un espacio de almacenamiento para copias de seguridad de las máquina de servicios del IFCA. Estas imágenes se almacenan directamente en formato "RAW data" y ante un hipotético fallo del sistema iSCSI, son capaces de iniciarse desde este sistema de ficheros compartido.
  - * Bloque: 200TB de bloque que sirven como almacenamiento persistente de las máquinas Cloud que así lo soliciten.
  - * Objeto: 50TB accesibles mediante protocolo "Swift", que proporciona almacenamiento en modo objeto a los diferentes proyectos Cloud.

- **Nearline**:

  * Robot del cintas magnéticas IBM TS4500: Una capacidad instalada de 1PB, ampliable hasta los 15PB. Con dos lectoras LT07. Este robot se usa como almacenamiento secundario "**nearline**", en combinación con

el software Spectrum Archive. Archive posibilita extender el espacio de nombres del sistema de ficheros Spectrum Scale a dispositivo de cinta, de manera que los metadatos son accesibles. Se produce un movimiento del dato a cinta, según los umbrales definidos en las políticas de migración, pero no del metadato, quedando este accesible desde el sistema de ficheros "Online". Un usuario externo no es capaz, sino es por la latencia de acceso al dato, de saber si este, se encuentra en disco o en cinta.

* Sistema de Copia de Seguridad: Basado en el software opensource Bácula, es implementado sobre un robot de cintas TS3500 de IBM, con dos lectoras LTO5 y una lectora LTO3. Este sistema realiza las copia de seguridad de las máquinas de gestión, de máquinas Cloud de proyectos que lo soliciten y de ciertas áreas definidas del sistema de ficheros de GPFS.

- **Gestión**: El núcleo de servicios del CPD del IFCA, se ejecuta sobre un entorno virtualizado mediante doce equipos, XEN y KVM. Estas doce máquinas importan los bloque iSCSI señalados anteriormente e inicializan de forma automática, cada una de las máquinas de gestión. Cualquiera de ellas puede acceder y gestionar cualquiera de los bloques, aumentando así la disponibilidad y redundancia del sistema (ver fig:4.5).

Gracias a la infraestructura proporcionada por el CPD del IFCA (ver fig:3.25) y a la experiencia adquirida, el grupo de computación del IFCA se encuentra actualmente capacitado para ofrecer nuevas posibilidades de recursos computacionales a la comunidad científica. Dichos recursos son:

- Sistema de transferencia masiva de Archivos.
- Almacenamiento de datos en Cinta (NearLine).
- Almacenamiento de datos en Disco (Online).
- Recursos Computacionales sobre HPC.
- Recursos Computacionales sobre Cloud.

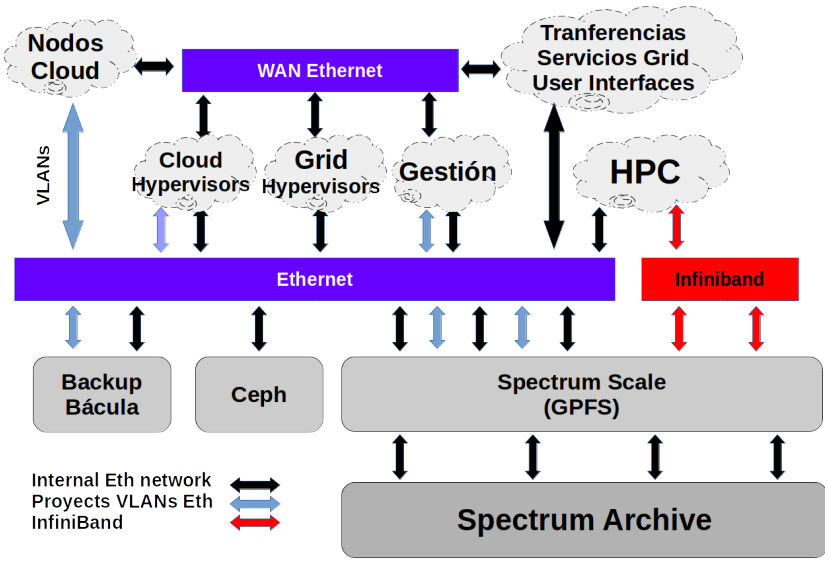


FIGURA 4.5: Descripción lógica de los servicios computacionales del IFCA.

El IFCA proporciona un amplio catálogo de recursos computacionales tanto a proyectos internos, del propio instituto, como a otras instituciones públicas como el Consejo Superior de Investigaciones Científicas (CSIC), universidades, museos o incluso a empresas del sector privado siempre que demuestren actividades de interés científico. Estos servicios quedan disponibles a través del "Catálogo de Servicios del CSIC" (ver fig:4.6).

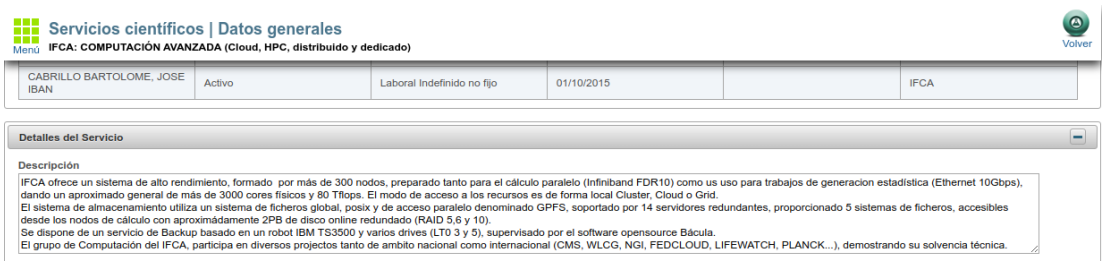


FIGURA 4.6: Vista del Catálogo de servicios del CSIC ofrecido por el IFCA.

4.4.2. Preservación de Bit del Proyecto Cabas en el IFCA

A principios del año 2014, parte de la Biblioteca Digital del CSIC, tenía que ser migrada por requerimientos de espacio. En ese momento el servicio TIC del CSIC, necesitaba liberar espacio en su infraestructura, y el espacio ocupado por la Biblioteca, aproximadamente unos 40TB, debía ser trasladado. Conociendo los recursos informáticos del IFCA, se propuso la posibilidad de almacenar y preservar estos datos en las instalaciones de nuestro CPD y ponerlos al servicio del CSIC. La experiencia de preservación, se estableció en partes claramente diferenciadas.

4.4.2.1. Transferencia de Datos

El proyecto comtempla la transferencia de archivos desde las instalaciones de TIC del CSIC en Madrid al IFCA situado en Santander. La transferencia de datos entre ambos sitios, presentaba varias dificultades.

La primera establecer un protocolo de transferencia soportado por ambas partes. Utilizar Gridftp (ver 2.4.2.3), por ambas partes, hubiera sido la solución tecnológica ideal, dadas sus múltiples ventajas, pero no es habitual que instituciones ajenas al ámbito Grid, estén familiarizadas con este tipo de tecnologías. Las posibilidades se limitaron a dos: usar rsync sobre un túnel ssh o vsftp. Rsync tiene la ventaja de que es capaz de continuar las transferencias desde el punto de interrupción, y además puede realizar comprobaciones de integridad, pero es extremadamente lento y dado el volumen de datos a transferir, no era viable. Por ello la opción escogida fue usar el protocolo File Transfer Protocol (FTP) implementado mediante el servicio Very Secure FTP (VSFTP).

La segunda, el volumen de datos a transferir, 40TB. El IFCA disponía de 2.5Gbps de ancho de banda soportado por ocho servidores configurados en Round-Robin para la gestión de la carga, frente a la conexión de 1Gbps de CSIC. Se decidió que las transferencias fueran nocturnas, evitando así las posibles interferencias durante los horarios de oficina. La tarea se completó en aproximadamente 30 días.

Los servicios del IFCA se configuraron para permitir este tipo de acceso, que hasta ese momento no estaba soportado, instalando servidores VSFTP sobre las ocho máquinas utilizadas por los experimentos como CMS para transferencia de datos. Los usuarios son autorizados en el servidor, mediante el servicio de gestión

de usuarios del IFCA basado en freeipa[74] y el demonio System Security Services Daemon (SSSD) ejecutado localmente en cada máquina.

4.4.2.2. Integridad y Verificación

Una vez que los datos se encontraban almacenados en las instalaciones del IFCA, se generan los valores "md5sum" de cada fichero descargado y se compara con sus valores en el sistema de origen, de modo que establecemos la confianza en la transferencia. Queda comprobado que el origen y el destino son iguales. Para este propósito, se creó un ejecutable python (ver apéndice 01), que más tarde fue extendido para dar soporte a todo el sistema de ficheros posix. Actualmente, este ejecutable es usado por la colaboración CMS, durante las campañas de "monitorización de espacio" y "consistencia de datos" (ver fig:4.7).

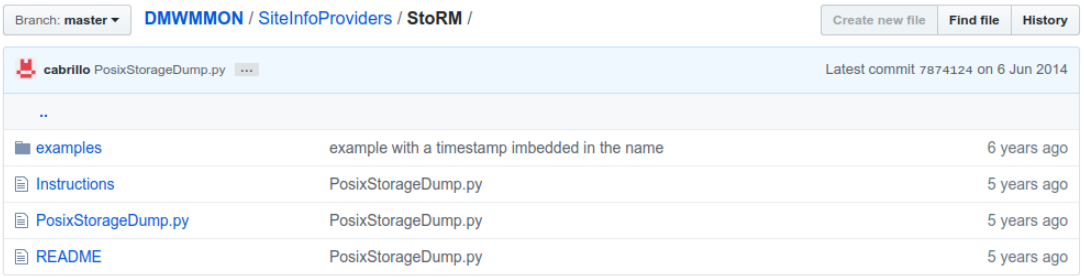


FIGURA 4.7: Fichero Posix Dump.

4.4.2.3. Preservación

Mediante el software de copias de seguridad Bácula[102], utilizado por el IFCA, se creo un nuevo catálogo para preservación, definiendo el periodo de caducidad a 15 años para los registros de esta nueva base de datos. Además, para evitar posibles pérdidas por re-escritura, se emplearon cintas LTO5 Write Ones Read Many (WORM), de modo que el dato una vez escrito no puede ser borrado. Dado el volumen de datos para la época, las operaciones de archivado se dividieron en lotes de 1TB aproximadamente, programando copias incrementales sobre cada lote de datos, pudiendo llegar a tardar entre 3∼4 días cada operación. El fichero con los valores de suma de verificación de cada lote era archivado en cada proceso.

Cada varios meses, los datos son leídos al azar y sus valores md5sum comprobados, verificando así el estado de las cintas y los datos.

4.4.2.4. Acceso

Para poder acceder a los datos, otra copia se mantiene "Online". Mediante los servidores VSFTP, los datos pueden ser almacenados y/o copiados desde exterior. Mediante el sistema "User Interface" mencionado anteriormente (ver 2.3.4.1), accediendo a través del protocolo ssh, los datos quedan accesibles, de forma que los usuarios de la Biblioteca pueden tratarlos localmente desde nuestro sistema (ver fig:4.8).

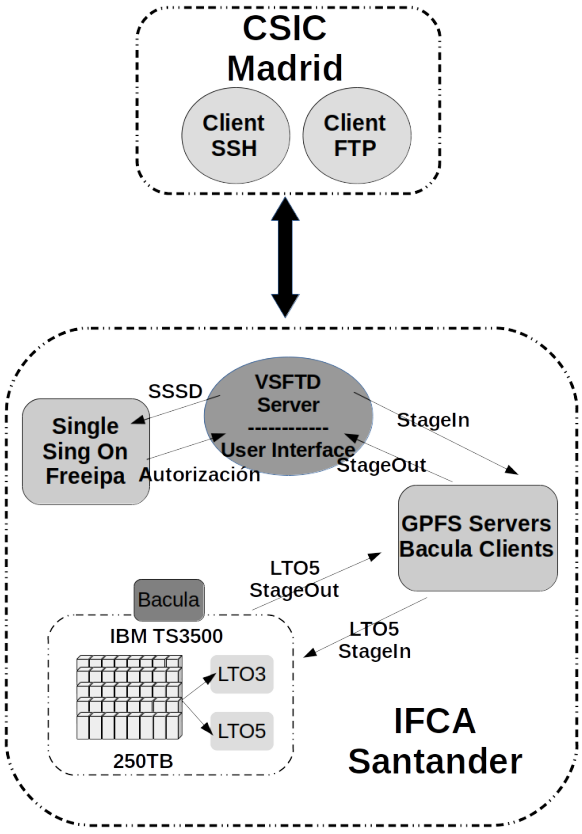


FIGURA 4.8: Implementación de la Preservación de Datos Bit de la biblioteca Digital del CSIC.

4.4.2.5. Cambio de Medio

El sistema de almacenamiento del IFCA fue renovado por completo en el año 2017. Durante éste proceso y debido a la incompatibilidad entre el sistema antiguo en ejecución (GPFS 3.4) y el actual (Spectrum Scale 4.3.2), se realizó una migración completa entre ambos. Alrededor de 1PB de datos fueron trasladados de un sistema a otro. Esta operación duró aproximadamente dos meses. Los datos de la biblioteca fueron migrados y la copia Online trasladada al nuevo sistema Spectrum Scale automáticamente. Actualmente el proyecto sigue activo, con algo más de 60TB de datos en dos copias completas "Online" y "Offline".

4.4.3. Iniciativa Opendata en el IFCA

Como hemos visto, CMS tiene una política de preservación de datos, reutilización y acceso abierto (ver 4.3.1), donde se define el enfoque de CMS para la preservación de los datos y el acceso a ellos en varios niveles de complejidad. La preservación de un análisis y sus datos correspondientes requiere una comprensión completa de los pasos involucrados en el mismo para producir los resultados, desde los datos en bruto, hasta la publicación. Para ello, se desarrolló y desplegó un marco de análisis global para preservar todo el conocimiento adjunto a esos análisis y datos. Como caso de uso para llevar acabo la experiencia de preservación[103], se tomo el siguiente ejemplo: La medición de la sección transversal de "ttbar" en el canal de dilepton a $\sqrt{s} = 8$TeV, realizada con 5.3fb de datos recopilados en 2012[104]. En algún momento en el futuro, acceder a estos análisis y datos implicará utilizar sistemas operativos obsoletos y versiones de sotfware no compatibles, que también se conservarán en un ambiente virtual basado en infraestructura Cloud del IFCA.

Una vez que un análisis es publicado, existe un alto riesgo de perder el conocimiento y la experiencia que hizo falta para llevarlo a cabo. CMS registraba la mayoría de los metadatos inmediatos: números de eventos y ejecuciones, condiciones de haz, software, versiones utilizadas en el reprocesamiento de datos, etc, pero no registraba la información práctica necesaria para poner los datos en contexto y analizarlos. Con la adopción de la política de preservación, la colaboración se compromete a buscar soluciones que permitan el registro de estos metadatos de

contexto. Es una tarea pesada, pero comenzar la preservación de los datos de análisis durante los momentos de actividad podría ayudar a preservar sus detalles mientras el conocimiento está aún reciente. CMS proporcionó el entorno necesario para ejecutar los diferentes análisis en forma de imágenes de máquinas virtuales desarrolladas por CERN para Experimentos LHC (CernVMs).

La experiencia del IFCA, se centrará en intentar comprobar el acceso y el entorno de ejecución de estás máquinas virtuales, usando la infraestructura Cloud basada en OpenStack. Siguiendo todos los pasos que un usuario interesado debería seguir, desde la configuración inicial de la infraestructura, hasta la ejecución de análisis con datos de 2010. Para llevar a término esta experiencia fue necesaria la personalización de las imágenes de CernVM, la creación de instancias de las máquinas virtuales y dotarlas del acceso al software necesario para ejecutar el análisis.

4.4.3.1. Preservación del Análisis

Como hemos visto anteriormente, CMS se organizan en una jerarquía de niveles de datos (ver 3.2.1) y garantiza que las condiciones de toma de datos sean eficientes y de calidad, mediante los marcos DQM que vimos anteriormente (ver 3.3.1), mientras que los objetos de física son revisados y aprobados por el POG.

Los datos RAW necesarios para llevar a cabo esta experiencia, se reconstruyeron con la versión de software CMSSW_5_2. Los datos RECO y las muestras fueron reconstruidos con CMSSW_5_3. Ambas versiones ejecutadas bajo Scientific Linux 5.

Para cualquier análisis necesitamos comparar la teoría con los datos medidos. Para ello se intenta igualar, tanto como sea posible las muestras MC, con las condiciones de datos fuera de línea, es decir, el campo magnético, el conocimiento de la alineación del detector, las últimas calibraciones del detector, etc. Todas estas condiciones, son almacenadas en una base de datos, y quedan identificadas y accesibles mediante una etiqueta global.

Diferentes grupos divergirán y tendrán diferentes enfoques para realizar el análisis. Pero la idea básica es siempre la misma: pasar del formato más pesado que contiene todos los eventos y toda la información de los mismos, a un formato

de datos más ligero que contendrá un subconjunto de eventos y solo la información de aquellos con interés para el análisis a realizar. Usando la Infraestructura de computación Grid de CMS (CRAB) y asignando la salida de los datos al Storage Element (ver 2.3.5) del IFCA, obtuvimos los ficheros "ROOT TTree" de nuestro análisis. A partir de los "ROOT TTre", obtenemos los histogramas, tablas con resultados significativos para el análisis y otros ficheros "ROOT TTree" más pequeños, en tamaño y complejidad, que contendrán solo la selección final de eventos.

4.4.3.2. Publicación de los Resultados

Los documentos públicos e internos de CMS son accesibles desde iCMS[105]. El primer paso para llegar a la publicación es obtener la pre-aprobación del análisis por el Physics Analysis Group (PAG). Después de la aprobación previa, el Comité de Revisión de Análisis (ARC) se hace cargo del proceso: una presentación oral seguida de una discusión detallada de los resultados, preguntas, comentarios y respuestas. A través de la Collaboration Wide Review (CWR) toda la colaboración examina el procedimiento y resultados, estos son almacenados en CDS[106] de forma privada, solo accesible para la colaboración. Muchos grupos de diferentes instituciones intercambian preguntas, comentarios y respuestas con los autores correspondientes. Una vez que todos los problemas se abordan adecuadamente, el siguiente paso es la lectura final de la publicación, seguida de la última: la presentación a una revista revisada por pares. A cada análisis le es asignado una entrada en la Analysis Database Interface (CADI) de CMS[107]. Una vez que la documentación es publica, se proporciona el correspondiente acceso mediante CERN Document Server (CDS) o arXiv (ver fig:4.9).

4.4.3.3. Recuperación del Entorno contextualizado

El IFCA, dispone de una Infraestructura como servicio (IaaS) accesible mediante un Cloud privado gestionado durante la realización de esta experiencia por OpenStack, Havanna, y Pike como versión actual en ejecución. El acceso a lo recursos puede realizarse mediante el dashboard que proporciona OpenStack, accesible desde la dirección **https://portal.cloud.ifca.es** o mediante comandos

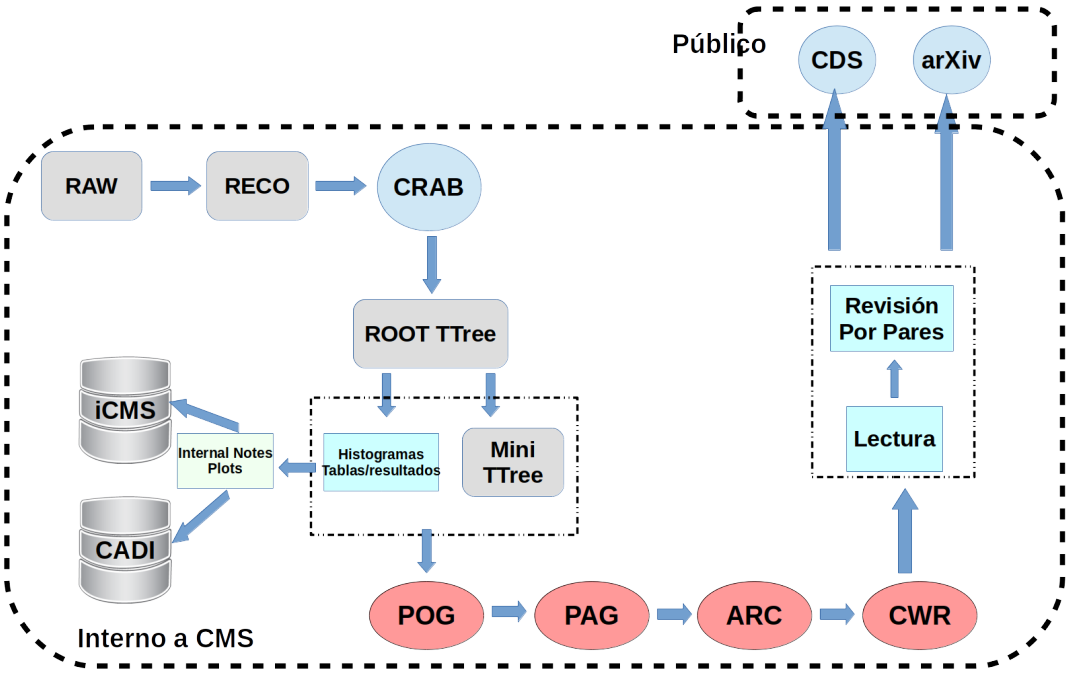


FIGURA 4.9: Flujo de desde la toma de datos hasta la publicación.

OpenStack desde la propia linea de comandos.

Un nuevo "Tennat" (proyecto/grupo en terminología OpenStack) llamado "Ifca.es:preservation" se creó en la infraestructura OpenStack local del IFCA. La configuración del "proyecto" originalmente consistía en crear un grupo de seguridad predeterminado, permitiendo solo conexiones ICMP y SSH. Para evitar acceso mediante contraseña en las máquinas virtuales, se creó un par de claves ssh. La clave pública se registra en la infraestructura de OpenStack y es inyectada de forma automática en máquinas virtuales cuando éstas son creadas.

Las imágenes virtuales, proporcionadas por el CERN, las CernVM, no se ejecutaban en el hypervisor Xen v4.0.1 (.hdd) y el resto de formatos de imagen proporcionados por el CERN (vmdk, vdi y ova) no eran compatibles con la versión de OpenStack ejecutada durante el desarrollo de la experiencia.

Después de una actualización a Xen v4.4, se creó una nueva imagen de la CernVM tomando la imagen ".hdd" y actualizando el gestor de arranque (grub). Esta imagen arrancó y la inicialización de la imagen CernVM fue correcta. Sin

embargo, la imagen fallaba durante la contextualización en la configuración de CVMFS, dejando así la misma inutilizable. Descubrimos varios fallos:

- Un cambio en "amiconfig", que introdujo un error de sintaxis en Python 2.4 (versión utilizada en SLC5); "amiconfig" entraba en excepción.
- Otro error impedía que el plugin de "amiconfig" que se encarga de configurar varias Opciones de la CernVM, como cuentas de usuario, proxies CernVM-FS, servicios vicios, etc.

Solventados los fallos, la CernVm pudo ejecutarse y contextualizarse correctamente, configurando el CVMFS. Se creó así una nueva imagen disponible para todos los usuarios, denominada como "IFCA CernVM (1.17-5 SL5)". Las CernVMs, proporcionan un instalación mínima; durante el primer inicio se autoconfigura tomando su forma final gracias a ficheros de contextualización. Estos ficheros de contexto, son ficheros de texto plano, pasados a la VM durante su inicialización. A través de la interfaz web de CernVM Online[108] estos ficheros pueden ser creados (ver fig:4.10) desde el dashborad de OpenStack.

Para iniciar la nueva instancia, pulsamos el botón "Lunch" (ver fig:4.11), rellenando las siguientes opciones mínimas:

- Details tab: Availability zone: nova
- Instance name: Nombre que le damos a la VM.
- Flavor: Recursos predefinidos a nivel de hardware que solicitamos a la IaaS (nº CPUs, RAM, HDD, Ephimeral). De las posibles que el IFCA dispone, se elige el flavor: m1.medium (2 cores, 4GB RAM, 40GB of disk), suficientes para el propósito de este análisis.
- Boot Source → Boot from image → Image (Seleccionamos IFCA CernVM (1.17-5 SL5)).
- Access & Security tab: Seleccionamos la clave pública que será embebida en la imagen.
- Post-Creation tab: Seleccionamos el fichero obtenido desde la interfaz web de CernVM Online.

FIGURA 4.10: Servicio Cern VM Online.

Una vez que la VM se está ejecutando, el siguiente paso es proporcionarle una dirección externa, para poder acceder desde el exterior. Para ello seleccionamos "Associate Floating IP". Esto nos proporciona la posibilidad de acceder a la máquina a través de ssh, de otra manera no podríamos acceder, ya que a cada instancia iniciada, se le asigna una IP privada dentro del rango de una vlan definida para el proyecto, solo accesible desde el portal Cloud, para los usuarios del mismo.

Por defecto, las máquinas virtuales no contienen ningún almacenamiento persistente, por lo que cuando la VM se destruye, cualquier cambio escrito por

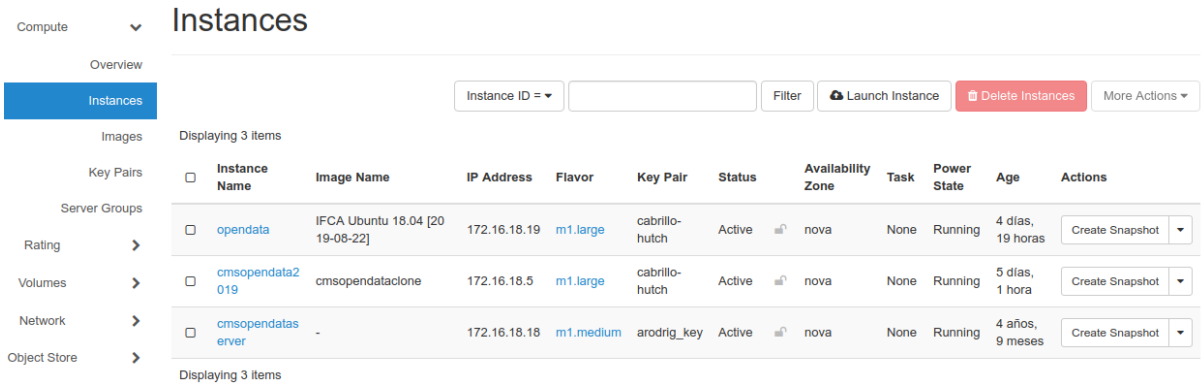


FIGURA 4.11: Portal OpenStack: Sub-menu Instancias.

los usuarios se pierde. OpenStack proporciona la opción de añadir bloques de almacenamiento externo a las máquinas virtuales (ver fig:4.12), proporcionando un almacenamiento persistente que puede ser conectado a VM en ejecución para almacenar el análisis de usuario y datos. Esto es posible gracias a la integración de Ceph (ver 2.4.2.2) y OpenStack, dando acceso a los nodos de computo definidos en OpenStack al BE de almacenamiento Ceph. El nuevo dispositivo es mostrado a la máquina en ejecución como un "dispositivo" de bloque en la siguiente unidad libre (xvdc/d... en formato XEN, en estos momentos el hypervisor empleado en los nodos de computo es KVM, por lo que los dispositivos de bloque tienen el formato vdc/d/..), al que hay que dar formato y montar en un directorio para ser accesible por los usuarios.

4.4.3.4. Solución implementada por el IFCA

Fruto de esta experiencia, el IFCA ha implementado un entorno funcional de datos abiertos de la colaboración CMS, orientado a formación, mediante la infraestructura Cloud existente en el IFCA. Una nueva máquina "https://opendata.ifca.es:8000", es utilizada para tareas educativas y/o de formación, siendo empleada para ejercicios prácticos en la asignatura de cuarto curso de grado Física de la UC. La practica consiste en realizar un análisis de física de partículas, empleando datos reales de la colaboración CMS tomados durante el año 2009 y disponibles públicamente a través de la iniciativa opendata del CERN. El análisis se basa en la medida de la masa del bosón Z. La prática es ejecutada a lo largo de varios

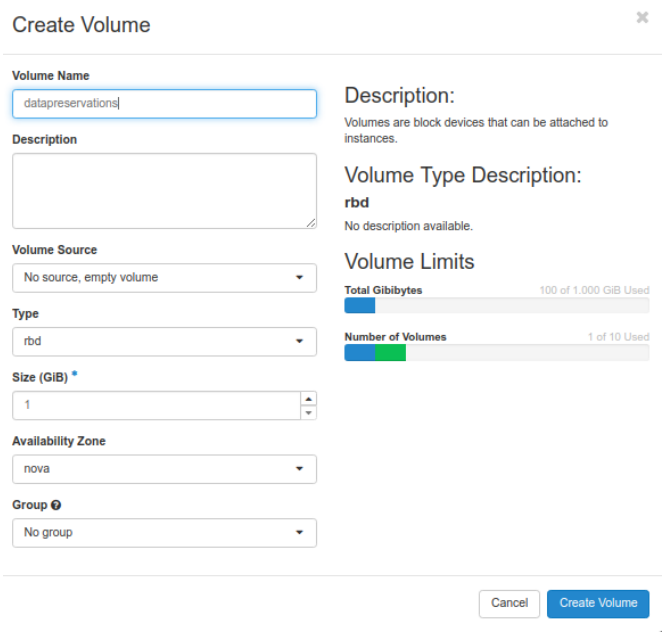


FIGURA 4.12: Portal OpenStack: Sub-menu Volumes.

jupyter notebooks, que se utilizan como base durante las explicaciones teóricas del profesor. Los notebooks documentan la metodología de análisis y a su vez se utilizan como método de evaluación, puesto que al final de la practica el alumno entrega los notebooks con los ejercicios resueltos. El software utilizado para realizar el análisis es ROOT compilado bajo python3, ejecutando imágenes Docker de Ubuntu 18.04 (ver fig:4.13) con el software de JupyterHub[109] preinstalado sobre la propia máquina virtual.

Uno de los mayores problemas al que nos enfrentamos en esta experiencia, es el acceso a datos en gran cantidad y de gran tamaño empleando la red, sobre todo a la hora de realizar ejercicios en vivo, como es el contexto actual. Aunque el protocolo Xrootd intenta solventar este problema, tiene sus limitaciones ya que aumenta los tiempos de ejecución y disminuye la eficiencia de los trabajos. Otra de las posibilidades que se barajaron, era proporcionar un acceso directo al sistema de ficheros del centro Tier-2, pero esto implicaba problemas de seguridad, al incluir una máquina dentro de la red de datos de IFCA.

Las nuevas versiones de OpenStack, nos proporciona nuevas posibilidades, pudiendo evitar las limitaciones que teníamos hasta este momento. Mediante

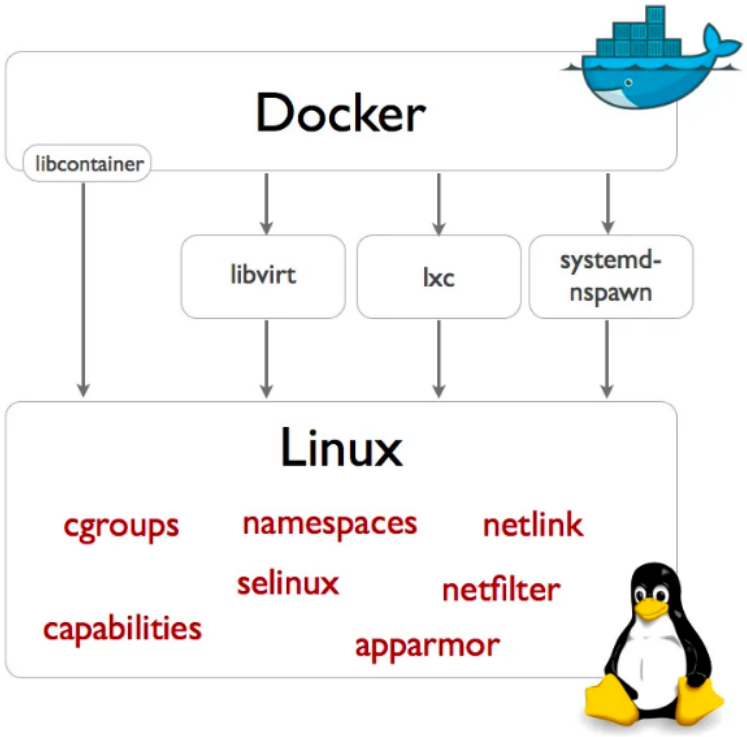


FIGURA 4.13: Descripción servicio Docker.
[110]

el servicio "Manila"[111], OpenStack es capaz de gestionar partes definidas del sistema de fichero Spectrum Scale y exportarlos mediante protocolo CIFS o NFS de manera redundante (Ganesha), de manera que el acceso a los datos sería local, evitando así las latencias a las que estaríamos sometidos usando protocolos como Xroot (ver tab:4.3).

Los metadatos de contexto (CVMFS), pueden ser accedidos desde los proxies instalados en el Tier-2, para los nodos de cómputo Grid, pero en este caso accedidos a través de la red pública del IFCA (ver tab:4.4).

Mediante el uso de "Dockers" o "Contenedores" unidos a herramientas como "JupyterHub" y "SystemdSpawner"[112] obtenemos un entorno encapsulado de experimentación ideal, para este tipos de entornos formativos, proporcionando aislamiento y seguridad a la infraestructura subyacente (ver fig:4.14), a la vez que se proporcionan las herramientas necesarias para que un grupo nutrido de estudiantes realicen tareas de análisis en un entorno casi real (ver tab:4.5).

TABLA 4.3: Configuración del espacio GPFS+NFS mediante el servicio Manila.

```
manila:~# manila create NFS 50 --name opendata

manila:~# manila list
+--------------------------------------+------------+------+-----------------+-----------+
| ID                                   | Name       | Size | Share Proto     | Status    |
+--------------------------------------+------------+------+-----------------+-----------+
| 70a48a78-d70f-470d-a345              | opendata   | 50   | NFS             | available |
+--------------------------------------+------------+------+-----------------+-----------+

manila:~ #  manila show 70a48a78-d70f-470d-a345
+-----------------------------+-----------------------------------------------------------+
| Property                    | Value                                                     |
+-----------------------------+-----------------------------------------------------------+
| id                          | 70a48a78-d70f-470d-a345                                   |
| size                        | 50                                                        |
| availability_zone           | nova                                                      |
| created_at                  | 2019-10-04T08:15:00.000000                                |
| status                      | available                                                 |
| name                        | opendata                                                  |
| description                 | None                                                      |
| project_id                  | 276c19f252d24f0ebe0aa9ff39307f56                          |
| snapshot_id                 | None                                                      |
| share_network_id            | None                                                      |
| share_proto                 | NFS                                                       |
| metadata                    | {}                                                        |
| share_type                  | b0fdd0ff-248d-41c4-a549-947d866fd0ac                      |
| is_public                   | False                                                     |
| snapshot_support            | False                                                     |
| task_state                  | None                                                      |
| share_type_name             | default_share_type                                        |
| access_rules_status         | active                                                    |
| replication_type            | None                                                      |
| has_replicas                | False                                                     |
| user_id                     | 6e57d91da967475b887ef485ca4db611                          |
| create_share_from_snapshot  | False                                                     |
| revert_to_snapshot_support  | False                                                     |
| share_group_id              | None                                                      |
| source_share_group_snapshot | None                                                      |
| mount_snapshot_support      | False                                                     |
| share_server_id             | None                                                      |
| host                        | manila.cloud#GPFS                                         |
| export_locations            | id = fdb0638b-912b-4dbd-9f36-6489190886d0                 |
|                             | path=172.16.18.252:/gpfs/ces/share-c014bf8d-ccc6          |
|                             | preferred = False                                         |
|                             | share_instance_id = c014bf8d-ccc6                         |
|                             | is_admin_only = False                                     |
+-----------------------------+-----------------------------------------------------------+

[root@gpfs03 ~]# ifconfig eth1.2018
eth1.2018:0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
   inet 172.16.18.252  netmask 255.255.255.0  broadcast 172.16.18.255
   ether ac:1f:6b:b2:25:df  txqueuelen 1000  (Ethernet)

[root@gpfs03 ~]# mmces address list

Address          Node                    Group       Attribute
-------------------------------------------------------------
10.10.4.13       gpfs03.ifca.es          none        none
10.10.4.14       gpfs04.ifca.es          none        none
172.16.18.252    gpfs03.ifca.es          none        none
172.16.18.253    gpfs04.ifca.es          none        none

[root@gpfs03 ~]# mmnfs export list
Path                                          Delegations  Clients
-------------------------------------------------------------------
/gpfs/ces/share-c014bf8d-ccc6                 NONE         172.16.18.6/32
```

Tabla 4.4: Acceso a CVMFS y GPFS desde la máquina Opendata.

```
root@opendata:~# df -h
Filesystem                          Size  Used Avail Use% Mounted on
udev                                3.6G     0  3.6G   0% /dev
tmpfs                               730M  708K  729M   1% /run
/dev/vda1                            29G  5.5G   24G  20% /
tmpfs                               3.6G     0  3.6G   0% /dev/shm
tmpfs                               5.0M     0  5.0M   0% /run/lock
tmpfs                               3.6G     0  3.6G   0% /sys/fs/cgroup
/dev/vda15                          105M  3.6M  101M   4% /boot/efi
/dev/vdb                             20G   44M   19G   1% /mnt
tmpfs                               730M     0  730M   0% /run/user/0
cvmfs2                               25G  5.1M   25G   1% /cvmfs/cms.cern.ch
cvmfs2                               25G  5.1M   25G   1% /cvmfs/grid.cern.ch
172.16.18.252:/gpfs/ces/share-c014b  50G     0   50G   0% /data
```

Tabla 4.5: Acceso a CVMFS y GPFS+NFS desde la imagen Docker.

```
jovyan@05fd0b338804:~$ df -h
Filesystem                          Size  Used Avail Use% Mounted on
overlay                              29G  5.5G   24G  20% /
tmpfs                                64M     0   64M   0% /dev
tmpfs                               3.6G     0  3.6G   0% /sys/fs/cgroup
shm                                  64M     0   64M   0% /dev/shm
cvmfs2                               25G  5.1M   25G   1% /cvmfs/cms.cern.ch
cvmfs2                               25G  5.1M   25G   1% /cvmfs/grid.cern.ch
/dev/vda1                            29G  5.5G   24G  20% /home/jovyan
172.16.18.252:/gpfs/ces/share-c014b  50G     0   50G   0% /home/jovyan/share
tmpfs                               3.6G     0  3.6G   0% /proc/acpi
tmpfs                               3.6G     0  3.6G   0% /proc/scsi
tmpfs                               3.6G     0  3.6G   0% /sys/firmware
```

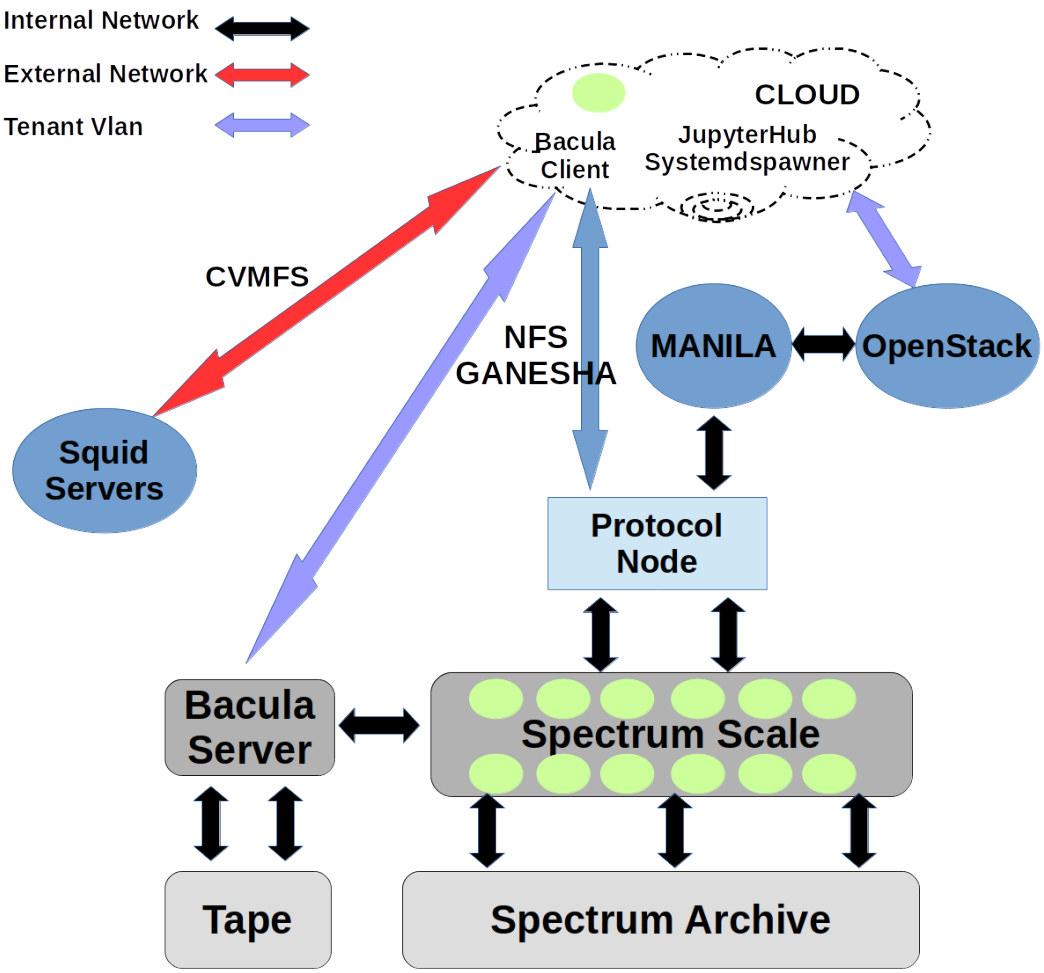


FIGURA 4.14: Esquema de Implementación de mejoras en el servicio de CMS Opendata del IFCA.

Conclusiones

Se ha implementado un Tier-2 en producción, dando cobertura a las necesidades de movimiento de datos, produccion de MC, análisis y almacenamiento, en el que he participado destacadamente como responsable de los sistemas de red, almacenamiento y del Tier-2 de CMS para el IFCA. Todo ello dentro de los compromisos adquiridos por el IFCA con la colaboracion Compact Muon Solenoid (CMS), contribuyendo al anuncio el 4 de Julio del 2012 de la observación de una nueva partícula "consistente con el **Bosón de Higgs**".

Se ha diseñado e integrado una infraestructura HPC, pudiendo ser empleada como recurso oportunístico por los usuarios locales del Tier-2 de CMS, proporcionando un recurso extra durante elevadas cargas de trabajo, o ante el requerimiento no programado, de análisis pesados en cortos periodos de tiempo.

Se ha presentado el diseño de una infraestructura de red singular gracias a las sinergías que confluyen en el CPD del IFCA. Fruto de este trabajo, se realizó una presentación en CHEP2013, mostrando la integración de los recursos del experimento CMS y el computador HPC de la Universidad de Cantabria, Altamira, que permite a los usuarios de la colaboración el uso puntual de recursos computación de alto rendimiento (HPC) normalmente no accesibles para ellos.

Se han realizado mejoras sobre la infraestructura anteriormente citada. Se posibilita el uso de recursos HPC de forma segura y transparente para los usuarios Grid, empleando cores libres de máquinas en uso, rentabilizando de este modo el Supercomputador Altamira.

Se habilita la opción de que usarios autorizados, ejecuten tareas instensivas sobre recursos HPC en modo Grid. El uso de estos recursos será contabilizado por las respectivas VOs, generando valores evaluables a la hora de justificar la necesidad de solicitud de nuevos recursos para el experimento.

Se ha participado activamente en las inicitivas de preservación presentadas en esta tesis, dando lugar a la publicacion de articulos en revistas como "**Implementing the data preservation and open access policy in CMS**" en el Journal of Physics: Conference Series, Volume 513, Track 4 y a la presentación de un poster en la conferencia de Infraestructure Grid Europea (EGI) en el año 2015, en ISCTE-IUL Portugal "**Implementation of the IFCA CMS Open Data Portal using EGI FedCloud resources**"

La experiencia adquirida, a través de la participación en diferentes grupos de trabajo en el ámbito de la preservación, llevados a cabo dentro de la colaboración CMS, ha posibilitado el desarrollo de iniciativas propias dentro del IFCA. Ejemplo de ello son los casos presentados de Preservación a Nivel de Bit (ver 4.4.2) y Open Access (ver 4.4.3).

En conclusión, esta tesis ofrece una solución adecuada y probada para la preservación de datos, desarrollada en el contexto del experimento CMS del LHC del CERN, que va mas alla de la simple preservación del bit. Utilizando tecnicas cloud, desplegadas en entornos de computación de alto rendimiento, y adaptables a otros muchos entornos de grandes colaboraciones científicas, permitiendo avanzar hacia la preservación del conocimiento.

Agradecimientos

En primer lugar deseo expresar mi agradecimiento a mis directores de tesis doctoral, Dra. Alicia Calderón y el Profesor de investigación Jesús Marco de Lucas, por la dedicación y el continuo apoyo que me han brindado durante la realización de este trabajo y por la dirección y el rigor que ha facilitado a la misma.

Agradezco todos mis compañeros del IFCA, especialmente al Grupo de Computación Avanzada por su apoyo personal y humano, por obligarme siempre a reciclar mis conocimientos. Este trabajo de investigación es fruto de ideas, proyectos y esfuerzos previos que corresponden a otras personas. En este caso mi más sincero agradecimiento a Miguel Ángel Nuñez e Aida Palacio, con quien he compartido el día a día durante estos años, al Dr. Fernando Aguilar por proporcionarme su tesis, a la Dra. Lara Lloret y al Dr. Pablo Martinez, por atender mis consultas y facilitarme parte de su conocimiento, su tiempo y sus ideas, y al Dr. Jordi Duarte por su predisposición a la hora de abordar nuevas iniciativas. Por su orientación, corrección y atención a mis consultas sobre metodología, mi agradecimiento al Dr. Rafaél Rodriguez.

Gracias a mis amigos (Angel, Alberto y Fernando), que siempre me han prestado un gran apoyo moral y humano, necesarios en los momentos difíciles de este trabajo y esta profesión. Pero sobre todo, gracias a mi mujer Cristina, mi inagotable editora, y a mis hijas, Lara, Elsa y Jimena por su paciencia, comprensión y solidaridad con este proyecto, por el tiempo que me han concedido, un tiempo robado a la historia familiar. Sin su apoyo este trabajo nunca se habría escrito y, por eso, este trabajo es también el suyo. A todos, muchas gracias.

Índice de figuras

Índice de cuadros

Apéndice 01

A continuación se muestra el código empleado para la generación de los valores checksum, file size y timestamp. Inicialmente nació para comprobar la integridad de los datos del proyecto de preservación de la Biblioteca Digital del CSIC, y posteriormente fue refinado para las campañas de monitorización de espacio y consistencia cíclicas de CMS. Los sistemas posix pueden almacenar los valores checksum, file size y timestamp, como un atributo extendido. Este codigo busca si estos valores están definidos y si no, los calcula y almacena.

```
############## PosixStorageDump.py ###############
#!/usr/bin/python

'''
Posix StorageDump v1.5
Multiprocces
Support for adler32, md5 and sha256 hash.
Silent mode
Python > 2.5
author: Iban Cabrillo
'''

import os
import sys
import glob
import time
import hashlib
import datetime
import argparse
```

```
import fileinput
import subprocess
from multiprocessing import Pool

def hashfile(file, hash, blocksize=65536):
    '''
    Calc the md5 or sha256 (pass in hash value) chksum value
       for a giving file
    '''
    if hash == 'md5':
        hasher = hashlib.md5()
    elif hash == 'sha256':
        hasher = hashlib.sha256()

    with open(file, 'rb') as f:
        for block in iter(lambda: f.read(blocksize), ""):
            hasher.update(block)
    return hasher.hexdigest()


def hashadler32(file, blocksize=65536):
    '''
    Calc the adler32 chksum value for a giving file
    '''
    from zlib import adler32

    val = 1
    fp=open(file)
    while True:
        data = fp.read(blocksize)
        if not data:
            break
        val = adler32(data, val)
        if val < 0:
            val += 2**32
```

```
    return hex(val)[2:10].zfill(8).lower()
def get_local_cksum(surl):
    '''
    Get the cksum store at local file level. If the file has no
        cksum value we should calc it.
    '''
    if hash == 'adler32':
        name = 'user.storm.checksum.adler32'
    else:
        name = 'user.checksum.%s' % hash

    print surl
    output, error = subprocess.Popen(['getfattr', '--only-
       values', '--absolute-names', '-n', name, surl],
    stderr=subprocess.PIPE, stdout=subprocess.PIPE).communicate
       ()
    #print 'output:[',output,']'

    if len(output) == 0:
        if hash == 'adler32':
            value = hashadler32(surl).rstrip('\n').lstrip('0')
            #Makes name avaible to be used under StoRM
        else:
            value = hashfile(surl, hash).rstrip('\n')

        if verb:

            print "No checksum value found for file %s.
               Processing..." % surl.replace(localpath,'')
            #Calc de adler32 value
        setcksum, error = subprocess.Popen(['setfattr', '-n',
           name, '-v', value, surl],
        stderr=subprocess.PIPE, stdout=subprocess.PIPE).
           communicate()
        return value
```

```
    else:
        return output.rstrip('\n')

def get_local_timestamp(surl):
    '''
    Get the mtime store for local file as extra attribute. If
       the file has no this
    value stored we calc it.
    '''

    #Look for the adler32 value
    output, error = subprocess.Popen(['getfattr', '--only-
       values', '--absolute-names', '-n', 'user.timestamp',
       surl],
    stderr=subprocess.PIPE, stdout=subprocess.PIPE).communicate
       ()
    #print 'output:[',output,']'
    #print 'error:[',error,']'

    if len(output) == 0:

        if verb:

            print "No timestamp value found for file %s.
               Processing..." %surl.replace(localpath,'')

        #Calc de timestamp value
        timestamp = str(os.stat(surl).st_ctime).rstrip('\n')

        # Set the timstamp value for the file.
        settimestamp, error = subprocess.Popen(['setfattr', '-n
           ', 'user.timestamp', '-v', timestamp, surl],
        stderr=subprocess.PIPE, stdout=subprocess.PIPE).
           communicate()
```

```
        return timestamp

    else:

        return output.rstrip('\n')


def get_local_size(surl):
    '''
    Get the size store for local file as extra attribute. If
       the file has no this
    value stored we calc it.
    '''

    #Look for the adler32 value
    output, error = subprocess.Popen(['getfattr', '--only-
       values', '--absolute-names', '-n', 'user.size', surl],
    stderr=subprocess.PIPE, stdout=subprocess.PIPE).communicate
       ()
    #print 'output:[',output,']'
    #print 'error:[',error,']'

    if len(output) == 0:

        if verb:

            print "No size value found as extra attribute for
               file %s. Processing..." %surl.replace(localpath
               ,'')

        #Calc de timestamp value
        size = str(os.stat(surl).st_size).rstrip('\n')

        # Set the timstamp value for the file.
```

```
        setsize, error = subprocess.Popen(['setfattr', '-n', '
            user.size', '-v', size, surl],
        stderr=subprocess.PIPE, stdout=subprocess.PIPE).
            communicate()
        return size

    else:

        return output.rstrip('\n')

def print_storage_dump(lfn, size, ctime, cksum):
    '''
    Create a file with keys: lfn, size, timestamp, cksum
    '''

    f = open(outfile, 'a')

    f.write(lfn)
    f.write('|')
    f.write(size)
    f.write('|')
    f.write(ctime)
    f.write('|')
    f.write(cksum)
    f.write('\n')


def threats(surl):
    '''
    Print the correct format for txt file.
    for user's files the cksum value is omited.
    '''

    if hash == 'adler32' or hash == 'md5' or hash == 'sha256':
```

```
            print_storage_dump(surl.replace(localpath,''),
                get_local_size(surl), get_local_timestamp(surl),
                get_local_cksum(surl))
        else:
            print_storage_dump(surl.replace(localpath,''),
                get_local_size(surl), get_local_timestamp(surl), 'N/
                A')


def merge_files():

    file_list = glob.glob('/tmp/DumpFile*%s*.txt' %datetime.
        date.today())

    #print file_list
    with open('DumpFile.%s.txt' %datetime.date.today(), 'w')
        as file:
         input_lines = fileinput.input(file_list)
         file.writelines(input_lines)

    print "writing output file as DumpFile.%s.txt" %datetime.
        date.today()

def getopts():
    '''
    Get the command line arguments
    '''

    parser = argparse.ArgumentParser(description='Process
        inline variables')

    parser.add_argument('-a', '--all', action='store_false',
        dest='dumpall', help='Storage Dump all the FS under path
         (overwrite -d value)')
```

```
    parser.add_argument('-v', '--verb', action='store_true',
        dest='verb', help='Default silent mode. overwrite -v
        versose')
    parser.add_argument('-m', '--1month', action='store_true',
        dest='dumpolder1', help='Storage Dump files older them 1
         Month  under path (overwrite -d value)')
    parser.add_argument('-c', '--cksum', dest='cksum', help='
        Look chksum value if it doesn\'t exist then calc it (
        values adler32 (default), md5, sha256)')
    parser.add_argument('-d', '--days', dest='days', default
        ='7', help='Dump files at FS newer than \"days\" (
        default value 7)')
    parser.add_argument('-n', '--ncores', dest='ncores',
        default='4', help='Number of cores to be used (default
        4)')
    parser.add_argument('-p', '--paths', dest='paths', nargs
        ='+', required=True, help='Path to look for files,
        Mandatory')

    return parser.parse_args()

if __name__ == '__main__':


######################Globals#####################
##################################################

    try:
        myargs = getopts()
        all = myargs.dumpall
        month = myargs.dumpolder1
        hash = myargs.cksum
        days = myargs.days
        paths = myargs.paths
        ncores = myargs.ncores
```

```
        verb = myargs.verb
        print all, verb, month, hash, days, ncores, paths

    except KeyError:
        print "missing some mandatory parameters, please run <
            StorageDumps.py -h >"
        sys.exit()

    try:
        for path in paths:
            outfile = '/tmp/DumpFile.%s.%s.txt' %(filter(None,
                path.split('/'))[-1],datetime.date.today())
            localpath = path[:path.rfind("/store")]
            if not os.path.isfile(outfile):
                for tupla in os.walk(path, followlinks=True):
                    if tupla[2]:
                        po = Pool(processes=int(ncores))
                    #try:
                        if all:
                            t = datetime.date(2000,01,01)
                            if time.mktime(t.timetuple()) <= os
                                .path.getctime(tupla[0]):
                                #for file in tupla[2]:
                                lista = [ tupla[0]+'/'+f for f
                                    in tupla[2] ]
                                print lista
                                result = po.map_async(threats,
                                    lista)
                                result.wait()

                        elif month:
                            t = datetime.date.today() -
                                datetime.timedelta(days=30)
                            if time.mktime(t.timetuple()) >= os
                                .path.getctime(tupla[0]):
```

```
                                #for file in tupla[2]:
                                lista = [ tupla[0]+'/'+f for f
                                    in tupla[2] ]
                                result = po.map_async(threats,
                                    lista)
                                result.wait()

                        else:
                            t = datetime.date.today() -
                                datetime.timedelta(days=int(days
                                ))
                            #print tupla[0], os.path.getctime(
                                tupla[0]), time.mktime(t.
                                timetuple())
                            if time.mktime(t.timetuple()) <= os
                                .path.getctime(tupla[0]):
                            #for file in tupla[2]:
                                lista = [ tupla[0]+'/'+f for f
                                    in tupla[2] ]
                                result = po.map_async(threats,
                                    lista)
                                result.wait()
                    #except OSError:
                        #print "Path no accesible %s" %surl
                        #pass
                          po.close()
                          po.join()
        merge_files()

    except IndexError:
        print "Maybe a empty line at some file"
        pass
```

```
usage: PosixStorageDump.py [-h] [-a] [-v] [-m] [-c CKSUM] [-d
    DAYS]
                           [-n NCORES] -p PATHS [PATHS ...]

Process inline variables

optional arguments:
  -h, --help            show this help message and exit
  -a, --all             Storage Dump all the FS under path (
      overwrite -d
                        value)
  -v, --verb            Default silent mode. overwrite -v
      versose
  -m, --1month          Storage Dump files older them 1 Month
      under path
                        (overwrite -d value)
  -c CKSUM, --cksum CKSUM
                        Look chksum value if it doesn't exist
                           then calc it
                        (values adler32 (default), md5, sha256)
  -d DAYS, --days DAYS  Dump files at FS newer than "days" (
      default value 7)
  -n NCORES, --ncores NCORES
                        Number of cores to be used (default 4)
  -p PATHS [PATHS ...], --paths PATHS [PATHS ...]
                        Path to look for files, Mandatory
```

```
############################################################################
```

```
############ calc_adler32.py #############
#!/usr/bin/env python
BLOCKSIZE=32*1024*1024
import sys
from zlib import adler32
```

```
for f in sys.argv[1:]:
        val = 1
        if f=='-':
                fp=sys.stdin
        else:
                fp=open(f)
        while True:
                data = fp.read(BLOCKSIZE)
                if not data:
                        break
                val = adler32(data, val)
        if val < 0:
                val += 2**32
        print hex(val)[2:10].zfill(8).lower()
```

##

Apéndice 02

A continuación, se muestra el código que realiza las comprobaciones del estado de las colas Cloud. Dependiendo de la existencia de trabajos encolados y de la disponibilidad de cores libres en los nodos en uso HPC, mueve jobs de una cola a otra.

```
#!/usr/bin/python

'''
Rebalance Jobs in Slurm Queues
author: Iban Cabrillo
v 1.0.2
2019/11/07
'''
from __future__ import division
from datetime import datetime
import os
import sys
import datetime
import itertools
import subprocess

def get_partition_state():
    '''
    Return a list with the partitions that are not in dain or
       down
    '''
```

```
    #Look for the adler32 value
    cmd="sinfo| grep -v PARTITION |grep -v 'down'|grep -v drain
        | awk {'print $1'}|uniq"
    output, error = subprocess.Popen(cmd, stderr=subprocess.
        PIPE, stdout=subprocess.PIPE, shell=True).communicate()
    #print 'output:[',output,']'
    #print 'error:[',error,']

    return output.rstrip('\n')


def cal_queue_weight(queues):
    '''
    Calc de wight between queue_cores/total_cores
    '''
    queue_weights = {}
    g_cores_idle = get_total_cores(queues,'I')
    for k, v in g_cores_idle.items():
        g_cores_idle[k] = int(v)
    total_idle_cores = sum(g_cores_idle.values())

    g_cores_total = get_total_cores(queues,'T')
    for k, v in g_cores_total.items():
            g_cores_total[k] = int(v)
    total_enable_cores = sum(g_cores_total.values())
    for queue in queues:
        if int(g_cores_idle[queue]) > 0:
            weight= 10*(float(g_cores_idle[queue])/
                total_idle_cores)
            queue_weights[queue] = int(round(weight))
        elif int(g_cores_idle[queue]) == 0:
            weight=10*float(len(get_pd_jobs(queue)))/
                total_enable_cores
            queue_weights[queue] = int(round(weight))
```

```
    return queue_weights

def get_node_list(state,queue):
    '''
    return the node list for a partition and a state (both
       strings)
    '''

    #look for node list
    cmd="sinfo | grep -w %s| grep -w %s| awk {'print $6'}" %(
       state,queue)
    output, error = subprocess.Popen(cmd, stderr=subprocess.
       PIPE, stdout=subprocess.PIPE, shell=True).communicate()
    #print 'output:[',output,']'
    #print 'error:[',error,']'
    return output.rstrip('\n')

def get_queue_cores(node_list,queue,core_info):
    '''
    Return two dics, one for individual Node/request core
       estatus of a queue
    and other with the total cores of required status of a
       queue
    '''
    core_stats = {"A":0,"I":1,"O":2,"T":3 }
    pair_dict = {}
    total_dict ={}
    #print node_list.split('\n')
    if len(node_list) != 0:
        for i in range(len(node_list.split('\n'))):

            cmd="sinfo --nodes=%s -o %%C | tail -n1" %(
               node_list.split('\n')[i])
            output, error = subprocess.Popen(cmd, stderr=
               subprocess.PIPE, stdout=subprocess.PIPE, shell=
```

```
                True).communicate()
            pair_dict[node_list.split('\n')[i]] = output.split
                ('/')[core_stats[core_info]].rstrip()

        cmd="sinfo --nodes=%s -o %%C | tail -n1" % ','.join(
            node_list.split('\n'))
        output, error = subprocess.Popen(cmd, stderr=subprocess
            .PIPE, stdout=subprocess.PIPE, shell=True).
            communicate()
        total_dict[queue] = output.split('/')[core_stats[
            core_info]].rstrip()

    else:
        total_dict[queue] = '0'
        pair_dict["Nonodes"] = '0'
    return pair_dict,total_dict

def get_pd_jobs(queue):
    '''
    Return pending job list for a given queue order by priority
    '''
    cmd="squeue -p %s -o %%i\" \"%%A --start | awk {'print $2'}|
        tail -n+2" %queue
    output, error = subprocess.Popen(cmd, stderr=subprocess.
        PIPE, stdout=subprocess.PIPE, shell=True).communicate()
    return output.rstrip('\n')

def get_cloud_pd_jobs():
    '''
    Return qeued cloud jobids order bay priority
    '''
    list= []
    cmd="squeue -o %i\" \"%A\" \"%P --start|grep cloud| awk
        {'print $2'}|tail -n+1"
```

```
    output, error = subprocess.Popen(cmd, stderr=subprocess.
       PIPE, stdout=subprocess.PIPE, shell=True).communicate()
    if output == "":
        return list
    else:
        list = output.strip('\n')
        return list

def get_total_cores(queues,core_info):
    '''
    Return the total cores in a determinate state (A/I/O/T)
    for a given list of enable (no down or drain) queues
    '''
    queue_cores = {}
    available_cores=''
    for queue in queues:

        #Calc de individual anf total cores for each avaible
           queue for a given state
        available_nodes = get_node_list('alloc',queue) + "," +
           get_node_list('idle',queue) + "," + get_node_list('
           mix',queue)
        ind_cores,tot_cores = get_queue_cores(available_nodes.
           rstrip(',').lstrip(','),queue,core_info)
        queue_cores[queue] = tot_cores[queue]
    return queue_cores

def update_pd_job(jobid, partition):
    '''
    Change job from partition
    '''
    #cmd="scontrol update jobid=%s partition=%s MinMemoryNode
       =20000 MinProcs=8" %(jobid,partition)
    cmd="scontrol update jobid=%s partition=%s" %(jobid,
       partition)
```

```
    output, error = subprocess.Popen(cmd, stderr=subprocess.
        PIPE, stdout=subprocess.PIPE, shell=True).communicate()
    print "Requeued jobid=%s to partition=%s" %(jobid,
        partition)
    return output

if __name__ == '__main__':


################################Globals
    ####################################
################################################################################

    gqueues=['cloudcms', 'cloudcmshp', 'cloudcmsfj']
    hpc=['compute']
    g_dis_queues=[]
    weights={}
################################################################################


    #Return a list with active an not active queues and calc de
        wiegh for each queue.
    try:
        g_actv_queues = [value for value in gqueues if value in
            get_partition_state().split('\n')]
        g_dis_queues = [value for value in gqueues if not value
            in get_partition_state().split('\n')]
        weights = cal_queue_weight(g_actv_queues)
        queue_list=[]
        j = 0

        #create the active queues iterate list
        for k,v in weights.items():
            queue_list=list(itertools.repeat(k,int(v)))+
                queue_list
```

```
#Balance the jobs in down queues between active queues
if g_dis_queues != []:
    for g_dis_queue in g_dis_queues:
        for i in range(len(get_pd_jobs(g_dis_queue).
            split('\n'))):
            #look for pending jobs in down queues
            get_pd_jobs(g_dis_queue).split('\n')
            #Move jobs to enable queues
            if get_pd_jobs(g_dis_queue).split('\n')[i]
                != "":
                update_pd_job(get_pd_jobs(g_dis_queue).
                    split('\n')[i],queue_list[j])
            if j == len(queue_list)-1:
                j = 0
            else:
                j += 1
else :
    for g_actv_queue in g_actv_queues:
        get_pd_jobs(g_actv_queue).split('\n')
        for i in range(len(get_pd_jobs(g_actv_queue).
            split('\n'))):
            get_pd_jobs(g_actv_queue).split('\n')
            if get_pd_jobs(g_actv_queue).split('\n')[i]
                != "":
                update_pd_job(get_pd_jobs(g_actv_queue)
                    .split('\n')[i],queue_list[j])
            if j == len(queue_list)-1:
                j = 0
            else:
                j += 1

#If there is Cloud pending jobs and there are free
    cores in hpc, then send move jobs to this queue
#Get pending jobs in cloud queues
g_pd_list = get_cloud_pd_jobs()
```

```
#Look only for free cores on used hpc machines not in
    full empty ones.
ind_cores_hpc,free_cores_hpc = get_queue_cores(
    get_node_list('mix','compute'),'compute','I')
minor = min(free_cores_hpc,len(g_pd_list))

#Get if there are enought resources available (cores
    idle)
g_cores_Idle = get_total_cores(gqueues,'I')

#Get Total cores in use
g_cores_used = get_total_cores(gqueues,'A')

if g_cores_Idle == 0 and free_cores_hpc['compute'] > 0
    and len(g_pd_list) > 0:
     print "minor=%s" % minor
     for i in range(minor):
         #update_pd_job(g_pd_list[i],'compute')
          pass
else:
     print
        "#############################################################"
     print "##### Time : %s      ########" % datetime.
        datetime.now().strftime("%d/%m/%Y %H:%M:%S")
     print "No job movement between partitions"
     print "Slots available at gqueues:%s" %
        g_cores_Idle
     print "Slots used at gqueues:%s" % g_cores_used
     print "Slot available at hpc:%s" % free_cores_hpc['
        compute']
              else:
                  j += 1
```

```
#If there is Cloud pending jobs and there are free
    cores in hpc, then send move jobs to this queue
#Get pending jobs in cloud queues
g_pd_list = get_cloud_pd_jobs()

#Look only for free cores on used hpc machines not in
    full empty ones.
ind_cores_hpc,free_cores_hpc = get_queue_cores(
    get_node_list('mix','compute'),'compute','I')
minor = min(free_cores_hpc,len(g_pd_list))

#Get if there are enought resources available (cores
    idle)
g_cores_Idle = get_total_cores(gqueues,'I')

#Get Total cores in use
g_cores_used = get_total_cores(gqueues,'A')

if g_cores_Idle == 0 and free_cores_hpc['compute'] > 0
    and len(g_pd_list) > 0:
     print "minor=%s" % minor
     for i in range(minor):
         #update_pd_job(g_pd_list[i],'compute')
         pass
else:
     print
        "#####################################################"
     print "###### Time : %s      ########" % datetime.
        datetime.now().strftime("%d/%m/%Y %H:%M:%S")
     print "No job movement between partitions"
     print "Slots available at gqueues:%s" %
        g_cores_Idle
     print "Slots used at gqueues:%s" % g_cores_used
     print "Slot available at hpc:%s" % free_cores_hpc['
        compute']
```

```
            print "Cloud pending jobs:%s" %len(g_pd_list)
            print
               "#################################################################"

    except IndexError:
        print "No disable queues in cloud"
        pass
```

Referencias

[1] F.˜Aguilar˜Gómez, "Data Management in a Cloud Framework: application to the LifeWatch ESFRI". PhD thesis, 11, 2017.

[2] Researchgate, "Digital Curation Centre Lifecycle Model. @ONLINE". `http://www.dcc.ac.uk/resources/curation-lifecyclemodel`, June, 2019. [Online; accessed jun-2019].

[3] Cern, "LHC MACHINE OUTREACH @ONLINE". `https://lhc-machine-outreach.web.Cern.ch/lhc-machine-outreach/`, October, 2018. [Online; accessed Oct-2018].

[4] Cern, "HIGGS. @ONLINE". `https://home.cern/resources/faqs/cern-and-higgs-boson`, June, 2019. [Online; accessed 07-jun-2019].

[5] Fermilab, "FERMILAB: TEVATRON @ONLINE". `https://www.fnal.gov/pub/tevatron/tevatron-accelerator.html`, June, 2019. [Online; accessed Jun-2019].

[6] O. E. para la Investigación Astronómica en el Hemisferio Sur, "ESO. @ONLINE". `https://www.eso.org`, June, 2019. [Online; accessed 07-jun-2019].

[7] Esa, "ESA. @ONLINE". `https://www.esa.int`, June, 2019. [Online; accessed 07-jun-2019].

[8] ESGF, "Earth System Grid Federation @ONLINE". `https://esgf.llnl.gov`, October, 2018. [Online; accessed 26-Oct-2018].

[9] M. Unican, "Meteo Unican. @ONLINE". `https://meteo.unican.es/trac/wiki/ESGF`, September, 2019. [Online; accessed 07-sep-2019].

[10] Top500, "Top500. @ONLINE". https://www.top500.org, September, 2019. [Online; accessed 25-sep-2019].

[11] L. Collaboration, "The LifeWatch Erin. @ONLINE". https://www.lifewatch.eu, April, 2019. [Online; accessed 17-Apr-2019].

[12] Daria.eu, "Dariah Collaboration. @ONLINE". https://www.dariah.eu, September, 2019. [Online; accessed 05-Sep-2019].

[13] Invenio, "Invenio Software. @ONLINE". https://invenio-software.org, September, 2019. [Online; accessed 10-Sep-2019].

[14] arXiv, "arXiv. @ONLINE". https://arxiv.org, September, 2019. [Online; accessed 07-sep-2019].

[15] INSPIRE, "HEPNames Searchn. @ONLINE". http://inspirehep.net/collection/HepNames?ln=es, December, 2019. [Online; accessed 011-Dec-2019].

[16] lhc closer, "LHC layout. Figure @ONLINE". https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_layout, October, 2018. [Online; accessed 27-Oct-2018].

[17] L. Closer, "Imanes y Detectore. @ONLINE". https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.magnets___detectors_i/idioma/es_ES, June, 2019. [Online; accessed 07-Aug-2019].

[18] A. Collaboration, "The Atlas Experiment. @ONLINE". https://atlas.Cern.ch, October, 2018. [Online; accessed 27-Oct-2018].

[19] C. Collaboration, "The CMS Experiment. @ONLINE". https://cms.Cern.ch, April, 2019. [Online; accessed 15-Apr-2019].

[20] A. Collaboration, "The Alice Experiment. @ONLINE". https://aliceinfo.Cern.ch, April, 2019. [Online; accessed 15-Apr-2019].

[21] L. Collaboration, "The LHCb Experiment. @ONLINE". https://lhcb.web.Cern.ch, April, 2019. [Online; accessed 15-Apr-2019].

[22] C. Collaboration, "The Cern Experiment. @ONLINE". https://home.Cern.ch/, April, 2019. [Online; accessed 15-Apr-2019].

[23] Egi, "EGI. @ONLINE". https://wiki.egi.eu/wiki/PROC14_VO_Registration, June, 2019. [Online; accessed 07-jun-2019].

[24] C. Collaboration, "MONARC. @ONLINE". https://monarc.web.cern.ch/MONARC/, June, 2019. [Online; accessed 05-jun-2019].

[25] G. W. Group, "GLUE v. 2.0 – Reference Realisation to LDAP Schema. @ONLINE". https://redmine.ogf.org/attachments/161/ogf-glue-2-to-LDAP-v7.3-draft5.pdf, June, 2014. [Online; accessed 07-jun-2019].

[26] OASIS, "OASIS Standard. @ONLINE". http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html, January, 2013. [Online; accessed 4-Dec-2019].

[27] OSG, "HTCondor-CE Overview. @ONLINE". https://opensciencegrid.org/docs/compute-element/htcondor-ce-overview/, September, 2019. [Online; accessed 4-Sep-2019].

[28] dCache Org, "Introducion to dCache. @ONLINE". https://www.dcache.org/, June, 2019. [Online; accessed 01-jun-2019].

[29] I. io, "StoRM. @ONLINE". https://italiangrid.github.io/storm, June, 2019. [Online; accessed 01-jun-2019].

[30] C. Collaboration, "DPM. @ONLINE". https://lcgdm.cern.ch/dpm, June, 2019. [Online; accessed 01-jun-2019].

[31] EGI, "UMD products ID cards. @ONLINE". https://wiki.egi.eu/wiki/UMD_products_ID_cards, July, 2019. [Online; accessed 07-Jul-2019].

[32] EGEE, "YAIM. @ONLINE". https://twiki.cern.ch/twiki/bin/view/EGEE/YAIM, December, 2019. [Online; accessed 15-Dec-2019].

[33] Puppet, "Unparalleled infrastructure automation and delivery. @ONLINE". https://puppet.com, July, 2019. [Online; accessed 07-Jul-2019].

[34] RedHat, "RedHat Ansible. @ONLINE". https://www.ansible.com, July, 2019. [Online; accessed 07-Jul-2019].

[35] EGI, "EGI-QC. @ONLINE". https://github.com/egi-qc, July, 2019. [Online; accessed 07-Jul-2019].

[36] EGI.EU, "The EGI marketplace. @ONLINE". https://marketplace.egi.eu/, September, 2019. [Online; accessed 4-Sep-2019].

[37] Egi, "EGI. @ONLINE". https://www.egi.eu/about/projects/, June, 2019. [Online; accessed 03-jun-2019].

[38] ~López~García, "Scientific cloud computing : improved resource provisioning, interoperability and federation". PhD thesis, 2, 2016.

[39] Crossgrid, "CROSSGRID. @ONLINE". http://www.eu-crossgrid.org/, June, 2019. [Online; accessed 05-jun-2019].

[40] C. Collaboration, "DATAGRID. @ONLINE". https://eu-datagrid.web.cern.ch/, June, 2019. [Online; accessed 05-jun-2019].

[41] MICINN, "Plan Estatal de Investigación Científica y Técnica y de Innovación. @ONLINE". http://www.ciencia.gob.es/portal/site/MICINN/menuitem.7eeac5cd345b4f34f09dfd1001432ea0/?vgnextoid=83b192b9036c2210VgnVCM1000001d04140aRCRD, June, 2019. [Online; accessed 07-Aug-2019].

[42] E. collaboration, "EGI Accounting portal. @ONLINE". https://accounting.egi.eu, June, 2019. [Online; accessed 05-jun-2019].

[43] Nordugrid, "Nordugrid. @ONLINE". http://www.nordugrid.org, August, 2019. [Online; accessed 24-Aug-2019].

[44] I. Cabrillo et al., "Direct exploitation of a top 500 Supercomputer for Analysis of CMS Data", *Journal of Physics: Conference Series* **513** (jun, 2014) 032014, doi:10.1088/1742-6596/513/3/032014.

[45] C. Collaboration, "EGI Accounting portal. @ONLINE". https://Cernvm.Cern.ch/portal/filesystem, June, 2019. [Online; accessed 05-jun-2019].

[46] O. Collaboration, "OPENSTACK KVM. @ONLINE". https://docs.openstack.org/nova/latest/admin/configuration/hypervisor-kvm.html, June, 2019. [Online; accessed 05-jun-2019].

[47] D. Bonacorsiet˜al., "The CMS experiment workflows on StoRM based storage at Tier-1 and Tier-2 centers", Technical Report CMS-CR-2009-080, CERN, Geneva, (May, 2009).

[48] Ceph, "Ceph. @ONLINE". https://ceph.io, June, 2019. [Online; accessed 07-Aug-2019].

[49] O. Collaboration, "OPENSTACK DOCS. @ONLINE". https://docs.openstack.org/latest, June, 2019. [Online; accessed 05-jun-2019].

[50] P. O. B. J. P. B. F. D. M. L. T. P. D. Petravick, "The Storage Resource Manager Interface Specification Version 2.2. @ONLINE". https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.pdf, September, 2009. [Online; accessed 07-Aug-2019].

[51] T. Berners-Lee, F. R.T., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax",.

[52] W. Liu et al., "GridFTP GUI: An easy and efficient way to transfer data in grid", volume 25, pp. 57–66. 09, 2009. doi:10.1007/978-3-642-11733-6_7.

[53] C. s. CMS Collaboration:R. Carlin, "Status of CMS. @ONLINE". https://indico.cern.ch/event/843657/contributions/3542213/attachments/1927114/3205801/CERN-RRB-2019-101.pdf, October, 2019. [Online; accessed 17-Oct-2019].

[54] C. Collaboration, "CMS: The computing project. Technical design report",.

[55] C. Collaborattion, "GEANT4. @ONLINE". https://geant4.web.cern.ch/, August, 2019. [Online; accessed 24-Aug-2019].

[56] C. Collaboration, "Fts3 Docs at cern. @ONLINE". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormats, June, 2019. [Online; accessed 24-Aug-2019].

[57] red.es, "Migración de los 100 GB de la red GÉANT troncal y transatlántica. @ONLINE". http://www.rediris.es/difusion/publicaciones/e-boletin/1/n5.html, June, 2019. [Online; accessed 07-jun-2019].

[58] C. Grandi et al., "CMS computing model evolution", *Journal of Physics: Conference Series* **513** (jun, 2014) 032039, doi:10.1088/1742-6596/513/3/032039.

[59] N. M. et al., "Distributed data transfers in CMS", 2010.

[60] M. A. et al., "Experience building and operating the CMS Tier-1 computing centres", 2010.

[61] Hepix, "How to Run HEP-SPEC06 Benchmark. @ONLINE". https://w3.hepix.org/benchmarking/how_to_run_hs06.html, November, 2019. [Online; accessed 07-Nov-2019].

[62] Cern, "LHCOPN - IP parameters, addresses and routing. @ONLINE". https://twiki.cern.ch/twiki/bin/view/LHCOPN/LHCopnRoutingDoc, November, 2019. [Online; accessed 07-nov-2019].

[63] WLCG, "REBUS: VO Requirements @ONLINE". http://wlcg-rebus.cern.ch/apps/pledges/requirements/, June, 2019. [Online; accessed 12-Jun-2019].

[64] C. Collaboration, "Fts3 Docs at cern. @ONLINE". http://fts3-docs.web.cern.ch/fts3-docs/, June, 2019. [Online; accessed 07-jun-2019].

[65] A. Kiryanov et al., "FTS3 - A File Transfer Service for Grids, HPCs and Clouds", p. 028. 03, 2016. doi:10.22323/1.239.0028.

[66] C. Collaboration, "WebFTS Simplifying power. @ONLINE". https://webfts.cern.ch/submit.php, August, 2019. [Online; accessed 28-Aug-2019].

[67] R. J. B. T. B. D. H. J. S. I. T. L. and Wu, "PhEDEx high-throughput data transfer management system Computing in High Energy Physics", 2006.

[68] R.~Egeland, T.~Wildish, and C.-H. Huang, "PhEDEx Data Service", volume 219,6, p. 062010. 2010.

[69] C. Collaboration, "PhEDEx CMS Data Transfers. @ONLINE". `https://cmsweb.cern.ch/phedex/`, June, 2019. [Online; accessed 07-jun-2019].

[70] G. S. Chahal, "Data Monte Carlo preparation in CMS", 2018.

[71] J. M. D. Silva et al., "Efficient monitoring of CRAB jobs at CMS", *Journal of Physics: Conference Series* **898** (oct, 2017) 092036, `doi:10.1088/1742-6596/898/9/092036`.

[72] Mellanox, "In-Network Computing and Next Generation HDR 200G InfiniBand. @ONLINE". `https://www.mellanox.com/pdf/whitepapers/WP_In-Network_Computing_Next_Generation_HDR_200G_IB.pdf`, September, 2019. [Online; accessed 12-Sep-2019].

[73] Mellanox, "Mellanox Comunity. @ONLINE". `https://community.mellanox.com/s/article/what-is-rdma-x`, September, 2019. [Online; accessed 12-Sep-2019].

[74] Freeipa, "Freeipa. @ONLINE". `https://www.freeipa.org/page/Main_Page`, August, 2019. [Online; accessed 30-Aug-2019].

[75] Nordugrid, "JURA Accounting Technical Details. @ONLINE". `http://www.nordugrid.org/arc/arc6/tech/accounting/jura.html`, August, 2019. [Online; accessed 24-Aug-2019].

[76] DPHEP, "DPHEP Data Preservation in High Energy Physics. @ONLINE". http://hep-project-dphep-portal.web.cern.ch/, September, 2019. [Online; accessed 07-jun-2019].

[77] Y. Totsuka, "JADE Experiment", in *Proceedings: 3rd Tristan Workshop, KEK, Tsukuba, Feb 28-Mar 1, 1977*, pp. 189–196. 1977.

[78] D. Collaboration, "DESY. @ONLINE". `http://www.desy.de/forschung/anlagen__projekte/hera/index_ger.html`, June, 2019. [Online; accessed 07-jun-2019].

[79] J. Szuba, "HERA Data Preservation plans and activities", *Journal of Physics: Conference Series* **331** (12, 2011) 072032, doi:10.1088/1742-6596/331/7/072032.

[80] S. Amerio et al., "Data preservation at the Fermilab Tevatron", *Nucl. Instrum. Meth.* **A851** (2017) 1–4, doi:10.1016/j.nima.2017.01.043, arXiv:1701.07773.

[81] C. Collaboration, "Indico. @ONLINE". https://getindico.io/, August, 2019. [Online; accessed 30-Aug-2019].

[82] C. Collaboration, "CMS data preservation, re-use and open access policy. @ONLINE". https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=6032, May, 2012. [Online; accessed 8-jul-2019].

[83] Hepforge, "Hepforge. @ONLINE". https://rivet.hepforge.org/, August, 2019. [Online; accessed 2-Aug-2019].

[84] Hepdata, "HEPData. @ONLINE". https://www.hepdata.net/, November, 2019. [Online; accessed 10-Nov-2019].

[85] C. C. Org., "Creative Commons. @ONLINE". https://creativecommons.org/, December, 2019. [Online; accessed 18-Dec-2019].

[86] O. C. D. Finland, "Open CMS Data Finland. @ONLINE". https://twiki.cern.ch/twiki/bin/view/HIPCMSExperiment/CMSOpenDataProject, August, 2019. [Online; accessed 2-Aug-2019].

[87] HEPiX, "HEPiX. @ONLINE". https://www.hepix.org, August, 2019. [Online; accessed 2-Aug-2019].

[88] C. Colaborration, "Cern Analysis Preservation Portal. @ONLINE". https://analysispreservation.cern.ch/, September, 2019. [Online; accessed 18-Sep-2019].

[89] Reana, "REANA. @ONLINE". http://reana.io/, September, 2019. [Online; accessed 18-Sep-2019].

[90] X. D.-T. e. a. Chen, "Open is not enough", *Nature Physics* **15** (feb, 2019) doi:https://doi.org/10.1038/s41567-018-0342-2.

[91] M. D. e. a. Wilkinson, "he FAIR Guiding Principles for scientific data management and stewardship", *Nature Physics* (Mar, 2016) doi:http://dx.doi.org/10.1038/sdata.2016.18.

[92] C. Collaboration, "EOS Service. @ONLINE". http://information-technology.web.cern.ch/services/eos-service, September, 2019. [Online; accessed 11-sep-2019].

[93] K. Lassila-Perini et al., "Implementing the data preservation and open access policy in CMS", *Journal of Physics: Conference Series* **513** (jun, 2014) 042029, doi:10.1088/1742-6596/513/4/042029.

[94] IFCA, "Clases Magistrales de Física de Partículas. @ONLINE". https://ifca.unican.es/es-es/educacion-y-divulgacion/clases-magistrales-de-fisica-de-particulas, December, 2019. [Online; accessed 2-Dec-2019].

[95] Quarknet, "Ouarknet. @ONLINE". https://www.i2u2.org, September, 2019. [Online; accessed 10-Sep-2019].

[96] IPPOG, "IPPOG. @ONLINE". http://ippog.org/, August, 2019. [Online; accessed 30-Aug-2019].

[97] I2U2, "I2U2. @ONLINE". https://quarknet.i2u2.org, September, 2019. [Online; accessed 10-Sep-2019].

[98] i2u2, "iSpy Webgl. @ONLINE". https://www.i2u2.org/elab/cms/ispy-webgl/, August, 2019. [Online; accessed 2-Sep-2019].

[99] physicsmasterclasses, "physicsmasterclasses. @ONLINE". https://cms.physicsmasterclasses.org/cms.html, July, 2019. [Online; accessed 8-jul-2019].

[100] Cisco, "Networking Fundamentals. @ONLINE". https://www.cisco.com/c/dam/global/fi_fi/assets/docs/SMB_University_120307_Networking_Fundamentals.pdf, September, 2019. [Online; accessed 2-Sep-2019].

[101] NetApp, "NetApp. @ONLINE". https://www.netapp.com/us/info/what-is-dynamic-disk-pools-technology.aspx, September, 2019. [Online; accessed 12-Sep-2019].

[102] Bacula, "Bacula. @ONLINE". http://bacula.org, September, 2019. [Online; accessed 18-Sep-2019].

[103] A. Y. R.-M. et al., "Experience with preservation of a Global Analysis Framework based on Virtual Machines for the CMS experiment at LHC", 2015. [Internal Note].

[104] C.˜S. C. J. A. C. I. J. C. A. C. S. H. D. C. J.˜F. et˜al., "Measurement of the tt¯ production cross section in the dilepton channel in pp collisions at s = 8 TeV",.

[105] CMS, "CMS Collaboration. @ONLINE". http://cms.cern.ch/iCMS/jsp/iCMS.jsp?mode=single&part=publications, August, 2019. [Online; accessed 20-Aug-2019].

[106] Cern, "Cern Document Server. @ONLINE". http://cds.cern.ch, October, 2019. [Online; accessed 20-Oct-2019].

[107] CMS, "CMS Collaboration. @ONLINE". http://cms.cern.ch/iCMS/analysisadmin/analysismanagement, August, 2019. [Online; accessed 20-Aug-2019].

[108] Cern, "Cern. @ONLINE". https://cernvm-online.cern.ch/context/new, August, 2019. [Online; accessed 22-Aug-2019].

[109] JupyterHub, "jupyterHub. @ONLINE". https://jupyter.org, August, 2019. [Online; accessed 22-Aug-2019].

[110] Docker, "Docker. @ONLINE". https://www.docker.com/, September, 2019. [Online; accessed 18-Sep-2019].

[111] Cern, "Cern. @ONLINE". https://docs.openstack.org/manila/latest/, August, 2019. [Online; accessed 22-Aug-2019].

[112] systemdspawner, "Systemdspawner. @ONLINE". https://github.com/jupyterhub/systemdspawner, August, 2019. [Online; accessed 22-Aug-2019].

Acrónimos

ALICE

A Large Ion Collider Experiment. 7, 10, 33

AOD

Analysis Object Data. 76, 80, 82, 132

API

Interfaz de Programación de Aplicaciones. 54, 56

ATLAS

A Toroidal LHC ApparatuS. 7, 8, 10, 26, 63

BE

BackEnd. 66, 67, 157

BSC

Barcelona Supercomputer Center. 104

CAF

CERN Analysis Facility. 87

CAP

CERN Analysis Preservation. 138

CDS

CERN Document Server. 153

CE

Computing Element. 59, 65, 115

CEDA

Centre for Environmental Data Analysis. 15

CERN

Laboratorio Europeo de Física de Partículas. 2, 6–8, 10, 20, 31, 37, 40, 73–75, 79, 83, 87, 128, 131, 133, 140, 154, 157

CMS

Compact Muon Solenoid. 2–4, 8, 10, 26, 58–60, 63, 65, 73–76, 79–87, 89, 92, 93, 97, 101, 103–105, 108, 109, 112, 113, 116, 129–135, 137, 139, 141, 142, 148, 149, 152, 157, 165, 166, 177

CMSSW

CMS Software. 98, 101, 102

CPD

Centro de Procesado de Datos. 1, 3, 59, 105, 113, 143, 145, 146

CRAB

CMS Remote Analysis Builder. 85

CSIC

Consejo Superior de Investigaciones Científicas. 147, 148

CVMFS

CernVM File System. 128, 155

CWR

Collaboration Wide Review. 153

DAS

CMS Data Aggregation System. 92

DC

Data Center. 37

DESY

Deutsches Elektronen-Synchrotron. 20, 118, 122, 123

DKRZ

German Climate Computing Centre. 15

DoE

Department of Energy. 14

DPG

Detector Performance Groups. 99

DPHEP

Preservación de Datos para Física de Altas Energías. 4, 118–123, 134, 137, 142

DPM

Disk Pool Manager. 66

DQM

Data Quality Monitoring. 98–100, 102

EGI

Infraestructure Grid Europea. 4, 39, 40, 52, 54, 58, 142, 166

ELT

Extremely Large Telescope. 12

EMI

European Middleware Initiative. 44

EoL

End of Life. 45, 59, 60

EOSC

European Open Science Cloud. 53, 54

ESA

Agencia Espacial Europea. 12, 13

ESD

Event Summary Data. 79

ESGF

The Earth System Grid Federation. 14–16

ESO

Observatorio Europeo Austral. 10–12

FE

FrontEnd. 47

FNAL

Fermilab National Laboratory. 82

FTP

File Transfer Protocol. 148

GFAL

Grid File Access Library. 41

GPFS

General Parallel FileSystem. 66, 67, 103, 105, 106, 108

HEP

Física de Altas Energías. 6, 20, 30, 117–119, 125, 129, 136

HERA

Hadron Electron Ring Accelerator. 118, 122–124

HPC

computación de alto rendimiento. 62, 63, 103–105, 108, 112, 113, 115, 116, 143, 165

IAAS

Infraestructura como Servicio. 55, 65

ICFA

Intenational Comitee For Future Acelerator. 119

ICTS

Instalación Científico Técnica Singular. 3, 104

IFCA

Instituto de Física de Cantabria. 1, 3, 4, 58–68, 72, 78, 79, 85–87, 97, 103–105, 107, 112, 113, 142, 143, 145, 146, 149, 153, 157, 159

INFN

Instituto Nazionale di Fisica Nucleare. 59, 60

IPCC

Intergovernmental Panel on Climate Change. 15, 16

IPPOG

International Particle Physics Outreach Group. 141

IPSL

Institut Pierre-Simon Laplace. 15

JSON

JavaScript Object Notation. 141

KVM

Kernel Virtual Machine. 144, 146

LCG

LHC Computing Grid. 40, 41

LEP

Large Electron-Positron collider. 7, 8, 117, 118

LHC

Gran Colosionador de Hadrones. 2, 4, 6–8, 25, 26, 31, 33, 37, 40, 43, 64, 66, 81, 82, 88, 93, 117, 118, 125, 131, 133–136, 141

LHCb

Large Hadron Collider beauty. 7, 10

MC

Monte Carlo. 74, 83, 85, 87, 88, 97, 102, 123, 124, 128, 135

MDST

Mini Data Summary Tape. 123, 124

ML

Machine Learning. 67, 133

MoU

Memorandum of Understanding. 82

MPI-M

Max Planck Institute for Meteorology. 15

NASA

National Aeronautics and Space Administration. 15

NCI

Australian National Computational Infrastructure. 15

NOAA

National Oceanic and Atmospheric Administration. 15

NSD

Network Share Disk. 67

NSF

National Science Foundation. 15

NTT

New Technology Telescope. 11

PAG

Physics Analisys Group. 100

PAP

Policy Administration Point. 44

PB

Petabyte. 37

PCMDI

Coupled Model Intercomparison Project. 15, 16

PDP

Policy Decision Point. 44

PEP

Policy Enforcement Point Server. 44

PETRA

Positron-Elektron-Tandem-Ring-Anlage. 121

PhEDEx

Physics Experiment Data Export. 80, 92, 93, 97

PIC

Port dÍnformació Científica. 82

POG

Physics Object Group. 97, 99–101

RAL

Rutherford Appleton Laboratory. 82

RDMA

Remote Direct Memory Access. 103, 104, 108, 143

REANA

Reusable Análysis. 138

RECO

RECOnstructed Data. 82

RES

Red Española de Supercomputación. 3, 63, 66, 104, 144

RIVET

Robust Independent Validation of Experiment and Theory. 135

SAM

Acceso Secuencial a Metadatos. 129

SE

Storage Element. 47, 59, 66

SGE

Son of Grid Engine. 60, 62, 63

SLAC

SLAC National Accelerator Laboratory. 20

SSSD

System Security Services Daemon. 149

TDR

Technical Design Report. 40

TMDB

Transfer Management Data Base. 93, 95

TriDAS

CMS Online Data Acquisition and Trigger System. 75, 81, 98

UC

Universidad de Cantabria. 1, 103, 104, 144, 157

UMD

Unified Middleware Distribution. 52

URI

Uniform Resource Identifier. 69

VM

Virtual Machine. 123, 139, 145, 155–157

VO

Organización Virtual. 22, 23, 39, 40, 43, 63, 64, 73, 74, 81–86, 89

VOMS

Virtual Organization Membership Service. 40

VOs

Organizaciones Virtuales. 39, 42, 58, 60

VSFTP

Very Secure FTP. 148, 150

WLCG

Worldwide LHC Computing Grid. 4, 40, 41, 77, 82, 83, 88

WMS

Work Load Management System. 42, 43, 45

WORM

Write Ones Read Many. 149

WWW

World Wide Web. 5, 21