UAB
Universitat Autònoma de Barcelona

Content Delivery Network solutions for the CMS experiment: the evolution towards HL-LHC

Carlos Perez Dengra

UAB

Universitat Autònoma de Barcelona

Content Delivery Network solutions for the CMS experiment: the evolution towards HL-LHC

A dissertation submitted by **Carlos Perez Dengra** at Universitat Autònoma de Barcelona to fulfill the degree of **Doctor of Philosophy**

Bellaterra, December 1st, 2023

Escola d’Enginyeria
Departament d’Arquitectura de Computadors i Sistemes Operatius

Content Delivery Network solutions for the CMS experiment: the evolution towards HL-LHC

A dissertation submitted by **Carlos Perez Dengra** at Universitat Autònoma de Barcelona to fulfill the degree of **Doctor of Philosophy.** This doctoral Thesis belongs to the research line of High Performance Computing of the Computer Science PhD Programme in the Universitat Autònoma de Barcelona under the supervision of Dr. Josep Flix and Dr. Anna Sikora.

Author	Advisor	Advisor
Carlos Perez Dengra	**Josep Flix Molina**	**Anna Sikora**
	Centro de Investigaciones Medioambientales, Energéticas y Tecnológicas (CIEMAT)	**Universitat Autònoma de Barcelona**

This Thesis has been funded by MINECO thanks to the BES-2017-082665 PhD formation grant within the FPA2016-80994-C2-1-R project, led by the CIEMAT institution. This work has been also supported by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21

Bellaterra, December 1st, 2023

Abstract

The High-Luminosity Large Hadron Collider (HL-LHC) program at CERN presents a major challenge to the field of High Energy Physics. Scheduled to become operational by 2029, the HL-LHC is designed to have an increased integrated luminosity[1] by a factor of 10 beyond the LHC's design value, hence it will increase proton-proton collisions to an unprecedented scale. This will significantly impact the way experimental data is stored and analyzed in the underlying worldwide distributed computing system, known as the Worldwide LHC Computing Grid (WLCG). The anticipated surge in data production by LHC experiments in the HL-LHC period and the computation required to process this data will be a challenge for the expected budget that worldwide funding agencies allocate to LHC computing. This situation has led scientists to search for new mechanisms to alleviate the expected increase of computational resources. Facing these challenges, the LHC experiments have launched an extensive Research and Development (R&D) program to reduce the resources requirements, and hence reducing the overall cost of compute and storage resources, both in terms of hardware and operations. Currently, the LHC computational resources are distributed across the WLCG, consisting of 170 centers in 35 countries that provide CPU and storage resources for the LHC experiments. This extensive R&D program is developed in conjunction with other non-LHC data-intensive sciences that have (or will have) similar computational challenges, since most of them use the same compute clusters as WLCG. The R&D projects that have emerged comprise efforts to integrate opportunistic resources, such as High Performance Computing centers (HPC) or commercial Clouds, in addition to improve application efficiencies through vectorization and porting suitable pieces of the code to Graphic Processing Units (GPUs), or even FPGAs. These projects aim to reduce the amount of compute resources elsewhere, and improve the application's efficiencies. However, the storage service context is more complex, since it cannot rely on opportunistic resources not-owned by the LHC experiments. The storage service involves many R&D projects for the whole data management, access, and orchestration areas. The most plausible proposed scenario involves reducing the complexity of the storage systems within WLCG by having a small number of centers holding

[1] Integrated luminosity quantifies the overall exposure to particle collisions at an accelerator detector and it is used to determine the total number of events produced in these collisions. It is then related to the total amount of data collected from particle collisions over a given period of time, and it is often expressed in units of inverse area, typically femtobarns (1b = 10^{-28} m^2, 1fb = 10^{-43} m^2).

LHC experiments data and serving them to other centers using the underlying high capacity private networks, while including regional data caches to serve data to CPU-only compute centers, or centers with small storage capacity. This would involve defining regions and grouping sites connected at low latencies and introducing new ways to transfer data between them. Additionally, the community also aims to introduce concepts such as storage classes or Quality of Service (QoS) to optimize the cost and efficiency at the sites.

The work in this Thesis has focused on efficiently evolving the distributed storage of the CMS experiment towards the HL-LHC by implementing cache systems (based on the XCache technology) between geographically close CMS WLCG sites. The case study specifically concentrates on the Spanish region, particularly the WLCG Tier-1 site in Barcelona (PIC) and the WLCG Tier-2 site in Madrid (CIEMAT), which are reliable and reference sites for CMS. Testbeds have been deployed at these sites, in a non-disruptive way, to investigate the benefits of incorporating these new cache elements into the existing storage system architecture. Before deploying a data cache as a production service running at scale, a comprehensive evaluation has been conducted to analyze how the experiment utilizes the data storage service and how data access patterns occur at these sites. This analysis also helps identify experimental datasets that could potentially benefit from caching techniques. Additionally, network utilization studies have been conducted, incorporating cache simulations based on real data accesses, to determine the optimal characteristics of a cache system for the region. Finally, the testbeds for a physical cache have been complemented by investigations into the efficiency enhancements observed in CMS jobs when using the cache techniques, using an analysis benchmark execution task, along with evaluations of CPU walltime work savings achieved during execution, which achieved a 10% reduction during the execution of analysis tasks. Additionally, research provides insights into determining the optimal data cache size and network limits, which has been found to be around 200 TBs, a server that should be equipped with a 25 Gbps network interface to serve all of the Spanish CMS Tiers. The main objective of this Thesis is to reduce storage resource requirements and maintain storage deployment within limited budgets, with no significant expected budget growth towards the HL-LHC. Additionally, the aim is to improve application efficiency by bringing data close to compute nodes. Corresponding estimates have been made for both storage costs savings and application efficiency improvements. These studies have a significant influence on the broader CMS collaboration and are carried out in conjunction with international CMS colleagues who are investigating similar concepts in various regions.

Resumen

El programa del Gran Colisionador de Hadrones de Alta Luminosidad (HL-LHC) en el CERN presenta un gran desafío para el campo de la Física de Altas Energías. Programado para estar operativo en 2029, el HL-LHC está diseñado para contar con una luminosidad[2] integrada que será incrementada en un factor 10 respecto al valor de diseño del LHC, lo que aumentará las colisiones de protones-protones a una escala sin precedentes. Esto tendrá un impacto significativo en la forma en que los datos experimentales se almacenan y analizan en el sistema de computación distribuida subyacente, el Grid Mundial de Computación del LHC (WLCG). El aumento anticipado en la producción de datos por parte de los experimentos del LHC en el período HL-LHC y la computación necesaria para procesar estos datos serán un desafío para el presupuesto esperado que las agencias de financiación de los países miembros y colaboradores asignan a la computación del LHC. Esta situación ha llevado a los científicos a buscar nuevos mecanismos para aliviar el aumento esperado de los recursos computacionales. Para enfrentar estos desafíos, los experimentos del LHC se han embarcado en un extenso programa de Investigación y Desarrollo (I+D) para reducir los requisitos de recursos y, por lo tanto, reducir los costes generales de los recursos de computación y almacenamiento, tanto en términos de hardware como de operaciones. Actualmente, los recursos computacionales del LHC se distribuyen en la Grid de Computación del LHC Mundial (WLCG), que consta de 170 centros en 35 países que proporcionan recursos de CPU y almacenamiento para los experimentos del LHC. Este programa se desarrolla en colaboración con otras ciencias intensivas en datos no-LHC que tienen (o tendrán) desafíos similares, ya que la mayoría de ellas utilizan los mismos clústeres de computación disponibles en todo el mundo. Los proyectos de investigación que han surgido incluyen esfuerzos para integrar recursos oportunistas, como centros de Computación de Alto Rendimiento (HPC) o Clouds comerciales, además de mejorar la eficiencia de las aplicaciones a través de la vectorización y el reenvío de piezas adecuadas del código a unidades de procesamiento gráfico (GPU) o incluso FPGA. Estos proyectos tienen como objetivo reducir la cantidad de recursos de computación en otros lugares y mejorar la eficiencia de las aplicaciones. Sin embargo, el contexto del servicio de almacenamiento es más

[2] "La luminosidad integrada cuantifica la exposición global a colisiones de partículas en un detector de acelerador y se utiliza para determinar el número total de eventos producidos en estas colisiones. Ésta se relaciona con la cantidad total de datos recopilados a partir de colisiones de partículas durante un período de tiempo determinado y ,a menudo, se expresa en unidades de área inversa, generalmente en femtobarns (1b = 10^{-28} m^2, 1fb = 10^{-43} m^2)."

complejo, ya que no puede depender de recursos oportunistas no controlados por los experimentos del LHC. El servicio de almacenamiento implica muchos proyectos de I+D para todas las áreas de gestión de datos, acceso y orquestación. El escenario propuesto más plausible implica reducir la complejidad de los sistemas de almacenamiento dentro de la WLCG al tener un pequeño número de centros que almacenan los datos de los experimentos del LHC y los sirven a otros centros utilizando las redes privadas subyacentes e incluyendo cachés de datos regionales para servir datos a centros de computación de CPU solamente. Esto implicaría definir regiones y agrupar centros conectados a bajas latencias e introducir nuevas formas de transferir datos entre ellos. Además, la comunidad tiene intención de introducir conceptos tales como las clases de *storage* o el QoS (*Quality of Service* o Calidad de Servicio en castellano) con el objetivo de mejorar costes y eficiencia en los sites.

El trabajo de esta tesis se centra en la evolución eficiente del almacenamiento distribuido del experimento CMS hacia el HL-LHC mediante la implementación de sistemas de caché (basados en XCache) entre centros de la CMS WLCG geográficamente cercanos. El estudio de caso se centra específicamente en la región española, en particular en el centro Tier-1 de la WLCG en Barcelona (PIC) y el centro Tier-2 de la WLCG en Madrid (CIEMAT), que son centros fiables y de referencia para el CMS. Se han desplegado plataformas de prueba en estos centros, de forma no disruptiva, para investigar los beneficios de incorporar estos nuevos elementos de caché en la arquitectura del sistema de almacenamiento existente. Antes de desplegar la caché como un servicio de producción en funcionamiento a escala, se ha realizado una evaluación integral para analizar cómo el experimento utiliza el servicio de almacenamiento de datos y cómo se producen los patrones de acceso a los datos en estos centros. Este análisis también ayuda a identificar conjuntos de datos experimentales que podrían beneficiarse potencialmente de las técnicas de caché. Además, se han realizado estudios de utilización de la red, incorporando simulaciones de caché basadas en accesos a datos reales, para determinar las características óptimas de un sistema de caché para la región. Finalmente, la plataforma de prueba para una caché física se han complementado con investigaciones sobre las mejoras de eficiencia observadas en los trabajos de CMS cuando se utilizan las técnicas de caché, utilizando una tarea de ejecución de análisis de referencia, junto con evaluaciones de los ahorros de trabajo de CPU en tiempo de ejecución, que alcanzan una reducción del 10% en las tareas de análisis. Además, la investigación proporciona perspectiva para determinar el tamaño óptimo de la caché de datos y los límites de la red, que ha sido ajustada aproximadamente en unos 200 TB sin exceder los 25 Gbps para cubrir las necesidades de CMS en España. El objetivo principal de este trabajo es reducir los requisitos de recursos de almacenamiento y mantener el despliegue de almacenamiento dentro de presupuestos limitados, sin un crecimiento significativo esperado hacia el HL-LHC. El objetivo es mejorar la eficiencia de las aplicaciones acercando los datos a los nodos de cálculo. Las estimaciones muestran ahorros de costos y mejoras en la eficiencia. Estos estudios tienen un impacto considerable en toda la colaboración CMS y se están llevando a cabo en colaboración con otros compañeros de la comunidad explorando conceptos similares en otras regiones.

“Curiosity can bring guts out of hiding at times, maybe even get them going. But curiosity usually evaporates. Guts have to go for the long haul. Curiosity's like a fun friend you can't really trust. It turns you on and then it leaves you to make it on your own - with whatever guts you can muster.”

- May Kasahara in the *Wind-up Bird Chronicle*.

Acknowledgements

La paradoja de acabar la tesis es acabar escribiendo la primera página; supongo que la única en la que verdaderamente uno puede expresarse más allá de una contribución científica. Como me dijo una amiga, cada tesis es un mundo y no debes comparar tu camino con el del resto. Unos cuantos años de tesis dan para muchas situaciones (buenas y malas): una pandemia, la pérdida de la cercanía con el acto picaportes que tanto tiempo nos ahorra en ciencia al discutir, cambios radicales,... y un largo sinfín de circunstancias que podrían darme para una novela de Dostoievski. Pese a todos los momentos de flaqueza, uno siempre saca fuerzas de donde sea, porque quien haya pasado por esto sabe que admite poco respiros y pocas excusas. A veces se saca fuerza de dentro pero, sobretodo, se saca gracias al apoyo de las personas que tienes alrededor; la que soporta los monotemas, la irascibilidad, las dudas, los miedos y, sobretodo, los que comprenden el drama que supone una tesis. Y es por eso que, supongo, esta sección existe.

Primero de todo, gracias a mi director Josep ‘Pepe’ Flix por confiar en mí y aguantar durante todos estos años los *highs and downs* de un proceso que no ha sido fácil y, a riesgo de no caer en el síndrome del impostor, creo que he podido conseguirlo sobretodo a su inestimable ayuda y apoyo en los momentos más difíciles. De verdad, gracias por todo. También a mi directora Anna Sikora, por su tremenda paciencia y guía incondicional durante todo este período, a veces, con la dificultad que a veces supone comunicar y transmitir las ideas del computing de CMS a la comunidad no-LHC de la UAB. Espero que después de este tiempo, con la culminación del documento, por fin lo hayamos conseguido.

Después tenemos a la gente del PIC. Primero, a Antonio Pérez-Calero Yzquierdo, por su valioso *feedback* siempre disponible y por haberme ayudado incansablemente durante todo este proceso, incluyendo las tardes largas de jueves con reuniones hasta horas intempestivas en el despacho. También a Andreu, por compartir su larga experiencia en este “mundillo CERN” durante todos estos años.

Tampoco debo dejar de agradecer la ayuda y atención del resto de los PIC members por haberme brindado siempre su ayuda cuando la he necesitado durante estos últimos años: a Carles, los Jordis, Elena, Vanessa, Francesc, Pau, Jorge, Martin, Ricard, Cristian, Ester, Gonzalo y M. Carmen.

Aunque estemos siempre lejos y nos veamos poco, tengo también que agradecer la atención y valioso feedback de la gente desde CIEMAT que, con su ayuda, esta tesis ha podido avanzar durante estos años, a Chema, Antonio y Fco. Javier. Desde la distancia, también gracias por su apoyo.

Ahora, por fin, puedo abrir el espacio para acordarme de los más cercanos que me han aguantado durante todos estos años. Si me dejo a alguno, perdonadme: a Mireia, Adri (por compartir casi toda la tesis juntos y por tus ánimos), Vicky, Juanmi, Edu, Oriol, Eli, Enric, Gerard, Sergio, Laura, Juan, Jona, Marcos, Cristian (los dos), Isaac, Carles, Bárbara, Adrian, Emilio, Angi, Davirón, Manuel, Pedro y Marc.

También a Estefanía, por haber sido el apoyo indispensable que he necesitado durante esta durísima última etapa en la que siempre he tenido un hombro en el que apoyarme y sostener la parte de mi vida de la cual no era capaz de ocuparme mientras invertía tiempo en esto. Esta tesis también tiene una gran parte de contribución tuya.

Sobre todo, gracias a mi familia (Barcelona y Huéscar) y especialmente a mis padres, Carlos y Montserrat, que desde que nací siempre han apoyado mis decisiones, por muy poco ortodoxas y extrañas que sean. También a mi hermana Raquel y a Javi, por haberme siempre apoyado incondicionalmente en cualquier momento y circunstancia. Y qué decir de mis sobrinos Iria y Aran, que en muchos momentos me han sacado la sonrisa cuando la necesitaba.

Y finalmente, aunque hayan pasado muchos años y sería imposible que lo viera, se lo dedico a mi tío Antonio, que bien seguro se hubiera sentido orgulloso de este momento.

Como dice la canción de los Beatles: *All these places have their moments; with lovers and friends I still can recall. Some are dead and some are living... In my life I've loved them all.*

Academic contributions

The R&D work performed during my PhD Thesis has resulted in several contributions to conferences and publications:

Data access patterns analysis

- Carlos Pérez [presenter]
- Oral at pre-GDB XCache Workshop, 8th Jul. 2019, CERN (Switzerland)
- [Link to the presentation]

CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2

- Antonio Delgado, José Flix [presenter], José M. Hernández, Carlos Pérez, Antonio Pérez-Calero, et al.
- Oral at Computing in High Energy Physics 2019, Nov. 2019, Adelaide (Australia)
- [Link to the presentation]
- Publication: EPJ Web Conf. 245 (2020), 04028 - 10.1051/epjconf/202024504028

PIC: Storage studies for CMS

- Carlos Pérez [presenter]
- Oral at HSF WLCG Virtual Workshop, 20th Nov of 2020
- [Link to the presentation]

A Spanish data cache service for CMS experiment

- Carlos Acosta, Francisco J. Calonge, Antonio Delgado, José Flix [presenter], José M. Hernández, Carlos Pérez, Antonio Pérez-Calero, and Anna Sikora
- Oral at I Workshop de Computing y Software de la Red Española de LHC, Apr. 2021
- [Link to the presentation]

A content delivery network for CMS experiment in Spain

- Carlos Pérez [presenter]
- Oral at the 16th RES users conference, Sept. 2022, Cáceres (Spain)
- [Link to the presentation]

New storage and data access solution for CMS experiment in Spain towards HL-LHC era

- Carlos Pérez [presenter], José Flix, Anna Sikora, on behalf of the CMS Collaboration.
- Poster to 20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2021), Nov. 2021, Daejeon (South Korea)
- [Link to the presentation]
- Publication: J.Phys.Conf.Ser.2438 012053 (2023) - 10.1088/1742-6596/2438/1/012053

Simulating a content delivery network solution for the CMS experiment in the Spanish WLCG Tiers

- Carlos Pérez [presenter], José Flix, Anna Sikora, on behalf of the CMS Collaboration.
- Oral at the Virtual International Symposium on Grids & Clouds (ISGC 2022), Mar. 2022
- [Link to the presentation]

Deploying a cache content delivery network for the CMS experiment in Spain

- Carlos Pérez [presenter], José Flix, Anna Sikora, on behalf of the CMS Collaboration.
- Poster at the 21th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2022), 26th Oct. 2022, Bari (Italy)
- [Link to the presentation]
- Proceedings submitted and accepted in March 2023 [not yet published].

Experience deploying XCache for CMS in Spain

- Carlos Pérez [presenter], José Flix, Anna Sikora, on behalf of the CMS Collaboration.
- Oral at the XRootd+FTS Workshop, 30th March. 2023, Ljubljana (Slovenia)
- [Link to the presentation]

A case study of content delivery networks for the CMS experiment

- Carlos Pérez [presenter], José Flix, Anna Sikora, Jordi Casals, Carles Acosta, Cecilia Morcillo, Antonio Pérez-Calero, Antonio Delgado, José M. Hernández, Francisco J. Rodriguez for the CMS collaboration.
- Oral at 26th International Conference on Computing in High Energy and Nuclear Physics, 8th May of 2023, Norfolk (USA)
- [Link to the presentation]
- Proceedings submitted in September of 2023.

Simulating a XCache for CMS Crab jobs executed in Spain

- José Flix, Paula Serrano, Carlos Pérez, Anna Sikora, for the CMS collaboration.
- Abstract submitted in November 2023 to International Symposium on Grids & Clouds (ISGC) 2024, which will be held in Taipei (Taiwan) on 24-29 March 2024.
- [About the Conference]

Content Delivery Network solutions for the CMS experiment: the evolution towards HL-LHC

- Carlos Pérez, José Flix, Anna Sikora, for the CMS collaboration.
- Submitted to Journal of Parallel and Distributing Computing (JPDC) in November 2023
- [About the Journal]

Table of contents

Chapter 1

Introduction

The energy levels of the subatomic world are so high, and the particles are moving at such incredible speeds, that the ordinary concepts of space and time no longer apply. The study of these phenomena is one of the most fascinating and challenging areas of modern science, and CERN is at the forefront of this research.

- Paul Dirac, pioneering English theoretical physicist and Nobel Prize in Physics laureate in 1933

The Standard Model (SM) of Physics [1,2,3] is the quantum field theory that describes three of the four fundamental forces known in the universe through the interaction and properties of the elementary particles that compose it, namely the electromagnetic, weak and strong interactions, excluding gravity. Today, the SM is considered one of the most successful theory of Physics, with numerous verifications carried out by various High Energy Physics (HEP) experiments around the world, during decades.

The methodology of the SM is founded on the principle that the fundamental interactions of nature occur through the exchange of mediator particles known as bosons. These bosons are defined as force carriers that mediate the strong, weak, and electromagnetic fundamental interactions, so they facilitate interactions between elementary particles, namely quarks and leptons. Photons (massless bosons) mediate the electromagnetic force between electrically charged particles. The massive $W^{\pm}$ and Z bosons mediate the weak interactions between particles of different flavors (all quarks and leptons). These three bosons along with the photons are grouped together, as collectively mediating the electroweak interaction. Gluons are massless bosons that mediate the strong interactions between quarks. Quarks are never found in isolation, so they combine themselves to form hadrons, the most stable of which are protons and neutrons (both of them composed by combination of three up and down quarks), the components of atomic nuclei. On the other hand, we have charged leptons, such as electrons, muons and tauons, as well as their associated neutrinos, known as neutral leptons. All of those already known particles and their basic properties are shown in Figure 1.1. Mass is the amount of matter in a particle that affects gravity and inertia. Charge is the electric property that determines interactions with electromagnetic fields. Spin is the intrinsic angular momentum that influences particle behavior and quantum states. Neutrinos are subatomic particles that are known to have very tiny, but non-zero, masses. The existence of neutrino masses was confirmed through experiments, most notably oscillation experiments that showed that neutrinos can change between different "flavors" (electron, muon, and tau neutrinos) as they travel, which can only happen if they have mass. The Nobel Prize in Physics 2015 was awarded to Takaaki Kajita and Arthur B. McDonald for their significant contributions to the experimental discoveries of neutrino oscillations, which provided strong evidence for neutrino mass [4,5].

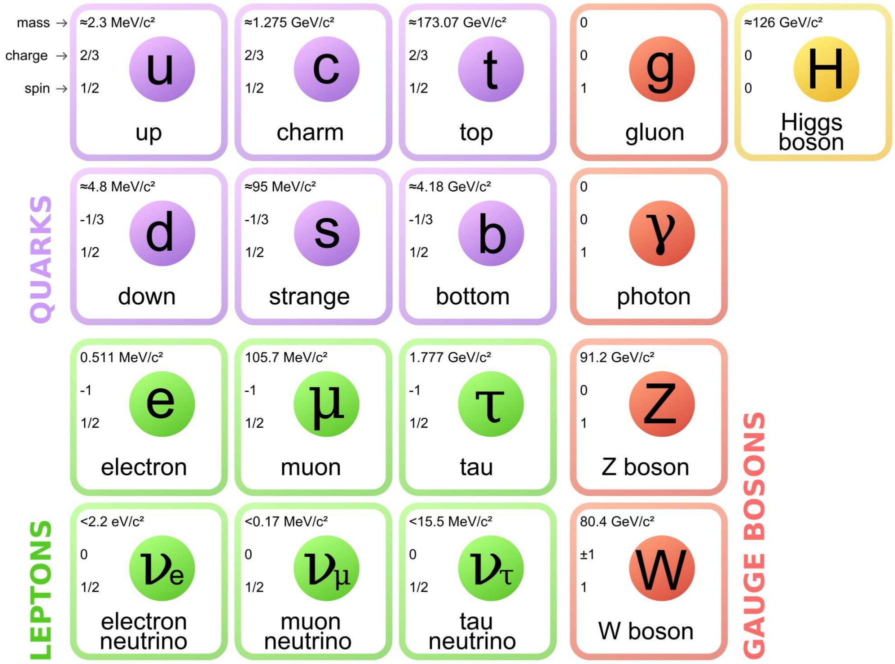


Figure 1.1: Fundamental particles predicted and experimentally verified within the SM.

As seen in Figure 1.1, there is an additional boson in the SM, the so-called Higgs boson, a massive boson which plays a unique role by explaining why the elementary particles acquire mass [6]. In quantum field theory, each quantum field has an associated boson. The Higgs field has always a non-zero value in space and remains constant, even when space is devoid of particles. As a result, all massive particles inevitably interact with this field, and through this mechanism, elementary particles acquire a mass proportional to the strength of their interaction. The interaction between all elementary particles and this field is mediated by the Higgs boson, which arises from the excitation of this field when it comes into contact with the particles.

The Higgs boson was expected to be very massive and very unstable, decaying almost immediately when created, so only a very high-energy particle accelerator can observe and record it. Its search was highly relevant during the last decades, and the LHC [7] at CERN (Switzerland, Geneva), together with its detectors, was instrumental for its discovery. Thus it was that in 2012, resulting from a collaboration involving thousands of scientists worldwide from the ATLAS and CMS experiments, the two multi-purpose LHC's detectors, the discovery of the Higgs boson was announced [8,9] the 4th of July of 2012. In Picture 1.1 the two Nobel laureates for its contribution to the prediction of the Higgs Mechanism, P. Higgs and F.Englert, can be seen at CERN's auditory receiving the official announcement for the Higgs boson discovery. The Higgs production is a very rare process that required billions of collisions to make the discovery to happen. The number of Higgs particles produced at the LHC is boosted when the accelerator collision energy and luminosity increases, and many detailed measurements of the Higgs particle have been made so far from the technical upgrades made at the LHC and its detectors. Its discovery finally confirmed the SM prediction and explained why particles acquire mass.

Picture 1.1: P. Higgs and F.Englert the day of the discovery announcement at CERN (4th July 2012). Credits: CERN.

Some of the open questions in Particle Physics today comprise, for example, the experimental evidence for the existence of supersymmetric (SUSY) partners for all of the known particles (massive partners that are theoretically viable), unveiling the nature and properties of dark matter (that dominates the gravitational effects in the universe, which existence is known but still undiscovered), the open question about matter and antimatter asymmetry in the universe (i.e. why the universe is only form of matter, if in the Big Bang the same amount of matter and antimatter was produced), and many more. Although the discovery of the Higgs boson is one of the most significant breakthroughs at the LHC to date, the LHC has continued to make new discoveries and observe rare processes, developing innovative ways to search for SUSY particles, setting or improving exclusion limits on the possible masses of these particles, and measure known particles' properties with high precision. The ongoing research is so rich that it is expected to produce many more discoveries in the future.

Throughout the entire process of generating collisions at the LHC detectors and producing the relevant Physics outcomes, it is necessary to count with a reliable computing infrastructure capable of efficiently managing the vast volumes of data generated at the LHC experiments. The deployment and operation of this computing infrastructure is crucial to ensure the success of experiments in order to advance in the research carried out at the LHC, as it was highlighted by CERN director Dr. Rolf Heuer on the announcement of the Higgs boson discovery.

1.1. The Large Hadron Collider (LHC)

The LHC was designed to occupy the space left by the Large Electron-Positron (LEP) collider [10], located at CERN near Geneva, Switzerland. With its 27-kilometer circumference, the LEP collider was the largest electron-positron accelerator ever built, with four detectors, ALEPH, DELPHI, L3 and OPAL [10] to observe the electron-positron collisions. LEP started operations in 1989, at an initial energy of 91 GeV, set to produce huge amounts of Z bosons. In the seven years that LEP operated at around 100 GeV it produced around 17 million Z particles. In 1995 LEP was upgraded for a second operation phase, and the collider's energy eventually topped 209 GeV in 2000 (close to the energetic regime of Higgs production). During 11 years of research, LEP's experiments provided a detailed study of the electroweak interaction. Measurements performed at LEP also proved that there are three – and only three – generations of particles of matter. LEP was closed down on 2 November 2000 to make way for the construction of the LHC in the same tunnel.

The LHC is the world's largest and highest-energy particle collider, with components built between 1998 and 2008 in a worldwide collaboration of over 10,000 scientists and hundreds of universities and laboratories, from more than 100 countries. The LHC dimensions can be observed in Figure 1.2 that displays the distribution of the LHC experiments throughout the accelerator spanning the Geneva city and part of the French territory. The accelerator primarily collides proton beams, but it can also accelerate beams of heavy ions: lead–lead collisions and proton–lead collisions are typically performed for one month a year. Four primary large-scale detectors, namely ALICE [11], ATLAS [12], CMS [13], and LHCb [14], are instrumental in observing the particle collisions generated at the LHC cutting-edge research facility. Among these, ATLAS and CMS serve as general-purpose experiments, while LHCb and ALICE specialize in investigating specific phenomena within particle Physics.

As versatile research instruments, ATLAS and CMS notably achieved a major milestone in 2012 by providing compelling evidence for the existence of the Higgs boson, thereby completing the experimental verification of the SM of Particle Physics. These detectors, designed with multifaceted capabilities, are now dedicated to ongoing data collection from proton and ion collisions, with the primary goal of uncovering Physics beyond the Standard Model (BSM). This pursuit includes endeavors to elucidate the enigmatic nature and presence of dark matter and to explore the conditions of the early universe during the immediate aftermath of the Big Bang. The LHCb experiment specializes in investigating the slight differences between matter and antimatter by studying a type of particle called the "beauty quark", or "b quark". In contrast, ALICE focuses on probing the properties of quark-gluon plasma, an exotic and dense state of matter that prevailed in the early universe. These distinct research objectives collectively advance our understanding of Fundamental Physics phenomena.

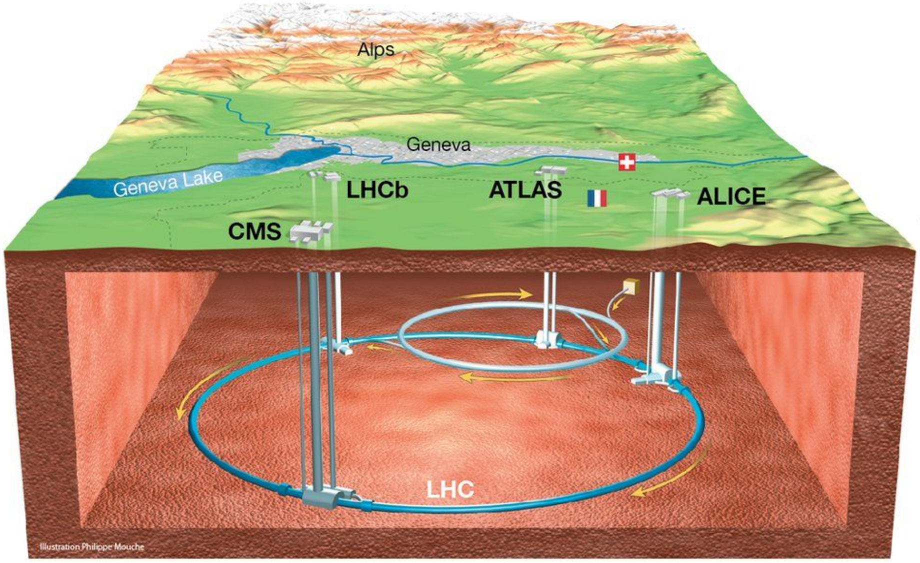


Figure 1.2: Schematic location of the LHC four experiments in relation to the Geneva city size and the border between Switzerland and France. Credits: CERN.

The first collisions at the LHC were achieved in 2010 at an energy of 3.5 teraelectronvolts (TeV) per beam, about four times the previous world record (Run1). The discovery of the Higgs boson at the LHC was announced in 2012. Between 2013 and 2015, the LHC was shut down and upgraded; after those upgrades it reached 6.5 TeV per beam (13 TeV total collision energy - Run2), reaching a peak luminosity which doubled the nominal design value. At the end of 2018, it was shut down for three years for further upgrades. In 2022 the accelerator has resume operations at a total collision energy of 13.6 TeV (Run3).

Luminosity is one of the most important aspects in the accelerator operational phase. The luminosity is the measure of the number of proton-proton collisions that occur per second in the LHC. It is calculated as the product of the beam intensity (number of protons per bunch) and the bunch crossing rate (number of bunches that collide each second). The LHC detectors are capable of recording around ∼95% of the delivered accelerator luminosity. As an example, Figure 1.3 (as taken from [15]) shows the total integrated luminosity at the CMS detector, for each year, since the LHC started operations. The largest recorded yearly luminosity translates to the largest amount of recorded collisions at the detector, and hence the largest amount of data to analyze. The total integrated luminosity has been steadily increasing since the LHC started operations, mainly due to the improvements to the accelerator's magnets and detectors, and the development of new techniques for colliding beams. These improvements have been deployed during the LHC technical shutdows years.

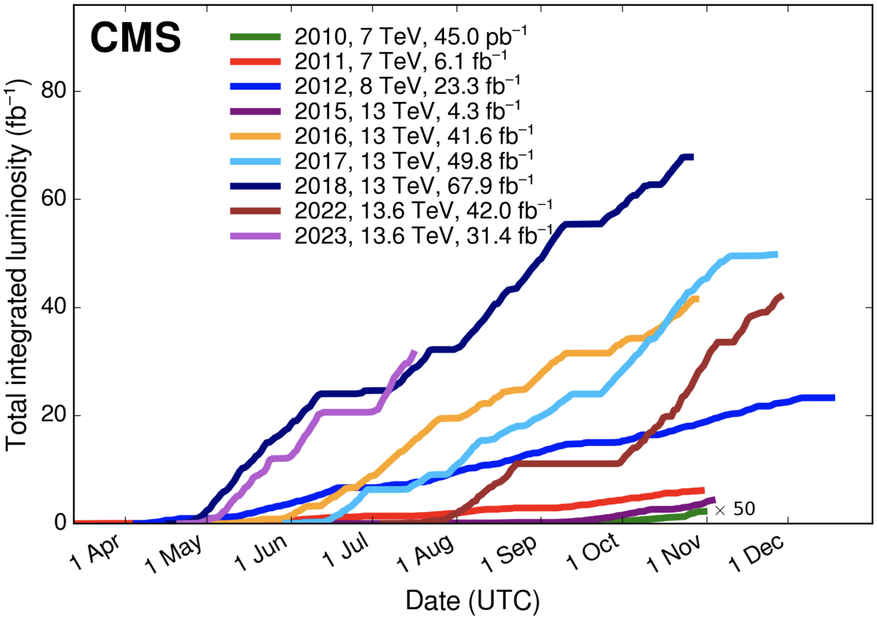


Figure 1.3: Total integrated luminosity achieved from 2010 to 2023 at the CMS experiment.

Moreover, in each of the proton-proton crossings at the detectors, many collisions occur, since the LHC accelerates trains of proton bunches, each bunch containing millions of protons. This maximizes the possibility of protons to collide at the center of the LHC detectors. The proton beams have been optimized at the collision points in order to maximize the number of collisions, hence, increasing the total integrated luminosity during these operation years. Figure 1.4 (also as taken from [15]) shows the distributions and averages for the number of interactions at each proton beam crossing (the so-called pile-up) at the CMS experiment, since 2011. As of today, the registered collisions at each beam crossing are a factor 5 higher as compared to initial LHC phase. It means that the recorded data is higher, and also that the collision images are more sophisticated to be analyzed, since each beam crossing contains many more collisions to analyze.

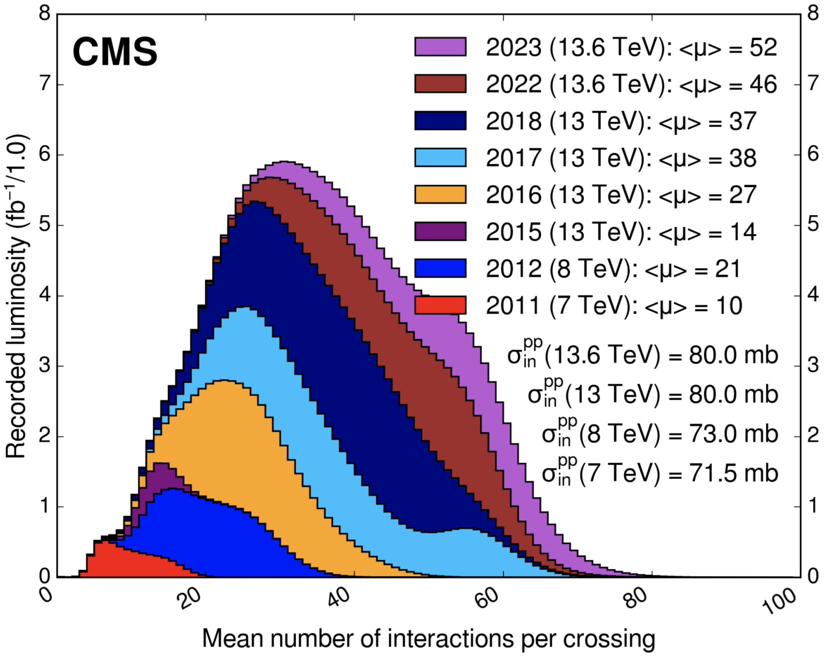


Figure 1.4: Recorded luminosity versus the mean number of interactions per crossing since 2011 until 2023 for the CMS experiment.

As a result of all of these improvements, the LHC experiments produce around 100 Petabytes of data per year, which requires a large computing infrastructure to handle this vast amount of acquired data. With more than 1 Exabyte of collision and simulated data samples produced so far, the data is stored on disk and magnetic tape and processed in a worldwide distributed computing infrastructure known as the WLCG [16].

1.2. The Worldwide LHC Computing Grid (WLCG)

Before building the LHC, its design parameters were known to produce more and more complex data than the previous experiences at CERN (in particular as compared to LEP experiments). The LHC experiments prepared estimates of computational and storage capacity requirements for the initial LHC phase and, for reasons of operability and budget, the computing of the LHC was proposed to be decentralized and distributed worldwide. In 1997 the MONARC [17] project emerged, as a solution to cover the large volumes of data produced by the LHC, with the need to distribute computational resources throughout the world and proposing a simple underlying architecture to manage data and processing activities. Parallel to this design, the concept of 'Grid', as conceived by Carl Kesselman and Ian Foster, was presented in 1999 [18]. This concept described the vision of a globally distributed infrastructure that could provide access to computing resources, applications and data on a large scale. In addition, it laid out the foundations for the introduction of concepts such as

Virtual Organizations (VOs), and the use of protocols for the massive transmission of data. The popularity of this novel concept and its coexistence with the MONARC project, with similar requirements and ideas, ended up causing a natural evolution of the LHC computational project towards the adoption of a Grid model. Finally, in the year 2000, a conference on HEP Computing was held in Padova [19], where the community agreed that the computing of the LHC would be regionally managed by several sites throughout the world, and a global community behind, under the umbrella of the WLCG, was born. The consolidation of the WLCG was carried out in a joint effort with funding from the European Union (EU [20]), the National Science Foundation (NSF [21]) from the US and some national Grid middleware projects, as well as the participation of computing sites. Finally, the infrastructure was developed in 2003 serving the LHC experiments, and commissioned and tested at scale prior to LHC starting producing collisions in 2010.

1.2.1. *Hierarchy of the WLCG infrastructure: the 'tiers'*

The WLCG is a distributed computational infrastructure organized in hierarchical levels (*Tiers*) in the data treatment. The sites are classified in Tiers according to their service level availability, processing, and storage capabilities. Three categories of Tiers are present: a Tier-0 at CERN, and 13 Tier-1 sites and 160 Tier-2s distributed world-wide. This hierarchy is represented in Figure 1.5. The Tier-0 site at CERN archives on robotic tape libraries the raw data from the LHC detectors, performs immediate reconstruction, and distributes a copy of the raw data and derived analysis data formats to the Tier-1 sites for custodial archival, via dedicated high-speed and private ~100 Gbps networks. On the other hand, the Tier-1 sites, around 13 big data centers worldwide, adhere to a 24x7 service level availability agreement, possess a robotic tape library for custodial storage of a proportional share of the data, and high throughput disk storage and compute processing systems for mass data reprocessing. Tier-1 sites run massive data processing campaigns to produce reduced analysis datasets for end-users, and massive simulation campaigns. The large storage systems deployed at Tier-1 sites ensures long-term data preservation. A shared fraction of compute resources at Tier-1s are as well used by end-users. In addition, the Tier-2 sites, around 160 sites worldwide spread in 42 countries, comply with a less demanding service availability (8x5) and are mainly dedicated to data analysis and production of simulated data.

The Tier-2 sites are often found in research institutions and universities. Several sites also offer dedicated resources for their local end-user communities, under the so-called Tier-3 category, composed of non-pledged resources, but opportunistically exploited. These Tier-3 resources are generally CPU only, because storage of WLCG has cataloged data which have to be managed by the experiments.

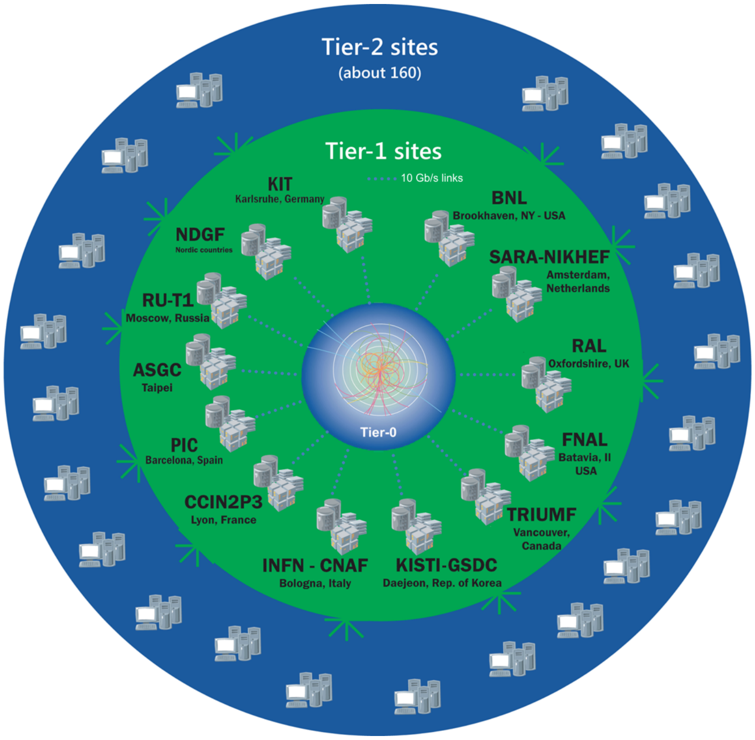


Figure 1.5: Structure and hierarchy of the WLCG Tiers. Source: WLCG.

Given the large data volumes transferred within WLCG sites, dedicated networking structures have been deployed, such as the LHC Optical Private Network (LHCOPN), and the LHC Open Network Exchange (LHCONE) network [22], which define private networks and provide dedicated bandwidth (10-100 Gbps range) to exchange data from the Tier-0 to the Tier-2 sites. The initially highly hierarchical organization of the data processing and analysis workflows, such as immediate reconstruction at the Tier-0, data reprocessing at Tier-1s, data analysis, and simulation production at Tier-2s, has blurred in the past years. Currently, many large and stable sites are capable of executing any workflow.

Spain has participated in the development, deployment, and operation of LHC computing since its inception. Spain contributes with a Tier-1 and three federated Tier-2 centers for the ATLAS, CMS, and LHCb experiments, offering about 4% of the total deployed capacity in WLCG. The Tier-1 center is located in Barcelona: the Port d'Informació Científica (PIC [23]), and supports the ATLAS, CMS and LHCb experiments, alongside various Astrophysics, Cosmology and Artificial Intelligence projects. ATLAS has Tier-2 sites at the Institute of Corpuscular Physics (IFIC [24]) of Valencia, the Autonomous University of Madrid (UAM [25]) and Institute for High Energy Physics (IFAE [26], co-located at PIC) in Barcelona. On the other hand, the CMS experiment has Tier-2 sites are both the Center for Environmental and Technological Research (CIEMAT [27]) in Madrid and the Physics Institute of Cantabria (IFCA [28]) in Santander. Finally, LHCb has Tier-2 sites at the University of Santiago de Compostela (USC [29]) and the University of Barcelona (UB [30]).

1.2.2. *Middleware: inter-connecting the sites*

The WLCG is a vast network of globally distributed computing centers managed by heterogeneous local services to access compute and storage resources. To maximize the transparent and efficient utilization of these resources, the system relies on middleware, which is a crucial abstraction layer within the architecture. These middleware services facilitate communication, coordination, and resource management across various components at the Grid sites. In both WLCG and the broader Grid computing environment, the middleware assumes a pivotal role, enabling the efficient and secure harnessing of distributed computing resources. Although there may be variations in the specific middleware components used by different LHC experiments, their collective goal is to establish a robust framework for job and data management, resource discovery, and security within the distributed computing landscape.

There are some common elements used by the LHC experiments, such as CERNVM-FS (CVMFS) [31], which is used for distributing software in a consistent manner across WLCG sites. The Globus Toolkit [32] is a set of open-source tools and services that provide a fundamental building block for security, data transfer, and execution management components. The File Transfer Service (FTS) [33] uses the Globus Toolkit and is extensively used in WLCG. Developed by CERN, the FTS is an Open Source software providing easy user interfaces for submitting secured transfers. With respect to job submissions, each of the WLCG sites provide Compute-Elements (CEs) in front of their compute batch systems. These CEs are used transparently by the experiments, regardless of the local batch scheduler that is used at the computing facilities. ARC-CEs [34] and HTCondor-CEs [35] are typically handling all of the experiment job submissions. These components, and many others, are typically integrated into the experiments' Workload Management Systems (WMS), which are in charge of orchestrating both data movements and job submissions. The WMS ensures that processing tasks are effectively and efficiently distributed across all hardware resources among all sites, and that the data is correctly placed across the infrastructure, on both disk and tape storages. The WMS typically handles central and end-user requests, and it plays a central role when accessing and managing the available resources. Some experiments have all of these elements integrated into a single service, such as DIRAC [36], which offers a complete Grid solution for communities that need to exploit distributed heterogeneous resources. DIRAC forms a layer between a community and various compute resources to allow optimized, transparent and reliable usage. The types of resources that DIRAC can handle include: Computing Resources (including Grids, Clouds, HPCs and Batch systems), Storage Resources (disk and tape) and Catalog Resources. Many communities use DIRAC, the oldest and most experienced being the LHCb collaboration. Other communities using DIRAC include, but are not limited to, Belle2 [37], ILC [38] and Cherenkov Telescope Array (CTA [39]). ATLAS and CMS have their custom-made WMS. Both of the experiments use Rucio [40] as a tool to manage the huge amount of data they generate across their heterogeneous and globally distributed storage systems. These services are open source and they are typically developed by the WLCG community, and have become popular to other HEP experiments, or even to other science domains.

Security and authentication in the Grid is also a critical factor. WLCG has relied on x509 certificates [41] since its inception, and it is currently transitioning to use token-based approaches [42] to secure access to experiment data and resources. The experiments make extensive use of GridFTP [43] (secured FTP), HTTPs [44], and XRootD [45] protocols for massively parallel data transfers across all of the Grid sites that are interconnected by multi-Gbps networks. To enhance security, WLCG has deployed two private networks: the LHCOPN and the LHCONE. In addition, monitoring also plays an important aspect. The systems are monitored in real-time, and alerts are delivered when a service is malfunctioning or when the resources are not used in the most optimal manner. The users can see the progress of submitted jobs, manage errors, and monitor data movement in real-time. Several monitoring systems are custom-made by the experiments, while many common services are monitored through the CERN MONIT [46] infrastructure.

By means of all of these elements, WLCG is able to manage around 1M of CPU cores and 1.5 ExaBytes of data distributed in around 170 worldwide distributed computing centers. WLCG is a prime example of how Big Data challenges are addressed in scientific research.

1.3. Computing challenges for the High-Luminosity LHC phase

The LHC operates in yearly periods characterized by incremental steps in the number of particles that collide in the detectors (i.e. luminosity) and total collision energy, known as Runs. Within these periodical phases of operations, the amount of experimental data to be stored, processed and analyzed gradually increases as a consequence of the higher number of collisions. Although resources are expected to increase for the current LHC period, the compute models of the experiments are not anticipated to undergo significant changes. However, there is an extensive research and development program aimed at increasing the number of collisions at the LHC by a factor of 10 by 2029, when the HL-LHC [47] is expected to start operating. Figure 1.6 (as taken from [48]) shows the LHC baseline plan for the next decade and beyond (the LHC lifetime is set up to the ~2040s). The upper line represents the energy of the collisions, while the lower lines represent the luminosity. The first long shutdown (LS1) in 2013-14 was necessary to allow for the optimization of beam energy and luminosity. The second long shutdown (LS2) in 2018-19 was focused on securing the luminosity and reliability of the LHC, as well as upgrading the LHC injectors. After the third long shutdown (LS3) that starts in 2025, the machine should finally be configured for the HL-LHC, reaching center-of-mass energies up to 14 TeV. In this ambitious upgrade, the LHC is set to undergo substantial changes, notably in luminosity. Two scenarios are considered, with collisions at 25 ns or (starting with) 50 ns, the latter implying a reduction in the number of proton bunches in the accelerator from $\sim$2,800 to $\sim$1,400 for the initial period. However, the noteworthy modification is anticipated in luminosity, projected to reach an impressive value of 7x10^{34} cm^{-2}s^{-1}, which is about five times the LHC's initial designed luminosity. This increase in luminosity and center of mass energy will enable researchers to observe rare processes and particles, as well as increase the chances of discovering new Physics beyond the SM. The increase in luminosity will lead to more pile-up at each proton beam crossing, transitioning from 60 pile-up events in Run-3 to 200 in the HL-LHC era, complicating the data reconstruction for all the experiments. To achieve the HL-LHC configuration, several upgrades are required, such as replacing accelerator magnets with higher magnetic field ones, installing new collimators, and upgrading the detectors, processes which will be done during the LS3 period.

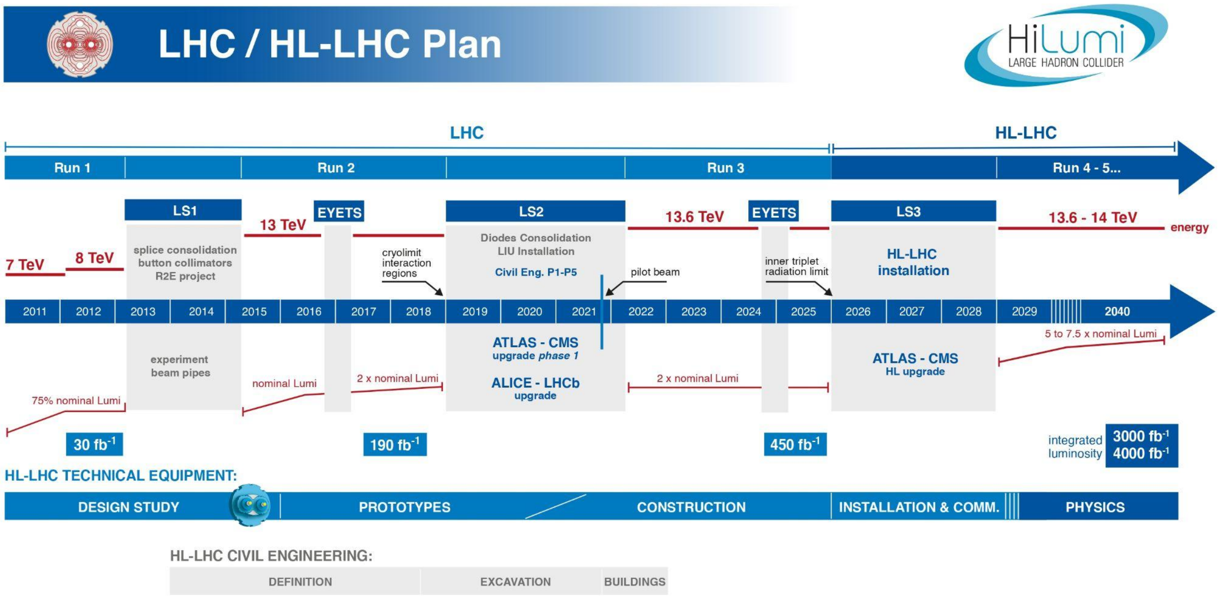


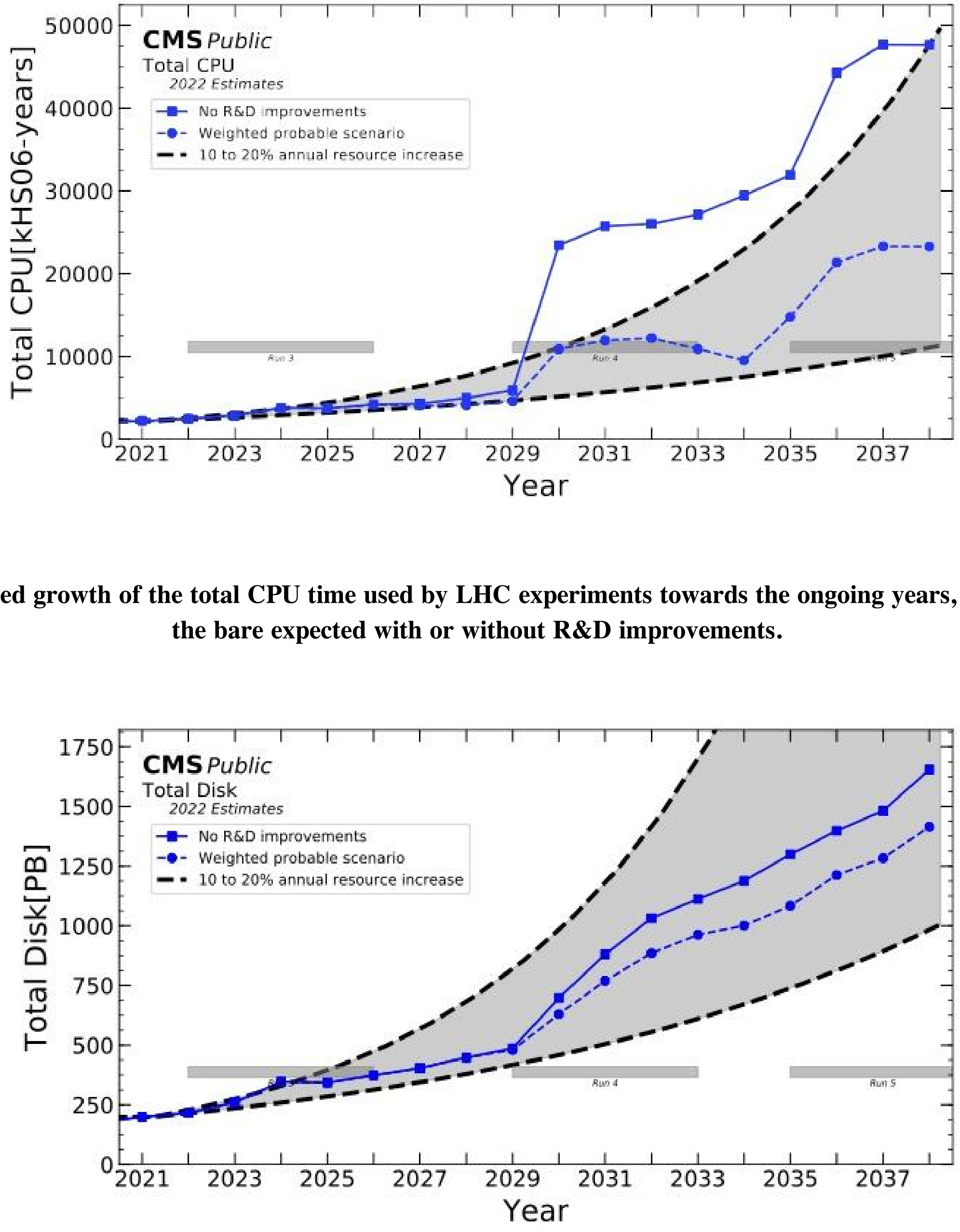
Figure 1.6: Projected HL-LHC plan (updated in December of 2022).

As a consequence of this update, which successfully will increase the number of collisions and, consequently, the volume of data to process, store, and analyze, the WLCG computational infrastructure needs significant advancements to align with the demands of the HL-LHC.

1.3.1. *The computing challenges of the HL-LHC*

The HL-LHC presents numerous challenges in terms of resource procurement and management, particularly given by the computing power and storage capacity to process and store the data that might go beyond the available budget that funding agencies assign to LHC computing worldwide. Additionally, the sheer volume of data generated by the LHC requires careful planning and optimization of resources to ensure that the scientific goals of the experiments can be achieved. In response to these challenges, LHC experiments have launched an extensive R&D program to develop novel ideas and techniques to evolve their computing models and services accordingly for the HL-LHC era [49]. These efforts involve not only the optimization of existing tools and technologies, but also the development of innovative solutions for data management and analysis. The ultimate goal is to ensure that the experiments are able to extract the maximum scientific output from the vast amount of data that will be generated during the HL-LHC phase. By taking proactive steps now to address these challenges, the LHC experiments are positioning themselves to make potential groundbreaking discoveries in the coming years.

The depicted plots in Figure 1.7 (as taken from [50]) showcase the future projection of the necessary computing resources for the HL-LHC compared with the estimated outcome within a flat-budget model, where no additional funding is expected. Two scenarios have been considered for resource needs: the baseline scenario, which assumes no improvement from ongoing R&D activities, and the second scenario, which incorporates the most likely outcome of these efforts. The blue curves and points represent the annual projected needs, calculated by summing the resource requirements of all WLCG sites for each scenario. To further illustrate projected resource availability, the gray band in the plots shows an example scenario that extrapolates the 2021 CMS pledged resources using an annual increase in available resources ranging from 10% to 20%, reflecting a flat budget scenario. Given technological evolution, with a flat financing profile, an increase in power or resources in those ranges can be obtained year after year. These estimates are based on a rough breakdown of CPU time, disk, and tape requirements for primary processing and analysis activities during a typical HL-LHC year. In relation to the projected capacity growth in a flat budget model for CMS (and also the rest of experiments), the increase in resources observed so far has been 5% instead of the 10% initially expected during the years previous to 2023. This discrepancy highlights the need to consider several external factors that influence the extrapolation of resources for scientific projects of this magnitude. Such factors include changing market dynamics and the constant evolution of technology, which can significantly impact the availability and allocation of financial resources. Therefore, it is essential to take these variables into account in future budget projections and resource management strategies to ensure the continued success of the LHC experiments computing infrastructure.



(a) Expected growth of the total CPU time used by LHC experiments towards the ongoing years, including the bare expected with or without R&D improvements.

(b) Expected disk drive storage expected growth towards the ongoing years, including the bare expected with or without R&D improvements.

Figure 1.7: CPU time and disk storage estimated annual requirements for CMS processing and analysis needs extrapolated from 2021 (updated estimates in December of 2022).

The proposed improvements to the CPU infrastructure include the use of GPUs and partial vectorization of applications, the integration of opportunistic resources (Clouds, voluntary computing), and the integration of HPC centers. However, the storage service poses a substantially more complex problem, since the sensitivity of the experiment's data can not be delegated to external service providers, and it creates a more complex scenario to be addressed in terms of data access and management at this expected magnitude. As shown in the Figure 1.7.b, the current model is still far from accommodating the increased data requirements, but experiments are underway to curb resource growth by introducing new techniques that rely on a more cost-effective model. Current research is focused on developing a more efficient and simplified distributed storage system that can scale to meet the demands of future experiments based on a Data-Lake model[3]. This model proposes a few centers to manage most of the LHC data, serving the demand of smaller centers with the use of simplified data caches, in which all compute resources will be embedded and/or plugged into.

The WLCG Data-Lake model [51] is conceived as an architecture designed to efficiently deliver and cache data across the WLCG sites, using a distributed storage infrastructure with a central namespace and geographically dispersed data nodes. This system mitigates the impact of latency, which arises from the division of data and computational resources, by taking advantage of its distributed architecture, cost-effective scaling, and diverse storage classes. Additionally, it optimizes hardware and operational costs through the strategic placement of caches close to where the data is required. In addition, it benefits from optimized data access and management across multiple sources, since it is able to get data closer to compute nodes on-demand when data-intensive applications are executed. In this scenario, the proposed data management approach outlined in the WLCG Data-Lake model aligns seamlessly with the concept of Content Delivery Networks (CDNs). A CDN is understood as an abstraction layer on top of the distributed storage architecture encompassing the necessary tools and services to equip the Data-Lake. In the industry, a geographically distributed CDN is often used to deliver content such as web pages, videos, or other digital plugins to end-users of any service. CDNs also offer a structured framework of centralized data centers, streamlined caching mechanisms and a unified integration of computing resources to effectively cover the needs of smaller centers, even centers with no storage element (SE) at all offering CPU-only resources to WLCG. The CDN’s network consists of a series of proxy servers and their data sites in a geographically distributed manner.

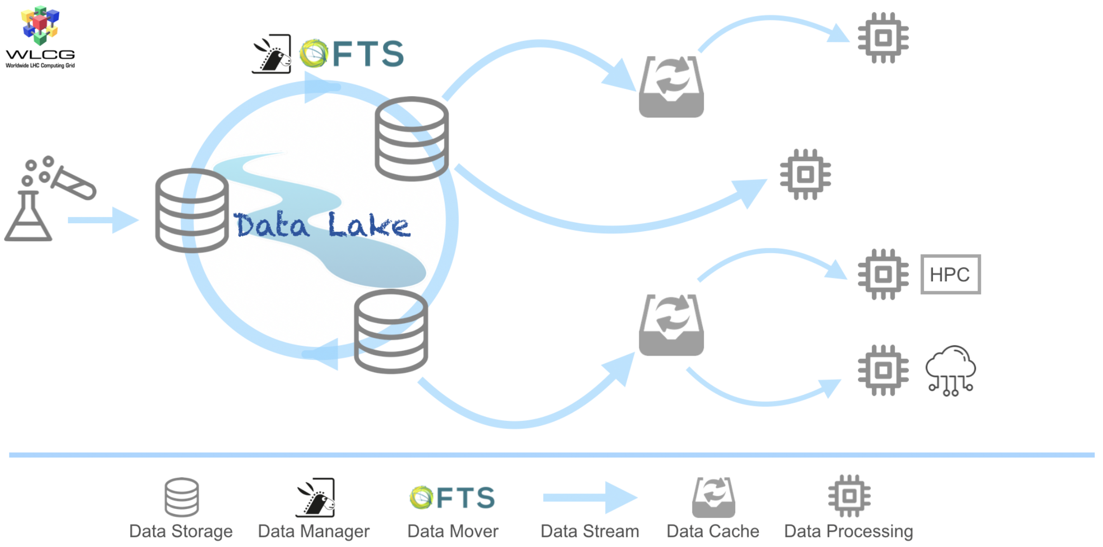
By deploying a CDN on top of the backbone of WLCG services through the deployment of data caches, data distribution could be optimized, improving the user experience and reducing

[3] The introduction of the 'Data-Lake' concept is attributed to James Dixon, making its first appearance in a blog post in 2010 [52]. In its initial definition, the concept referred to consolidated and centralized storage systems for raw or diversely structured massive data. This data could be taken from different sources without the requirement of predefined schemas or data transformations. However, the high flexibility of this concept causes ambiguity in its definition, and many of the architectures with similar aspects in their deployment can be interpreted as 'Data-Lake' models.

storage costs. Stateless caches are aimed to keep the data close to computing facilities, buffered from the main sites for their reutilization without operating any persistent storage (or a small amount) at the site. This is the point where the importance of caches lies within this design. Among the main benefits of deploying caching systems in the Data-Lake regions stand out improvement of CPU efficiency, storage savings, latency hiding, and even reduction of network usage. Indeed, an efficient low-latency CDN is crucial for data delivery to Data-Lake associated sites, optimizing CPU resource utilization, reducing job and queue times, and minimizing power consumption costs.

Despite the high volumes of data daily accessed by the sites, the network infrastructure that interconnects most of the sites in WLCG is high enough to fetch data and populate caches, hiding the latency for data reads, as if it were a local file. To avoid network congestion, cache sizes must be carefully planned. A structured framework of centralized data centers, optimized caching mechanisms, and unified resource integration can effectively address the needs of smaller centers. This new technology would also make it easier to integrate opportunistic resources, such as commercial Clouds or HPC centers, which could hold data caches or read files from the nearest or regional data cache. These heterogeneous physical facilities, regardless of their capacities, act together as a regional Data-Lake (spanning in the definition regions, countries and, even, continents) integrated to the Data-Lake global backbone, as shown in Figure 1.8. On top of that, one of the most interesting considerations of the model in terms of cost-effectiveness is that the natural evolution is for stable, large Tier-1 and Tier-2 centers to deploy the majority of storage in the Data Lake. This novel paradigm opens the door to run smaller sites without persistent storage systems deployed, which has been proved to perform successfully in WLCG sites with caching systems [53].

On the other hand, one of the benefits of this model is that storage can be deployed as a distributed service having access to multiple physical facilities but offering the users a single entry point. By concentrating their investments on running larger computing farms, larger storage systems, or both, sites within a WLCG Data-Lake can offer distributed storage as a service with transparent access to multiple physical facilities. To ensure the success of this model, a CDN that minimizes latency is needed, which will bring data to those sites attached to the Data-Lake. Lower latency ensures that CPU resources are utilized more effectively by minimizing idle waiting times and optimizing resource allocation. This leads to shorter job execution times, reduced queue times, and higher overall throughput in a distributed computing environment, while reducing the related power consumption costs. In order to evaluate and assess the feasibility of the proposals to evolve the storage and data management infrastructure, the DOMA (Data Organization, Management, and Access) task force [54] has been established as a collaboration between the LHC experiments and WLCG sites.



Datalakes, latency hiding and caching - Xavier Espinal (CERN)

Figure 1.8: Sketch of caching models leveraging data access from a consolidated storage infrastructure labeled as WLCG Data-Lake as conceived in [55].

This Thesis addresses the effects of the inclusion of CDNs in Spain and the resulting outcomes and tasks have been presented and discussed within CMS and in the DOMA task force, as well as presented in many main HEP computing conferences.

1.4. Research motivation and objectives

The key challenges towards the HL-LHC era encompass managing increased data rates and complexity, developing new algorithms, optimizing software performance and efficiently managing vast resources across the WLCG. Additionally, addresses storage limitations and higher I/O rates, scaling the current system, and accommodating the needs of the multiple experiments involved. This Thesis studies the evolution of data management and storage for the CMS experiment, and in particular the effects of including CDNs in the Spanish region. It evaluates novel ideas for storage management, organization, and access to data. The results will help improve application efficiencies and minimize storage costs at the region.

In this work, the primary objective is to prove that data access can be improved by deploying a CDN that can serve data to several WLCG sites that are placed at low latencies and short distances (up to ~500 km to ensure typical low latency provided by networks). Furthermore, these objectives encompass the benefits mentioned in the previous section: the CPU efficiency of applications run in the compute nodes, storage savings, latency hiding to access the data and the reduction of network traffic. The primary motivation is to contribute to the evolution of the CMS experiment's storage system and data management, addressing the computing challenges faced by the WLCG community in the context of HL-LHC.

Our research has focused on exploring the novel paradigm of the WLCG Data-Lake model, explained in detail in the previous section. The integral design of this model brings together several aspects that are the subject of research in this Thesis. This framework has allowed us to evaluate, identify and propose improvements for the CMS data management in Spain, also looking forward to overcoming the challenges posed in the HL-LHC era. Therefore, basing part of the research on the WLCG Data-Lake model has allowed us to demonstrate the viability of a similar solution for the Spanish region (and potentially similar scenarios) that coincides with the objectives outlined in this Thesis.

Looking forward to deploying WLCG Data-Lake concepts in the region, all the new elements have been tested, and the benefits evaluated using PIC and CIEMAT Tier-1 and Tier-2 sites in Spain as example. However the main focus of the Thesis is centered on the exploration of the potential benefits of CDNs. During the development of the Thesis, WAN connectivity between PIC and CIEMAT increased from 20 to 100 Gbps in RedIRIS [56], with a round-trip time (rtt) latency of $\sim$9 ms. Regarding these conditions, both sites are suitable candidates to perform the experimental procedures carried out in this Thesis. The procedures and techniques might be extended to include the remaining Spanish site, IFCA, placed in Santander. The optimal cache sizes at the sites, the feasibility of serving the whole region with a single cache, and the effects on network connectivity for the Spanish CMS sites are studied in detail. To achieve the mentioned objectives, we aim to study in detail the current data access and usage patterns and propose, develop, evaluate, and test novel ideas on storage management, organization, and access using the HL-LHC expectations of the CMS experiment as a baseline.

Our research begins by analyzing CMS data usage and access patterns from executed jobs and storage services at PIC and CIEMAT, identifying limitations, and exploring potential improvements. The impact of incorporating data caches in the region is evaluated through simulations and modeling of caching systems based on real access records at PIC and CIEMAT. This enables the identification of workflows with files suitable for caching, allowing for optimal cache dimensioning and configuration. In parallel, a testbed of a regional cache for PIC will be deployed, contributing to service robustness and monitoring. Job efficiency improvements and deployment strategies have been evaluated, including assessing the impact of caching on real execution tasks run by users in PIC and controlled job tests. The goal is to demonstrate that the performance gains observed in controlled tests translate to real-world user jobs. This research aligns with the objectives set by the RES (Spanish Supercomputing Network [57]) project to deploy a cache service for CMS in the Spanish region (grant DATA-2020-1-0039). Also, the cost benefits of transitioning to the new model are evaluated. Nevertheless, the crucial objectives of this work are to improve applications efficiencies and minimize the deployment costs for storage services in the region.

The impact of the proposed implementations explored in this Thesis have been positive and hence they are currently deployed in production. In this way, they can be evaluated at large scale and further improved during the LHC Run3. Finally, our work is expected to lead to improved data management and storage systems for the CMS experiment. The knowledge gained from this research has broader implications for other regions and other data-intensive scientific endeavors as they currently explore some of the ideas explained in this Thesis.

1.5. Organization of the document

This Thesis covers the potential benefits of caching systems for the CMS experiment in the Spanish region, extrapolating the results to other regions with computing sites holding similar requirements. In Chapter 2, an overview of the fundamental elements of the WLCG architecture is provided and how these components integrate to enable distributed computing for LHC experiments. Subsequently, Chapter 3 delves into the experiment at the core of this Thesis: CMS. This Chapter explains how data is generated within the detector, ultimately processed and made ready for analysis by scientists.

Chapter 4 exposes the computing challenges that WLCG faces in preparation for the HL-LHC era. The proposed solutions from the community are discussed in order to justify the choice of data management as the focal point of this Thesis. Moving forward, Chapter 5 presents various models and uses of CDNs employed within WLCG, evaluating which one aligns best with the requirements of the deployment intended for CMS in Spain. Subsequently, Chapter 6 offers an overview of the Spanish CMS sites, their computational resource allocation, and the technical details of the XCache [58] (the XRootD's data cache service employed during this Thesis) deployment in Spain.

Up to this point, the context and scientific framework of the Thesis has been exposed. Starting from Chapter 7, we present the unique research contributions. The physical deployment of XCache, as discussed in Chapter 6, can be considered as one of these contributions. Chapter 7 presents studies regarding data usage and access patterns at PIC and CIEMAT, both from a storage perspective and in the context of job execution. Chapter 8 details benchmark studies conducted in a controlled environment to demonstrate how XCache enhances job efficiency by bringing data for local use. Chapter 9 showcases the studies and results supporting the idea that XCache has significantly improved job efficiency for analysis tasks by end-users in a production environment. This chapter also provides estimations of the expected economic impact of XCache as an efficiency measure for delivering data to jobs. In Chapter 10, the results of cache simulations using real CMS data are presented. These simulations aim to evaluate and fine-tune aspects such as cache sizes, allowing us to assess the impact of different configurations on XCache without the need for production environment testing.

Finally, Chapter 11 serves as the culmination of our research and the proposed approaches presented in this Thesis. Furthermore, it explores potential ways for future work, given the new opportunities that have emerged within this field.

Chapter 2

The WLCG architecture

Several HEP and non-HEP experiments utilize the Grid technology to transparently and securely access storage and computing resources from different administrative domains that operate autonomously, which conform the Grid infrastructure. This access must be reliable from different locations and preferably through a single interface. To facilitate this, institutions and individuals, together with their available resources, are grouped into Virtual Organizations (VOs). A VO refers to a dynamic set of individuals or institutions defined around a set of resource-sharing rules and conditions. All these VOs share some commonality among them, including common concerns, services and requirements, but may vary in size, scope, duration, sociology, and structure. The experiments at the LHC have developed many common services to better integrate their workload management systems into the Grid infrastructure. Some of these services, such as FTS, Rucio, CERN Virtual Machine File System (CVMFS) or DIRAC are used in WLCG and even beyond by other sciences. Most of these services are built from collaborative efforts from scientists in the experiments, and they are Open Source projects.

2.1. What defines a grid-like system?

Since its first inception in the early 2000's the Grid has been deployed and adapted accordingly to include advances of security, data, computing and network technologies, as well as to adapt to the evolution of the underlying experiments needs. Grid infrastructures, in general, must accomplish these characteristics:

- **Scalable:** It should be able to continuously deploy resources to keep up with the growing demands from the experiments, while keeping the same level of service reliability.

- **Interoperable:** The system capacity to integrate heterogeneous resources and applications.
- **Secure:** A strong security for services and access framework, safe-keeping user data and protecting against unauthorized requests.
- **Available:** A Grid system is capable of persevering through service failures while maintaining uninterrupted operations.
- **Manageable:** It also has to be user-friendly, while maintaining unified and reliable interfaces.

In addition to these core characteristics, WLCG has enhanced its functionalities as a result of the advances in computing technology. Some of these cover:

- **Heterogeneity:** WLCG operates as a heterogeneous computing system, counting with different processors and computing resources alongside the predominant x86 architecture [59]. New elements include GPUs, the use of new CPU architectures, such as ARM [60] and POWER [61], or the feasibility of use of Field Programmable Gate Arrays (FPGAs) for offline processing. The use of heterogeneous computing has implications for software development, Physics validation and optimization.
- **Virtualization:** By adopting virtualization tools, a Grid system improves its resource management, isolation, and efficiency. It enables multiple virtual machines to share physical hardware, providing isolation between applications and reducing costs by deploying different services in the same physical machine. Some WLCG sites adopt novel technologies such as Singularity [62] and Kubernetes [63], which are containerization and orchestration technologies used to manage containerized workloads efficiently with scalability. Kubernetes, for example, allows sites to manage and deploy services using containers, an ideal choice to deploy them in the heterogeneous environment of hardware and WLCG.
- **HPC resources:** The integration of HPC resources into the WLCG is underway, with some successful cases reported. However, challenges remain, such as the integration of HPC systems with the WLCG workflow and data management services. HPC centers are more successful when their site architectures are similar to the generic x86 used in the WLCG Grid. The use of HPC resources has been reported by LHC experiments, contributing to the overall computing power. Nowadays, all LHC experiments have reported using some HPC resources with different degrees of success and technical difficulty. Some cases reported encompass CINECA in Italy integrating the Marconi A-2 HPC resources for running LHCb workflows [64], Berkeley lab for ATLAS in USA [65] and, currently, an up to 50 per cent of the computing power dedicated to simulation for

the CMS experiment is executed at HPC centers, such as Barcelona Supercomputing Center (BSC [66]) in Barcelona [67].

- **Cloud computing resources:** One of the earliest adopted paradigms in the industry and Grid frameworks is Cloud computing. Within this paradigm the computing resources and services, such as storage, computing and some applications are delivered over the Internet. Generally, they share similar characteristics with Grid-like environments, such as scalability, flexibility and user-friendly interfaces. By integrating Cloud computing, a Grid system boosts its on-demand resource accessibility, enhancing flexibility and scalability without managing the physical infrastructures. However, relying on Cloud computing poses some challenges, such as the high bandwidth utilization to connect users to services or the high costs of usage of services set by the providers. In the LHC community there is the example of the ATLAS-Google R&D TCO [68] project, which aims to improve the efficiency and scalability of data processing and analysis for large-scale experiments like ATLAS. This includes developing new tools and techniques for data management, automation, and virtualization, as well as exploring the potential of Cloud computing and Machine Learning.

2.2. The WLCG main infrastructure

The deployment and allocation of resources in WLCG is governed by a Memorandum of Understanding (MoU) [69] and the resources growth is reviewed and approved by the Resource Review Board (RRB [70]). The MoU is a framework for participating countries to establish their roles, responsibilities, and resource commitments in supporting the LHC experiments. The allocation of resources is officially confirmed on an annual basis through the RRB, where countries certify their commitments (the so-called pledges) through their respective funding agencies. If a new federation wishes to register, or leave WLCG, the national funding agency is required to sign or re-sign the MoU, in both cases, and procedures exist for both registering and leaving the international collaboration.

The WLCG infrastructure is adaptable and scalable, meeting the evolving demands of the LHC experiments, while maintaining higher reliability. The fundamental components of the WLCG can be categorized as:

- **Workload Management Systems (WMS):** The WMS schedules and oversees job execution at the WLCG sites, connecting users to appropriate resources and ensuring execution task completions.
- **Data Management Systems (DMS):** DMS is responsible for secure data storage and management, granting users access to essential data for their analyses, keeping its integrity and availability. The system includes a data catalog, which

keeps track of data location in real-time.

- **Authorization and Authentication:** these components ensure that WLCG resources remain accessible only to authorized users, reinforcing system and data security.
- **CVMFS:** a distributed file system used for the distribution of LHC experiment software versions. This system incorporates an effective caching mechanism for software versions, in a tiered structure, being CERN the source for software distribution to the rest of CVMFS endpoints.
- **Conditions Database (CondDB) [71]:** a central repository for conditions data of the state of the detector, calibration or related data within different periods of the LHC Runs. In the case of the larger experiments of the LHC, such as the four big detectors, they access the conditions data through a Frontier with squid caches deployed in WLCG sites where the compute nodes are located [72].
- **Monitoring Services:** performance insights are gained through monitoring services, enabling the measurement, analysis, and troubleshooting of system performance issues (tickets). These services are necessary in order to identify and resolve any operational issues. There are central WLCG and dedicated experiment monitoring views, and a ticketing system (GGUS [73]) which is used to track down problem resolutions.

2.3. The Workload Management

Workload management has significantly changed since the first inception of the WLCG. Initially, a broker-resource system was employed, where users would contact a service responsible for determining the optimal job execution location. This optimization was based on their specific requirements and the available resources. However, this model turned out to become complex when several sites were involved, facing numerous challenges in dynamically adjusting the priorities for the diversity of workloads. Within this model, placeholder jobs called pilots are submitted to available resources. When a pilot job starts running, it contacts a central task queue managed by the experiment, which decides which job should be running next. Another advantage of this model is that jobs can determine its environment and communicate that to the task queue, and priorities can be set dynamically within the experiment. Priorities can be dynamically set and real-time adjustments can be made based on the available resources and immediate requirements. This management is carried out by the WMS, the responsible component for distributing and managing computational tasks. The management system considers the requirements of the job (such as CPU time, memory, and data access needs), information about available resources, and Grid policies to make the decision. This WMS is also responsible for interconnecting the physical CPU resources, through the CEs. These CEs services act as the entry point for jobs coming from the Grid into a local computing resource or cluster. In terms of interaction with the WMS, pilot jobs are submitted through these CEs. Once the pilot job is submitted, the CE launches it on local resources. Then, pilot jobs report to a central task queue, managed by respective experiments, which tells them on which tasks to run based on set priorities.

Every LHC experiment employs its own dedicated middleware solutions for workload management. For example, the CMS experiment employs HTCondor and GlideinWMS [74], where GlideinWMS handles pilot job submission to Grid sites and HTCondor manages and schedules these jobs. ATLAS relies on the PanDA [75] system for managing both production processing and end-user analysis. The LHCb experiment employs DIRAC, which orchestrates scheduling, computation management, storage management, and dataset replications across the WLCG. ALICE, for these purposes, employs AliEn [76], which interfaces the local resources, Grid services, and job agents that run in compute nodes to download and execute the actual payload from the central task queue. All of these solutions help WLCG to efficiently distribute the vast amount of jobs that are daily executed by the experiments all over the world.

2.4. The Data Management System

The DMS at WLCG typically manages hundreds of Petabytes, handles data movement between computing sites (i.e. interconnects all of the available SEs), sets storage policies (such as disk or tape), applies data migration and replication rules, and populates the experiment data catalog, which holds million of files and data locations in real time. The mentioned SEs are intended for the storage of experiment data and provide uniform access to all storage resources. They can be composed of disk servers or pools (large file systems located on more than one disk server), or magnetic tape storage systems. In WLCG sites that run tape systems, a reserved disk buffer is used in front of the tape system to store data temporarily before it is written to tape. This size is not generally greater than 10% of the actual disk drive storage deployed, since it is intended to maximize the I/O throughput in tape servers. The European research projects leading to EGI [77] aimed at establishing a global data grid infrastructure for e-science. At the beginning, the main concepts applied within the Chevernak model [78] outlined the necessary data management services: SEs, file transfer services, catalogs and data orchestration services and performance multi-stream transfer protocols. Over time, dedicated middleware was also developed and deployed with efforts to achieve the anticipated functionality at necessary scales and costs, essentially because the translation of the model into actual deployment was not a trivial task to develop. Each of the mentioned elements constitute the infrastructure that defines the deployed data management model in WLCG.

The DMSes use several different file identifiers to uniquely identify and access files:

- **Global Unique Identifier (GUID):** This is a unique identifier that is assigned to each file when it is created in the catalog (typically a hash).
- **Logical File Name (LFN):** This is a human-readable name that is assigned to each file. Linux/UNIX rules apply for assigning a LFN.
- **Physical File Name (PFN):** This is a human-readable name which is set at each of the computer cluster storage systems to access the file locally. Linux/UNIX rules apply for assigning a PFN.
- **Storage URL (SURL):** This is a widely usable URL of the file within the SE where it is located.
- **Transfer URL (TURL):** This is a URL that can be used to actually transfer the file from/to an SE.

Within the WLCG framework, SEs facilitate data access through various protocols and interfaces. Main protocols that stand out in the WLCG context are:

- **GridFTP:** Securely transfers files across Grid storage servers. The wide support for GridFTP across all SEs in WLCG ensures compatibility and interoperability among the storage systems and data management tools used in the WLCG infrastructure.
- **HTTPs:** Secure variant of HTTP [79]. As widely supported as HTTP while improving its predecessor's security.
- **XRootD:** Widely used in the WLCG infrastructure, XRootD provides low-latency and high-throughput access to data through its protocol of the same name. The redirector infrastructure that CMS has built using XRootD directs client requests to the appropriate data servers. It evaluates each request and forwards it to the most suitable data server to provide the requested data to the client. This process distributes the load across multiple servers, reducing latency and improving overall system performance.

WLCG's SEs also hold the GSI protocol [80], through X.509 [81] certificates for user authentication and encrypted data transfers. GSI secures GridFTP, HTTPs and XRootD transmissions over the grid resources. However, the experiments are currently migrating X.509 to JSON Web Tokens (JWT [82]) authentication, as part of the evolution of the computing infrastructure and according to the phase-out of GSI service.

2.4.1. *The Storage Elements (SEs)*

The WLCG SEs are distributed storage systems that provide high-throughput, reliable, and secure storage for the vast amounts of data generated by the LHC experiments. The worldwide distributed SEs are interconnected by high-speed networks, allowing for efficient data transfers and data access from anywhere on the Grid. A set of protocols for data access are available, as well as many services that have been developed to efficiently manage data transfers across WLCG sites.

The Storage Resource Manager (SRM) [83] is a service deployed on each SE at the WLCG infrastructure. It provides an efficient interface to the heterogeneous storage solutions that are adopted by WLCG sites. This system manages control capacities of data ingestion and data exports from the SEs, and it is equipped with redundant interfaces for each of the protocols that are supported at the site. Since the DMSes of the experiments and services such as FTS can manage data transfers efficiently, the use of SRM to access data on disk servers has been reduced (the DMSes interact directly with the protocol interfaces). However, the SRM is still a valid service which is used to manage data on tape. A new tape REST API [84] is being developed, with the aim to deprecate the use of SRM service before the HL-LHC.

The WLCG SEs are typically implemented using a variety of storage technologies to handle disk or object storage arrays, Cloud storage services, and tape libraries, services that are provided by diverse vendors. The specific local storage technology used depends on the requirements of the SE, such as the type and volume of data being stored, the desired performance and reliability, and even the budget availability, software support, or other local constraints. All of the deployed SEs offer multi-protocol access to a scalable data-store with support for the authentication methods used in the Grid, such as X509 and JWTs.

The main SE technologies used at the WLCG sites are:

- **dCache [85]:** A distributed storage system designed and developed by a collaboration between Deutsches Elektronen-Synchrotron (DESY [86]), Fermi National Accelerator Laboratory (FNAL [87]) and the Nordic e-Infrastructure collaboration (NeIC [88]). This product started development in the 2000s, and it is very popular in WLCG, and non-HEP communities. dCache provides a single virtual filesystem tree with a variety of standard access methods (including POSIX [89] and NFSv4.1 [90]) with high-performance data access. Depending on the persistence model, dCache provides several methods for exchanging data with tape storage systems (deployed with several technologies) as well as space management, pool attraction, dataset replication, hot spot determination and recovery from disk or node failures.

- **EOS [91]:** An open-source distributed disk storage system developed at CERN, since 2011. The Elastic Object Storage (EOS) provides a service for storing large amounts of Physics data and user files, with a focus on interactive and batch analysis. It supports thousands of clients with random remote I/O patterns with multi protocol support (WebDAV [92], CIFS [93], FUSE [94], XRootD, and gRPC [95]). EOS includes tape storage in combination with the CERN Tape Archive (CTA[4] [96]) software. At CERN, the current EOS instance manages 900 PB of data, and this technology has been adopted by some Tier-1 and Tier-2 sites in WLCG.
- **StoRM [97]:** The Storage Resource Manager (StoRM) is a light, scalable, flexible, high-performance, file system independent, storage manager service (SRM) for generic disk based storage systems. StoRM is a product developed by CNAF [98], which is currently co-funded by the EOSC-hub project [99], and it works on any POSIX filesystems.
- **DPM [100]:** An Open Source software for disk-based solutions developed by EGI and WLCG. The Distributed Data Management (DPM) implements a lightweight storage solution, which has been adopted by a multitude of Tier-2 Grid sites. This service has been developed and supported as a collaborative effort, but the lack of recent support has set an End-Of-Life (EOL) by the end of 2023. Around 50 Tier-2 sites are migrating their SEs to other technologies, mainly dCache, EOS, or direct XRootD servers or data cache services.
- **Proprietary storage management systems:** Some sites adopted licensed products, such as GPFS [101] (IBM), or Lustre [102] (Sun Microsystems). These are advanced clustered file systems, mainly designed for HPCs, that have been integrated into the Grid.
- **Ceph [103]:** Ceph is a distributed open-source file system designed with aims to be focused in the Big Data area. Ceph is POSIX-compatible and has a fault-tolerance system on its data replication. As a software defined storage system, Ceph stands for being a very reliable storage system regardless of the infrastructure it is run on.

To manage tape resources locally, several services are also used in WLCG, some of which are open-source and collaborative efforts, such as CTA, Enstore [104] or HPSS [105]. Other sites with tape storage systems use proprietary softwares, such as TSM (IBM) [106]. These systems are typically accessed via SRM, however a new tape REST API is being developed, which will provide a standardized HTTP interface simplifying the access tape systems. Many functionalities are being implemented, such as staging files in large bulk batches. This has proven to improve the read performance, and increases efficiency and sustainability of

[4] Not to be confused with the Cherenkov Telescope Array [39].

tape-based storage systems, in the way they are used in WLCG. The experiment DMs are responsible for managing the disk and tape resources in WLCG. The DMs interact with the SEs (using SRM, accessing directly through the protocol interfaces, or using the tape REST API), and mainly use the FTS for reliable and large-scale data transfers across the network. Checksums and retries are provided per transfer and it is a flexible tool due to its multiprotocol support (Webdav/https, GridFTP, XRootD). It also allows parallel transfers optimization to get the most from the network without saturating the SEs. It is used far beyond the HEP community, with 24 FTS instances deployed worldwide, supporting ~40 VOs.

2.5. Hardware architectures deployed and operative systems

WLCG deploys several AMD and Intel x86_64 based architectures, since most of the software and services are designed to execute on these. The x86_64 architecture is used for both the server infrastructure and the compute nodes that perform the actual data processing tasks. However, the heterogeneous number of architectures enabled are planned to expand during the upcoming years. The rest of technologies, such as GPUs, ARMs or PowerPCs are still under test, since each architecture uses a different compiler and features. It is also necessary to point out that each new architecture employed needs to pass the Physics validation tests. This validation is the process of assessing the accuracy of a computational simulation by comparing its results with experimental data. It involves checking the effects of software changes in simulation software, such as full and fast simulation, to ensure that the relationship between computation and the actual data and should not rely on the architecture used to execute the software. FPGAs are currently employed to perform specific tasks on the HLTs, that require high-performance computing workloads. It is algo suggested that their usage could expand to Tier-1 sites if processing tasks get more strictly optimized, though this is not the experiment's priority at this time.

In general, the majority of software utilized within the WLCG has historically been Open-Source. Fermilab and CERN created Scientific Linux [107] to unify OS deployment across WLCG and provide a Linux-based OS that could meet the software needs of the experiments. However, this OS has been discontinued for various reasons. Notably, the rise of CentOS [108], an open-source extension of Red Hat Enterprise Linux (RHEL) [109] that gained prominence over the years. CentOS Linux presents several advantages over Scientific Linux, including reduced total cost of ownership, ongoing support, heightened stability, and reliability. Furthermore, CentOS has demonstrated compatibility with containerization, facilitating the operation of virtual machines and Singularity [110] images for running services independent of the underlying OS. Consequently, many have transitioned to CentOS due to its numerous benefits, although the landscape shifted when RHEL announced in 2021 that support would end beyond the final release of CentOS Linux 8 [111]. This announcement prompted WLCG experiments to carefully consider the future of a dependable OS. Presently, the explored options include RockyOS [112] and AlmaLinux [113], both aiming to ensure sustained software support in the years ahead.

WLCG uses several hardware technologies for disk and tape storage to meet the requirements of the massive amounts of data generated. High-capacity HDDs, sealed with Helium for improved performance, serve as the primary means of storing active and frequently accessed data. In recent years, SSDs, flash-based storage devices, have gained relevance, providing high-speed access to frequently accessed or critical data. In the context of tape storage, magnetic tape libraries from vendors like Oracle [114] (being discontinued!) and IBM [115] are common in WLCG sites. Specifically, the standard technology utilized for tape storage is Linear Tape-Open (LT0), with LT09 being the latest tape storage technology in use [116].

2.6. Network

The LHCOPN and the LHCONE are specific private network structures that provide dedicated bandwidth and ease of management for the WLCG infrastructure, built on top of the National Research and Education Networks (NRENs) from a multitude of countries that conform the WLCG. The LHCOPN is a private optical network that interconnects Tier-0 and the Tier-1 centers. The LHCONE is a collaborative private network connecting the Tier-2 sites with the rest of the WLCG network infrastructure.

RedIris, the Spanish NREN, plays a central role in providing multiple 100 Gbps connectivity between the Spanish WLCG sites and the WLCG infrastructure, through the LHCOPN and LHCONE networks.

2.6.1. *The LHC Optical Private Network (LHCOPN)*

The LHCOPN employs a star topology, where all Tier-1 data centers have a direct connection to the Tier-0 data center, optimizing data transfer performance and reliability. The network utilizes both single and bundled long-distance redundant 10/100/400 Gbps links, ensuring high-speed connectivity among data centers while maintaining security measures to safeguard data against unauthorized access.

The current topology, as well as the bandwidth between all the Tier-1s and the Tier-0 is displayed in Fig.2.1 (as taken from [117]). To efficiently manage traffic, the LHCOPN employs BGP routing. This approach guarantees even distribution of traffic across the network and prioritizes critical traffic, enhancing network performance.

Additionally, the LHCOPN is designed in dual-stack, supporting both IPv4 and IPv6 traffic. However, the plan is to complete the migration of underlying services to IPv6 entirely soon. Currently, nearly all traffic is done in IPv6, and IPv4 remains in use due to the limited compatibility of a few software applications with IPv6. The provision for Tier-1-Tier-1 transit via the Tier-0 data center further enhances adaptability and expandability of the network architecture.

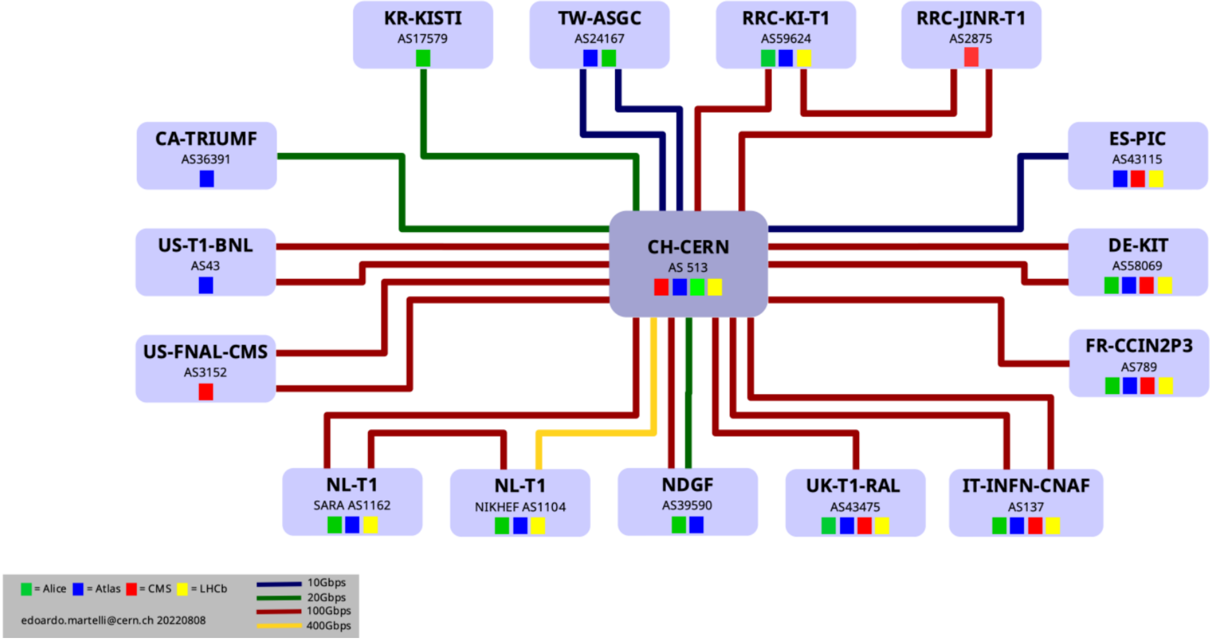


Figure 2.1: Topology of the LHCOPN interconnecting Tier-1 sites with the Tier-0 at CERN.

2.6.2. *The LHC Open Network Exchange (LHCONE)*

The LHCONE is a private network that connects Tier-1 and Tier-2 sites around the world. It is a collaborative effort among NRENs to share the use and cost of expensive network resources. The LHCONE is open to other HEP collaborations, and it serves any LHC site according to its needs and allows them to grow. In the case of this network, it employs high-speed links to ensure rapid and dependable connectivity across LHC sites, a critical aspect for seamless data transfer and analysis, especially with the substantial volume of data generated. This network stands out for its adaptability, allowing customization to the unique needs of LHC sites. Sites have the flexibility to utilize the LHCONE for connections to the Tier-0 data center, other Tier-1 sites, or both, including solutions accordingly, including the dual-stack to interconnect with sites, such as the LHCOPN. Most of the traffic in this route is IPv6 as well. The LHCONE also offers a cost-effective approach for LHC sites to access high-speed network resources. By sharing network resource usage and costs, the LHCONE significantly mitigates expenses for individual sites, promoting efficient resource utilization.

2.6.3. *New features in the WLCG network*

In the context of load balancing between the LHCONE and the LHCOPN, ongoing discussions are focusing on optimizing the distribution of loads between the two networks. The NOTED [118] project aims to improve WAN network bandwidth utilization by balancing traffic across multiple paths rather than overloading the best path. The project has developed a software toolbox and a prototype network controller to automate this load balancing process. The project has successfully demonstrated its effectiveness in improving network utilization for Rucio data transfers via FTS by balancing traffic across the LHCOPN and LHCONE connections. Concurrently, several working groups are actively addressing the implementation of packet marking schemes for IPv6. This initiative aims to enhance the understanding of IPv6 traffic, contributing to overall performance improvements and effective load balancing. Furthermore, efforts are being made to integrate Software Defined Networks (SDNs) into the new network functionality projects within the LHCONE and the LHCOPN. The SDNs provide a flexible and programmable approach to network management, from the underlying applications, enabling efficient traffic optimization and data delivery across WLCG sites.

2.6.4. *Global usage of LHC network resources*

The LHCONE and the LHCOPN networks have moved ~60 GB/s on average during the last year period, as shown in Figure 2.2. This translates to moving around ~5 PB/day in the global WLCG infrastructure using these underlying private networks. The computing model for LHC experiments during the HL-LHC era demands a substantial boost in network bandwidth capacity. ATLAS and CMS will generate about 350 PB of annual data per experiment, requiring higher real-time transfers to Tier-1 data centers as compared to current values. This requires a total bandwidth of 4.8 Tbps from CERN to the Tier-1s, including 1.25 Tbps transatlantic connectivity. To achieve this, the larger Tier-1 data centers are aimed to connect to the LHCOPN and LHCONE at 1 Tbps, for each of the networks.

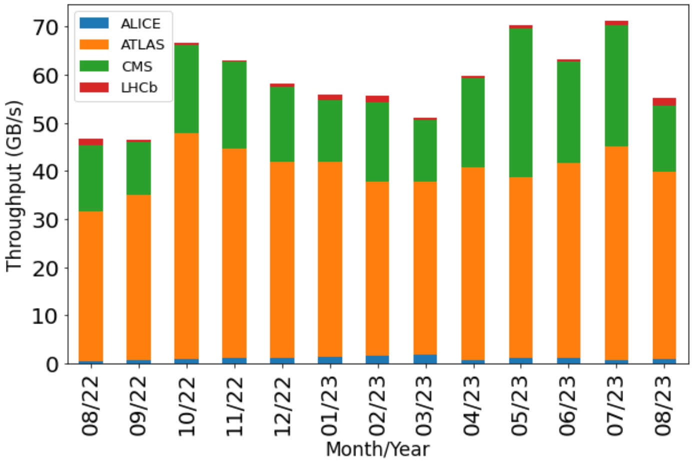


Figure 2.2: Monthly transfers by LHC experiments for the year 2022 and 2023. Source: CERN MONIT.

2.7. Security and authentication

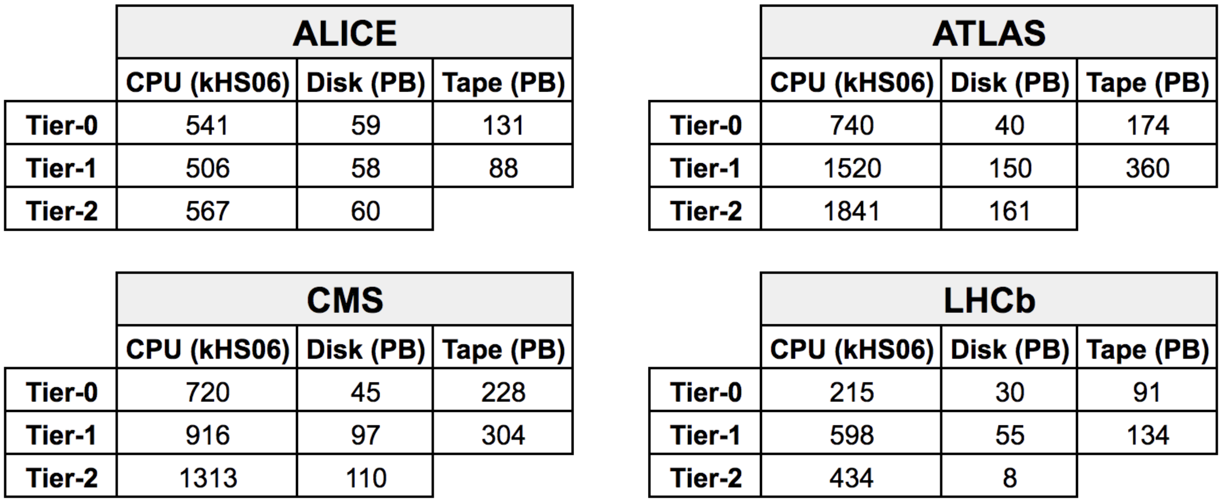
Security, authorization and authentication to access the WLCG infrastructure is based on commercial standards. X.509 certificates has been the method used within WLCG since its inception, and currently WLCG is transitioning to JWT, which aims to substitute the use of X.509 by the end of LHC Run3. These methods provide the critical foundation for verifying user identities and ensuring secure data transfers within the distributed and collaborative ecosystem of the WLCG. Each of these approaches offers distinct features and advantages, catering to diverse needs and striking a balance between security, convenience, and scalability.

The Virtual Organization Membership Service (VOMS [119]) is a Grid attribute authority which serves as a central repository for VO user authorization information, providing support for sorting users into group hierarchies and keeping track of their roles and other attributes. With the use of X.509 standardized digital certificates, the users can obtain a VOMS proxy containing their roles, group memberships and assertions used in the Grid environment for authorization purposes. This ensures that only authorized users can access the data and services provided by the Grid infrastructure for a particular VO (i.e. experiment). The X.509 certificates are issued by trusted certificate authorities (CAs). In addition to envisioning a novel frontier, JWT tokens are an evolving authentication method fitted for WLCG, a method that better suits when integrating external resources (Cloud, HPC, etc...). They are engineered for enhanced security and efficiency as compared to X.509 certificates, and are based on OAuth [120] & OpenID [121] standards. The INDIGO IAM [122] and EGI checkin services [123] are Identity and Access Management (IAM) services developed to expedite tokens and resemble the VOMS functionalities. WLCG has a clear timeline to migrate to JWT, a migration that affects many services that need development to support JWT [124].

2.8. Deployment of WLCG resources

CERN and the countries represented by their funding agencies hosting computing resource centers sign the MoU that governs the WLCG collaboration. The WLCG MoU defines the service levels expected by the participating sites and determines the resources to be provided by each center. It also establishes the existence and role of the LHC Experiments RRB, which meets twice per year, in spring and autumn. Through these meetings, the RRB oversees and officially approves both the computing resources required by the LHC experiments and those pledged by centers. The RRB is a high level committee, in which the CERN director, the RRB evaluation committee and national funding agencies are represented. Table 2.1. shows the resource types pledges at the Tier-level for all of the LHC experiments and for the year 2023 as taken from CRIC portal [125]. The total pledged resources for 2023 in WLCG are ~9.9M HS06 (which corresponds to approximately ~750k CPU cores), ~870

PB in Disk, and ~1500 PB in Tape. HEP-SPEC06 (HS06) is a widely used benchmark for evaluating the computational performance of systems in HEP. It measures the processing power of a CPU core using a standardized workload, allowing for effective comparison of computational capabilities across different systems [126]. The conversion of HS06 to walltime (core·h) depends on the specific workload and system efficiency, as walltime represents the actual time taken to complete a task using a given number of CPU cores. This conversion is influenced by factors like task complexity, code efficiency, and computing architecture. For analysis execution tasks in CMS, the approximate conversion is 12.06 HS06·core per 1 walltime core·hr at Tier-1s[5]. Figure 2.3 (left) shows how the resources are distributed in the Tiers. Around 20% of the CPU and disk resources are deployed at the Tier-0 at CERN. For Tape resources, around 40% of the space is deployed at the Tier-0 at CERN and the rest at the Tier-1 sites. Most of these resources are assigned to the ATLAS and CMS experiments, as seen in Figure 2.3 (right). The WLCG pledged resources are distributed worldwide. Spain currently pledges 4% of the WLCG resources at Tier-1 and Tier-2 sites, supporting the ATLAS, CMS, and LHCb experiments. The main countries providing resources to WLCG, aside CERN, are the United States of America, Italy, Germany, United Kingdom and France. The number of participating countries in WLCG exceeds 40 countries.



	ALICE		
	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier-0	541	59	131
Tier-1	506	58	88
Tier-2	567	60	

	ATLAS		
	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier-0	740	40	174
Tier-1	1520	150	360
Tier-2	1841	161	

	CMS		
	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier-0	720	45	228
Tier-1	916	97	304
Tier-2	1313	110	

	LHCb		
	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier-0	215	30	91
Tier-1	598	55	134
Tier-2	434	8	

Table 2.1: Resource types pledges at the Tier-level for all of the LHC experiments and for the year 2023. Source: CRIC Portal.

[5] While HS06 has been the primary benchmark for HEP experiments since 2009, it is being replaced by HEPScore to account for advancements in hardware technology. However, this work refers to results as HS06 due to the timing of the benchmark's deprecation.

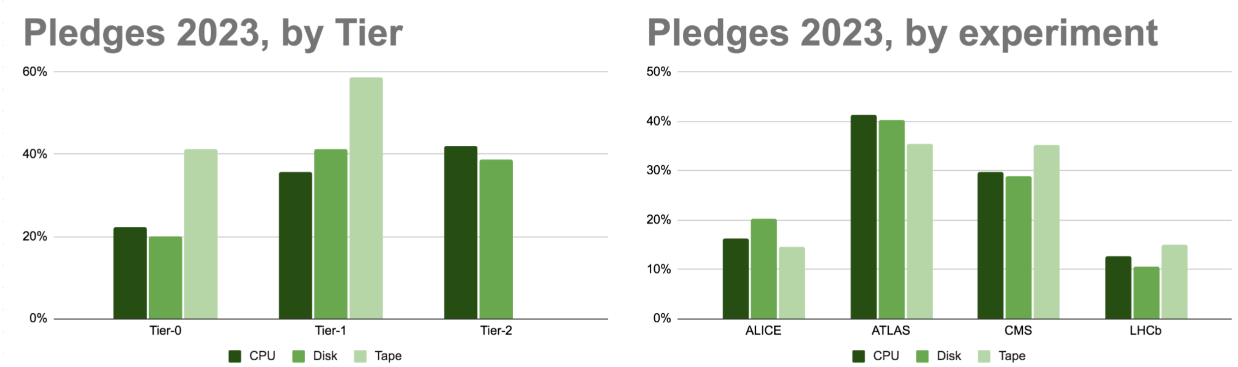


Figure 2.3: CPU, disk and tape pledges committed to LHC by tier and by experiment in WLCG.

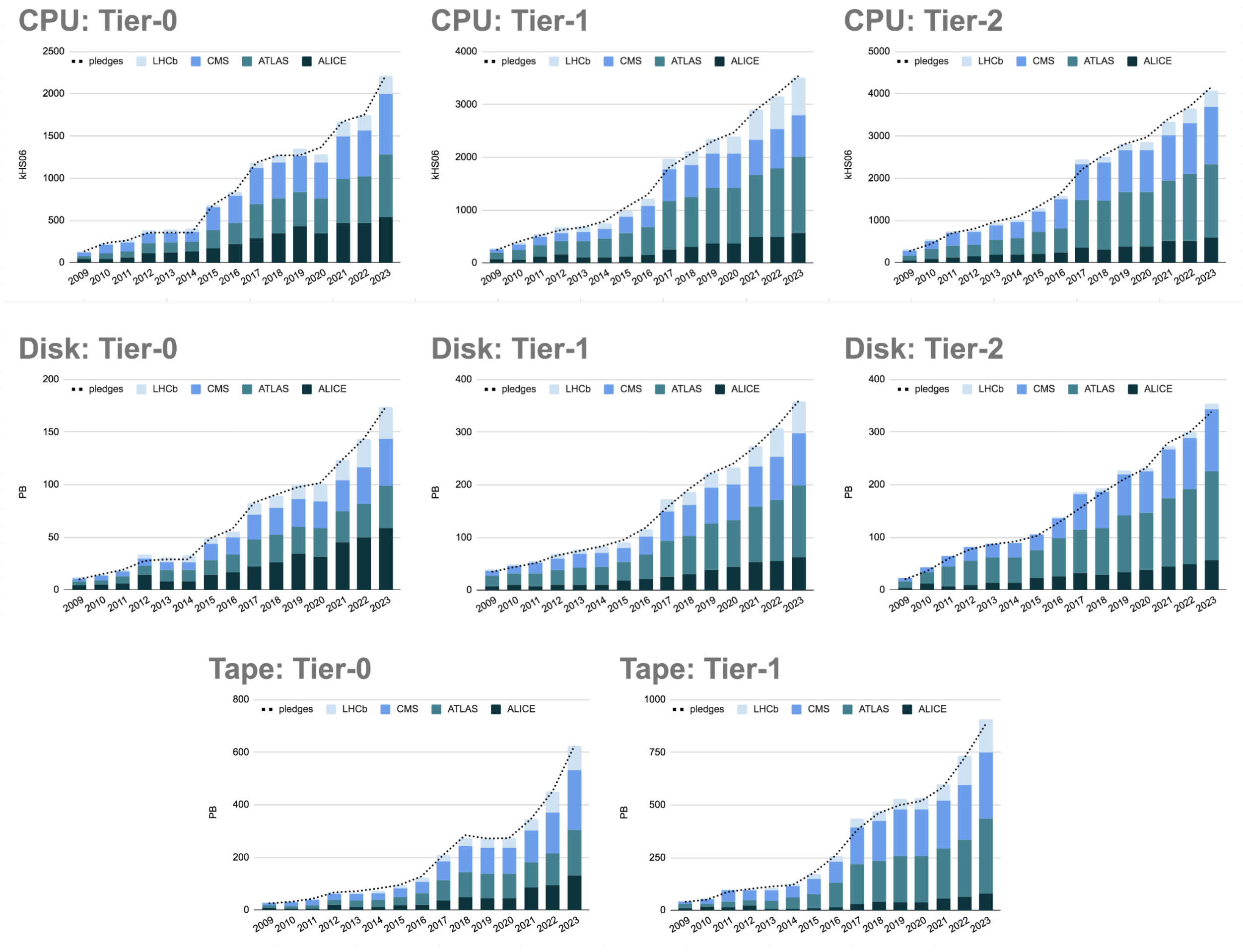


Figure 2.4: CPU, Disk and Tape resources deployed at the Tier-0, Tier-1 and Tier-2, since 2009. The dashed line represents the WLCG site's pledges. Source: CRIC Portal.

The resources in WLCG have been growing in time to accommodate the experiment's requirements, based on the LHC data delivery expectations. Figure 2.4 shows the resource deployment evolution for CPU, Disk and Tape, as compared to the pledge resources provided by all of the countries in WLCG. The pledges have typically fulfilled the requirements from the LHC experiments. It can be observed that the resource increases flattened when the LHC was in shutdown periods. For CPU resources, the experiments typically have access to opportunistic CPU usage beyond the pledges. This is due to the fact that some CPU servers are extended in operation far beyond their expected lifetime, or by the use of CPU cycles not reserved to WLCG at large scale clusters in the Universities where most of the Tier-2 sites are deployed, or by the use of HPC resources. For both ATLAS and CMS, the used CPU resources currently almost double the pledges committed to WLCG. The current CPU usage in WLCG corresponds to an average use of around 1M CPU-cores worldwide. This, together with the deployed >1 Exabyte storage, makes the WLCG to be the largest scientific computing infrastructure ever built.

Chapter 3

The CMS experiment

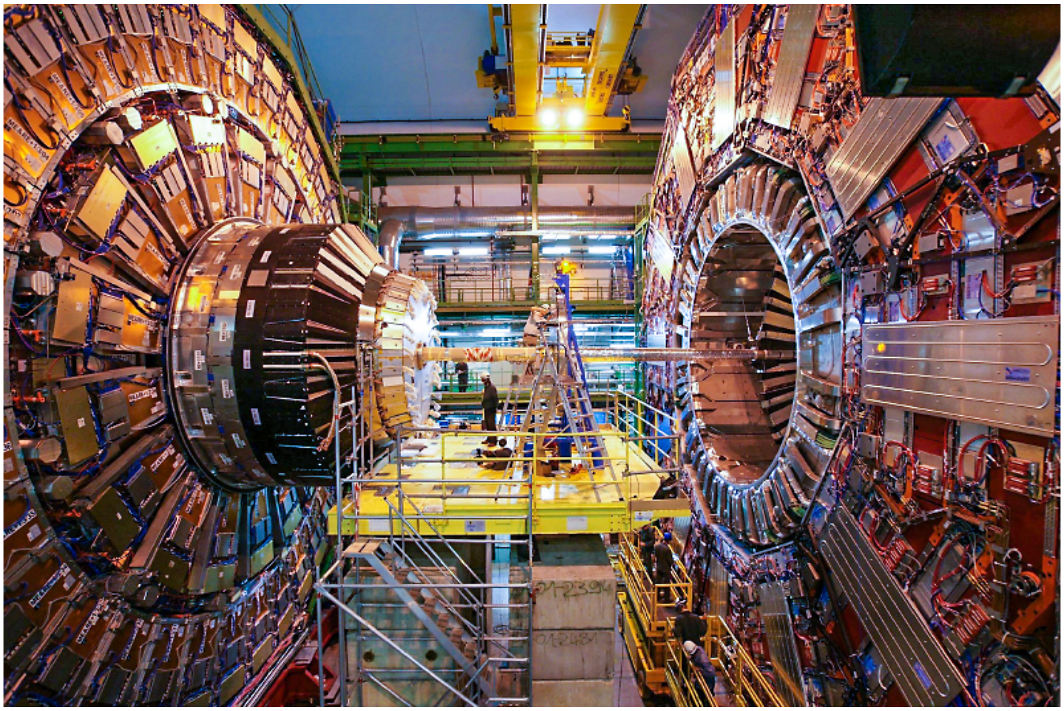
The CMS experiment is one of the most successful experiments in the history of Particle Physics. It has made significant contributions to our knowledge of the universe, including the discovery of the Higgs boson. The CMS experiment is still running and is making new experiments to deepen our understanding of particle Physics.

- Peter Higgs, Nobel Prize in Physics laureate in 2012, who proposed that broken symmetry in electroweak theory could explain the origin of mass of elementary particles, through the Higgs mechanism.

The Compact Muon Solenoid CMS is one of the two general purpose experiments in the LHC along with ATLAS. In 2012 CMS was able to find evidence for the existence of the Higgs boson, thus completing the only remaining experimental evidence to explain the Standard Model. Therefore, its multipurpose design intends to continue collecting data on the protons and ions collisions in search of BSM Physics: the nature and existence of dark matter or to delve into the conditions of the early universe in the phases immediately after the Big Bang. Particle detectors incorporate intricate engineering and the prevailing approach to their construction usually involves a layered design, with each layer optimized for maximum efficiency in detecting the different particles that are produced during particle collisions. Although the design can significantly vary upon the detector interests, HEP detectors typically comprise four main layers for unique purposes. The inner detector is responsible for measuring particle trajectories with precision, enabling accurate estimates of the momentum of charged particles by analyzing their curvature within a high magnetic field present at the detector, as well as their identification. Subsequently, the two calorimeter layers measure the energies of photons, electrons, and hadrons (such as protons, neutrons or any other particle composed by quarks) as they go through each of the layers. Finally, the outer muon spectrometers identify and measure the momentum of muons, a kind of leptons of special interest to study some physical processes of interest in HEP and BSM Physics. Nevertheless, each experiment customizes its own detector layers and tools, relying them to the specific Physics they aim to study or the precision required for their research objectives, specially in those cases where the detectors serve multipurpose functions. This chapter describes the CMS detector, which is displayed in Picture 3.1, its main components, and the data flow resulting from the experiment.

3.1. Main components of the CMS experiment

The HEP detectors typically have a layered structure, with each layer designed to detect specific groups of particles. In Figure 3.1 the CMS detector and its main components can be observed in detail. At first sight, its dimensions stand out, measuring 15 m in diameter and 21.6 m in length. Regarding its total weight, about 12500 tons, CMS is the heaviest of the four large LHC detectors, and it is located about 100 m deep underground.



Picture 3.1: The Compact Muon Collider (CMS) picture inside the cavern. Source: University of Florida [127].

In broad terms, the detector must be conceived as a great filter in each of its layers for each type of particle arising from the collisions of protons and ions. In each of those layers the particles stop, deposit their energy, and the sub-detectors measure the properties of these particles. Thanks to its specialization in each of the components of the structure, the energy and momentum of each of the resulting particles from collisions can be measured in order to identify them and, thus, identify which is the physical process that produces them.

The core of the CMS detector is formed by a superconducting magnetic solenoid (which gives its name to the detector) with a diameter of 6 meters. Its superconducting solenoidal structure is cooled to about -268.5°C and is capable of generating a magnetic field of 4T. This magnetic field is one of the largest constant magnetic fields ever generated by humans in a large volume. In terms of magnitude comparison, the Earth's magnetic field oscillates in the order of no more than 100 µT, and the spots on the surface of the Sun, which are active regions with very high electron flows, range between 0.1 and 1T. The central solenoid functions by bending the paths of charged particles, enabling the tracker to measure their momentum through the Lorentz force's strength-momentum relationship. In the beginning, particles emerging from the collisions meet the silicon tracker. A more detailed view of the CMS experiment layered is depicted in Figure 3.2.

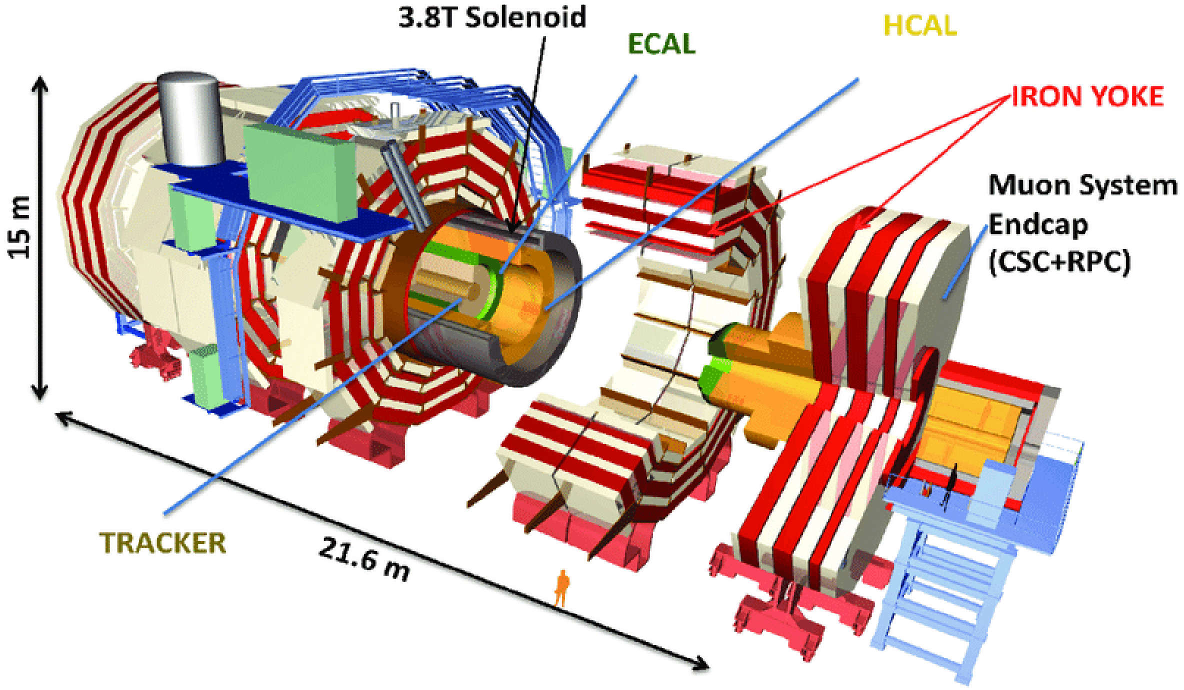


Figure 3.1: Schematic view of the CMS experiment [128].

The tracker is made entirely of silicon detection elements, including pixels at the core, which deal with the highest particle intensity, and silicon microstrip detectors surrounding them. As particles pass through the tracker, the pixels and microstrips produce tiny electric signals. These signals are amplified, detected, and stored in microchips that flush them in memory for several milliseconds. This data is then processed and converted into infrared pulses for transmission over 100-meter fiber optic cables to a radiation-free environment for analysis.

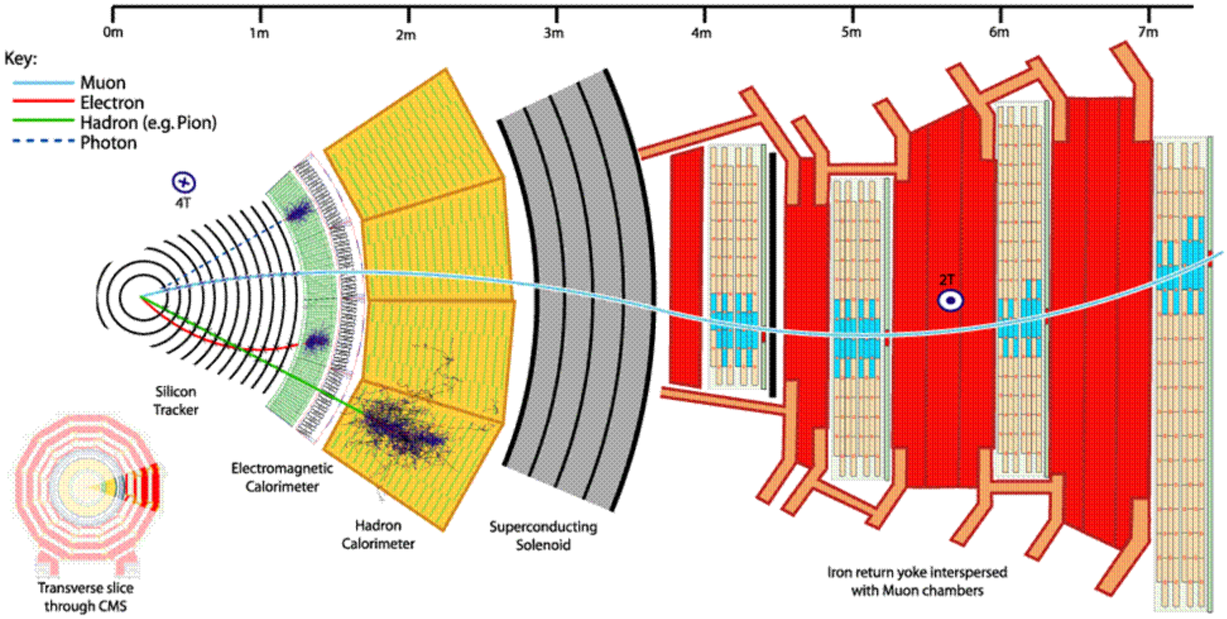


Figure 3.2: Layer organization of the su CMS experiment sub detectors. The particle specialization of each layer is also displayed through the components [129].

On top of that, the tracker has over 135 million separate electronic readout channels, covering an area about the size of a tennis court. It utilizes approximately 40,000 fiber optic links to transport the signals efficiently and with low power consumption. However, particles should be the least stopped as possible, in order to reach the next layer: the calorimeters. Since the tracker's main function is to measure the kinematic properties of the particles, calorimeters are designed to stop the movement of particles along their trajectory in order to measure their energy deposition and other characteristics.

Furthermore, to keep on recording the events produced at LHC (referred as snapshot of the collisions occurring at the detector), the calorimeters are necessary in order to study the Physics related to concrete particles. The first one is the Electromagnetic Calorimeter (ECAL), which specializes in detecting and measuring photons and electrons. The second is the Hadronic Calorimeter (HCAL), responsible for detecting hadronic particles containing quarks, such as protons or neutrons.

The ECAL uses lead tungstate crystals that produce light proportional to the particle's energy. The high-density crystals emit fast, well-defined bursts of light, enabling a compact and accurate detector. To work effectively, photodetectors are attached to the crystals to detect and convert the scintillation light into electrical signals, facilitating further analysis. The ECAL is divided into a barrel section and two end caps, comprising 61,200 crystals and nearly 15,000 crystals in the end caps.

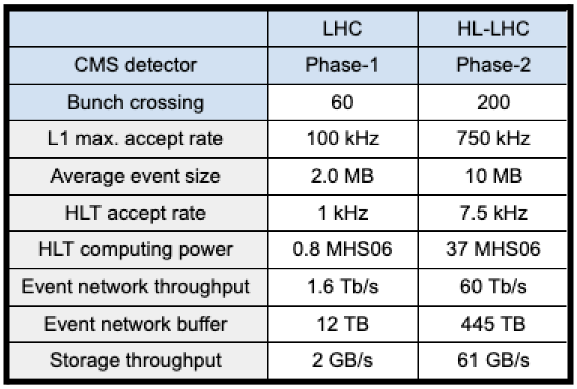
In addition to this, the Hadronic Calorimeter (HCAL) measures the energy of particles composed of quarks. The HCAL's hermetic design guarantees the detection of particles by effectively capturing particles generated from collisions. It employs a sampling calorimeter, which is composed of alternating layers of absorber and fluorescent scintillator materials. When a particle passes through these layers, it produces rapid light pulses that are collected by special optic fibers and readout boxes for analysis. Similar to the ECAL, the HCAL is organized into barrel, endcap, and forward sections. It has 36 barrel wedges and end cap wedges positioned inside the magnet coil. Two hadronic forward calorimeters are placed at either end of CMS to detect particles at shallow angles relative to the beam line. These sections use radiation-resistant materials to handle high particle energy levels.

As the name of the experiment says (Compact Muon Solenoid), the detection of muons is crucial to study the Physics processes involved in the collisions. The CMS experiment uses specialized muon chambers for muon detection. Muons, due to their properties, can penetrate several meters of material without significant energy loss. Due to their minimal interaction with other calorimeters, muons are strategically placed in the outermost part of the experiment. As it can be observed in Figure 3.2, due to its large size, inner tracker signals are much more precise than the previous ones. This allows muons, which interact little with matter, to produce larger signals and, therefore, be easier to detect. To measure a particle's trajectory, a curve is fitted to the hits recorded in the four muon stations located outside the magnet coil. The position of the particle is tracked through the multiple active layers of each station, and this information is combined with data from the CMS silicon tracker to improve precision and measure the particle's momentum. The CMS experiment includes a total of 1400 muon chambers, which are made up of different types of chambers, including drift tubes (DTs), cathode strip chambers (CSCs), resistive plate chambers (RPCs), and gas electron multiplier chambers (GEMs). Each type of chamber has its own strengths and weaknesses, so the different types are used together to create a system that is able to effectively filter the background noise.

To conclude, the data acquired by these sub-detectors is collected by a large number of channels attached to them. This total number of channels spans a total number of ~ 160 million among all the detector layers. The silicon tracker has the most channels, with 137 million, because it is the most sensitive sub detector (tracks the particle paths). ECAL and HCAL count with 18 and 3 million channels, respectively, because of the strong support that both sub detectors have from the silicon tracker measuring the trajectories. Moreover, muon chambers have 5 million channels to collect data from the most penetrating particles in the experiment.

3.2. The data acquisition system of the CMS experiment

In the nominal luminosity load regime of the LHC, proton or other particle bunches cross at a frequency of 40 MHz. This produces a large amount of data (~40 TB/s), a volume that outpaces any conventional data acquisition system. Given that individual events are in the megabyte range, the implementation of a selection system or trigger becomes necessary to cope with this order of data flow. The trigger architecture employed by the CMS acquisition system (DAQ) unfolds across multiple stages, employing dedicated hardware and algorithms specifically crafted for this task. Fortunately, the events of interest are orders of magnitude lower than the frequency of collisions. This circumstance empowers the CMS trigger system to act as a gatekeeper, enabling the collection of experimental data with precision and readying it for subsequent phases of processing, storage and analysis. The trigger system consists of two stages. The first level (L1) operates synchronously (online) with the LHC's bunch crossing frequency of 40 MHz. It is composed of custom hardware processors and utilizes information from calorimeters and muon chambers to select events within the microsecond range. Due to the increased complexity of the detector and pileup density towards HL-LHC, the trigger was subject to several upgrades to handle the increasing data streams. As a consequence, the maximum event rate allowed in the HL-LHC phase will be 750 kHz (a 7-fold increase over the current rate). Once these events pass the selection, they are processed by a two-stage system with a throughput of 100 GB/s. Upon reaching this point, the complete events are transferred to the HLT asynchronously (offline) within a computing farm, close to the detector and underground, which reduces the collection of events down to the order of O(1 kHz). Consequently, during the HL-LHC era, the events generated will have an average size of 10 MB (about a ten-fold increase of the average size in the LHC era). A detailed comparison of throughput, acceptance rates of the trigger and computing power between LHC eras can be observed in Table 2.1. These events are temporarily written to a disk buffer before being transferred to CERN's Tier-1 for subsequent offline processing. The current data delivery rate from the CMS detector is about 2 GB/s, aiming to increase up to 61 GB/s during the HL-LHC.



	LHC	HL-LHC
CMS detector	Phase-1	Phase-2
Bunch crossing	60	200
L1 max. accept rate	100 kHz	750 kHz
Average event size	2.0 MB	10 MB
HLT accept rate	1 kHz	7.5 kHz
HLT computing power	0.8 MHS06	37 MHS06
Event network throughput	1.6 Tb/s	60 Tb/s
Event network buffer	12 TB	445 TB
Storage throughput	2 GB/s	61 GB/s

Table 2.1: Acceptance rates and I/O comparison between LHC data acquisition in the LHC and HL-LHC.

3.3. The CMS Computing Model

The CMS experiment has maximized its use of WLCG resources within its computational framework. To meet distinct experiment requirements such as scalability and reliability, CMS has carefully developed and integrated its own tools throughout its evolution. As a result, the experiment uses the Grid resources with dedicated CMS services.

A clear example lies in CMS data management, where the experiment has its own independent data catalog in big Oracle’s instance at CERN, namely CMS Data Aggregation Service (DAS) [130]. This approach aligns better with the unique demands of the experiment. In terms of data transfers, CMS has engineered several applications optimized to precisely integrate to the experiment's characteristics, particularly those shared with the WLCG. Furthermore, CMS has developed its own suite of workload management tools and a robust tracking system to ensure efficient operations. Currently, 7 Tier-1 and 52 Tier-2 sites in WLCG support the CMS experiment all over 4 continents, with the exception of Oceania and Antarctica.

3.3.1. *The CMS data management*

CMS computing jobs process 100 GB each second. This data is globally stored in different locations to ensure both safety and accessibility. Mostly, this encompasses a Tier-0, Tier-1 for long-term storage and Tier-2 sites for data analysis support. Tens of thousands of samples were cataloged and petabytes of data were moved during the three runs of LHC, among the management of distributed resources and workflow management tools. The main components of the CMS Data Management System are Rucio, the data-transfer management system, and the Data Bookkeeping Service (DBS) and Data Aggregation Service (DAS). These core components are designed to work together and achieve important tasks such as data bookkeeping, data location catalog maintenance, and data placement and transfer management.

Firstly, Rucio manages global data transfers for CMS over the Grid in a robust, reliable, and scalable way. Rucio has an FTS implementation that allows managing file transfers, coordinating requests, monitoring progress, handling errors, and optimizing transfer efficiency across multiple FTS servers. Another component is the Trivial File Catalogue (TFC [131]) is a simple set of rules that is used to map logical file names to physical file names on each site. Following, the DBS is a metadata catalog that provides information about the datasets and files produced by CMS, while DAS aggregates views and provides them to users and services. These components work together to achieve some tasks as data bookkeeping, data location catalog maintenance, and data placement and transfer management.

3.3.2. *The CMS data distribution and movement*

The computing model employed by CMS prioritizes the presence of data in specific locations. These are situated within the sites of the different Tiers, in accordance with their usage and the type of workflows being executed. This adherence to the experiment's policies and priorities results in various types of data.

Managing data within the context of experiments is a multifaceted endeavor, demanding careful considerations at every juncture. The complex task of distribution involves decisions such as determining the optimal number of dataset replicas, selecting storage locations for these copies, and making informed choices between disk and tape storage. This orchestration is further complicated by the long-standing practice within WLCG VOs of aligning computing tasks with data location. This means that frequently accessed datasets must be replicated across multiple sites, with a preference for readily accessible disk storage. Moreover, the predictability of data transfer requests remains an ongoing challenge, relying on the fluctuating activities of various analysis groups and individual physicists. When big portions of archived data require reprocessing, a meticulous central coordination is imperative, including the pre-staging of data (transferring data from tape systems to disk servers) at the Tier-1 centers, where most of the reprocessing campaigns are executed.

One key component over the distribution of data in CMS is the Dynamic Data Management system (DDM) within Rucio. DDM improves the strategic allocation of data across the managed distributed storage resources by adhering to a defined set of policies outlined by Rucio users. These policies include many details, such as the targeted storage resources for data replication, the duration data should persist on each resource and the optimal timing for data deletion from these storage resources. PopDB [132] maintains historical dataset usage information, encompassing metrics like total and user accesses per dataset and total CPU time spent on each dataset during processing or analysis jobs. This dataset usage information is instrumental for DDM to replicate popular datasets across various sites. However, the replication process takes place afterwards and needs a sufficient accumulation of historical data to initiate the replication process.

Regular consistency checks emerge as a necessity to ensure consistency between the filesets and the actual content of the SEs, protecting against discrepancies. In addition, the preparatory act of storing tape files on disk groups, setting the stage for later reprocessing tasks, becomes crucial. Files are grouped to be later recalled from tape in bulk, maximizing the tape read throughput.

3.3.3. *The CMS data formats*

The first data classification involves dividing the data into two categories: Monte-Carlo (MC) simulated data generated using software like GEANT4, reprocessed later in the data reconstruction chain, and real data from proton and ion collisions recorded in the detector (Collision data or DATA).

MC simulation strongly relies on the package of GEANT4, simulating the passage of particles through matter using the MC methods. This type of data is used to calibrate the detector and to study events that are difficult to directly observe in collision data. All physical processes known are simulated, so that deviations between the simulated and real data serve as indications of New Physics. As an example, Figure 3.3 (as taken from [133]) displays a plot of the discovery of the Higgs, clearly showing that it does not appear in the simulations but does in the real data (as shown with the red curve).

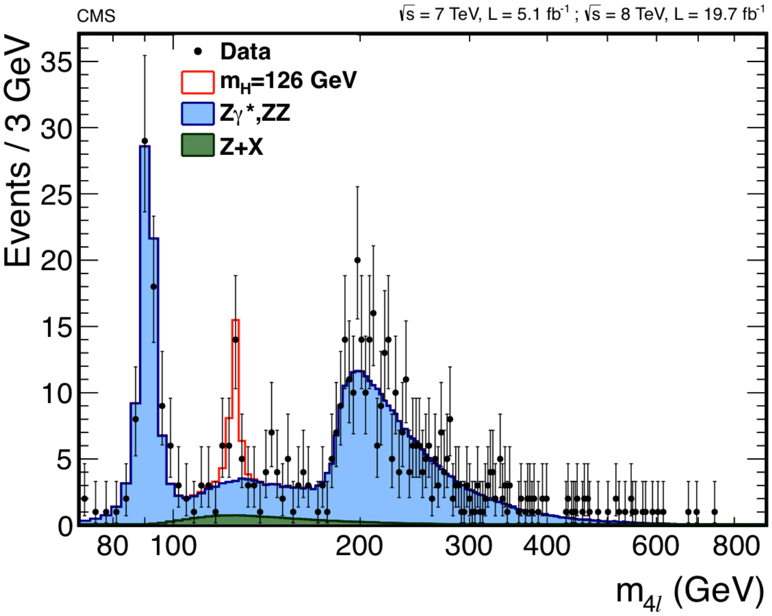


Figure 3.3: Higgs boson decay in four leptons: red peak analysis for SM verification and comparison with simulated processes.

Collision data (or RAW data) is the detector output of real collisions. This type of data is used to perform Physics analysis and to search for new particles and phenomena. To perform meaningful posterior statistical analysis, the number of simulated events is determined by the Physics goals and the statistical requirements of the specific analysis, and the MC precision is set in accordance to the detector subsystems resolutions. The different data formats used by CMS experiments are similar to the other LHC experiments. In LHC experiments, data formats for both real and MC data are meant to suit the requirements of each experiment and the analysis techniques utilized. These formats can differ in structure, organization, and the data they encompass. Nevertheless, there exists a fundamental shared structure and a set of common principles that direct the design of these formats across all LHC experiments. The data formats used for fine-tuning and calibration of the detector are the Express Data streams and the Alignment and Calibration (AlCa) streams. Within each of the Streams organization several data formats can be found. Among the different CMS Data Streams stand out the Physics Streams. This category is subdivided into Primary Datasets, each constituting a bedrock for in-depth Physics analyses. These events, grouped based on the exacting criteria established by the HLT, serve as the nexus of this analysis endeavor. This meticulous classification is orchestrated in two tiers[6]. The initial tier deploys hardware-based filters, swiftly sieving through events with rapid yet simple selections. The subsequent tier employs more intricate software-based filters, unearthing profound insights that require more time-intensive analyses. At the culmination of this intricate orchestration, events emerge in their original RAW format, the culmination of the selections guided by the HLT's criteria. Embedded within these events are the outcomes of the final HLT selection and the high-level objects borne from the processing stage. Starting from the original RAW data provided from the online acquisition system, a cascade of subsequent processing stages are performed. This event reconstruction yields a refined dataset through successive degrees of enhancement.

3.3.4. *The CMS data tiers*

The CMS events navigate through complex reconstruction and simulation chains, culminating in an information-rich synthesis. The culmination of this process includes a wide spectrum of formats, each keeping the of the experiment's meticulous step on the processing and reduction of the actual experimental data. Among the main Collision data tiers stand out the following:

- **RAW (Raw Data Format):** This format serves as a definitive record of the complete event information captured at Tier-0, situated within the confines of CERN. As an unadulterated snapshot, it preserves the raw detector information devoid of any processing. Notably, it undergoes only preliminary hardware filtering to identify events with potential interest. While the RAW format isn't intended for analysis use, it finds its home at Tier-0, a testimony to its inception. However, all the RAW data is also sent to Tier-1 sites for its safer preservation.
- **RECO (Reconstructed Data Format):** Representing the initial phase of processing at Tier-0, RECO embodies the conversion of RAW data into reconstructed physical objects within the intricate framework of the sub-detectors. This layer is similar to a canvas, spanning from "hits" to the comprehensive reconstruction of entities like leptons and hadrons. While RECO is indeed

[6] CMS refers to different data categories using the word ‘tier’, such as the WLCG uses it to characterize the different categories of their computing sites. In this Thesis, we will refer to them as ‘data tiers’ to refer to the data categories of CMS to avoid any confusion.

analyzable, its level of detail renders it for frequent or intensive study. The magnitude of data in this format can introduce significant analysis lag due to its considerable size.

- **FEVT (Full Event Data Format):** A fusion of both RAW and RECO, the FEVT format emerges as an amalgamation that brings together the raw event data and the fruits of its initial reconstruction. This dual presence reflects the comprehensive nature of this format, bridging the gap between pristine data capture and preliminary processing stages.
- **AOD (Analysis Object Data Format):** Providing more than 50% compression as compared to RECO, the AOD format encapsulates a condensed representation of low-level information extracted from the events. Designed for widespread analysis, this format strikes a harmonious balance between event size and the accessibility of vital information. AOD's strategic design optimizes analysis speed and flexibility, culminating in an invaluable tool for researchers. Complementary to this is the emergence of streamlined versions: MINIAOD, a lighter variant at around 15% the size of AOD, and NANOAOD, a remarkably compact rendition constituting less than 1% of AOD's dimensions.

Besides the collision experimental data, there is also the remaining set of MC-simulated data:

- **GEN (Monte Carlo Generated Event Format):** GEN captures the essence of Monte Carlo-generated events. Serving as a virtual experiment within the experiment, this format encapsulates events crafted through simulation to mimic real-world interactions. It is a cornerstone in theoretical exploration and hypothesis testing.
- **SIM (Simulation Data Format):** Resulting from the GEN data processing. Within SIM, the spotlight falls on Monte Carlo (MC) particles and their energetic imprints within the detector. As a manifestation of energy depositions, SIM provides a snapshot of MC particles' interaction with the experiment's sensitive elements.
- **GEN-SIM:** This format represents the first step in the data processing chain and contains the simulated event information, including the generated particles and their properties. It is a detailed snapshot of the simulated events, but it does not include the effects of the detector response.
- **DIGI (Digitized Data Format):** Resulting from the DIGI data processing. In DIGI, the transition is made from abstract "hits" to tangible detector responses. This format captures the transformation of these hits into data points that represent the detector's reactions to particles. To a large extent, it mirrors the essence of the RAW output generated by the detector itself.

- **GEN-SIM-DIGI-RECO:** Generated in a single workflow. This format includes the additional steps of digitization and reconstruction, which simulate the response of the CMS detector to the generated particles and reconstruct the Physics objects from the detector signals. It provides a more realistic representation of the events, including the effects of the detector.
- **AODSIM:** Derived from GEN-SIM-DIGI-RECO (analogue to AOD but for MC), the AODSIM format includes both the simulated and reconstructed objects, making it suitable for a wide range of Physics analyses. Likewise Collision data, AODSIM is subsequently derived and reduced to produce the MINIAODSIM and NANOAODSIM of simulated AOD data. The MINIAODSIM format is smaller and more streamlined, targeting approximately 10% of the size of the Run 1 AOD format. It is designed for more specific and specialized Physics analyses, where a smaller data size is desirable. On the other hand, NANOAODSIM is an even more compact version of the AOD format, targeting in this case $\sim 1\%$ of the size of the Run 1 AOD format. It is optimized for fast and efficient analysis, providing the essential information needed for specific Physics studies.

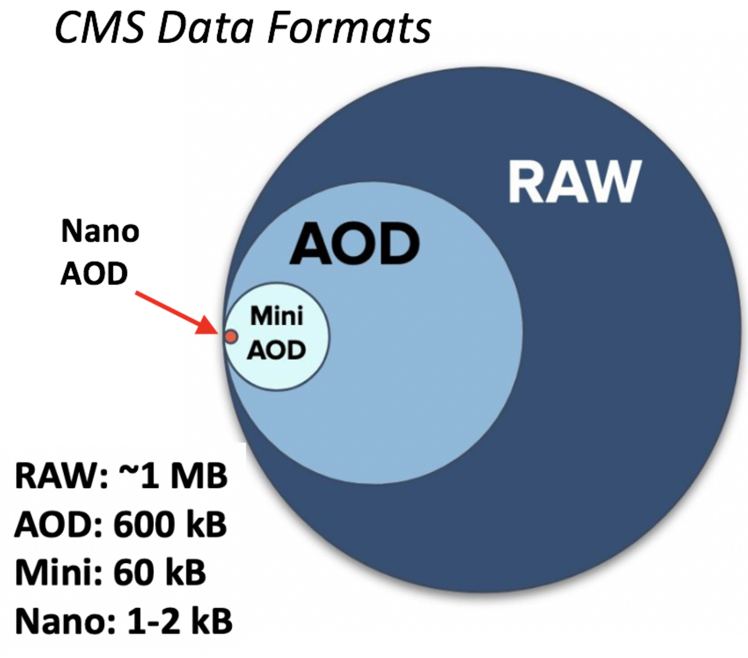


Figure 3.4: Typical event sizes of the usual CMS data tiers.

Event collections and analysis datasets play crucial roles in organizing and extracting valuable insights from large and diverse datasets. These events are snapshots of the particles bunch-crossing and all together are merged to produce the files which are organized in these data tiers. Event collections are distinct subsets of processed datasets, while analysis datasets are subsets of events tailored for specific analytical exploration. These components enable researchers to work with manageable and relevant data, enhancing the efficiency and effectiveness of their analyses.

The last organization in terms of storage for CMS are the assembled data in files and file blocks. Files serve as the fundamental building blocks of storage, encapsulating data for manipulation by various computational processes. Typically, CMS files have 2.5 GB on average including all data-tiers. In Figure 3.4 a visual recreation of the typical sizes per event in MB for each data tier can be observed and compared.

3.4. The CMS execution tasks scheduling

In CMS, central teams organize the processing of collision data and MC workflows. WMCore [134] manages the lifecycle of computational tasks, bridging the gap between CMS experiment requirements and the underlying computational resources. In conjunction, WMAgent [135] complements WMCore by overseeing task execution and resource management, optimizing task efficiency in the process. Submission infrastructure (SI [136]) uses GlideinWMS, a pilot-based workload management system, for resource allocation and matchmaking at the Grid sites. It employs a pull-based architecture, where jobs are initially submitted as placeholders and access real tasks from a central queue at the time of execution. By using this pilot approach, it provides scalability and allows schedule on a wider variety of resources, including a wider variety of resources in the WLCG, and it has been extended to use specific supercomputers, or Cloud resources, and even heterogeneous resources not using the x86 architecture (such as GPUs or ARM processors).

On top of that, glideinWMS employs a hierarchical framework where a central server oversees pilot jobs that are submitted to resource sites. Pilot jobs execute on local batch systems, and communication between resource sites and the central server completes the task scheduling loop. The central server sends payloads to pilot jobs, and sets the priorities for different payload executions. The CMS Global Pool [137] is a unified HTCondor pool that includes all WLCG computing resources dedicated to CMS, including Cloud and opportunistic resources. It serves as the primary computing resource provisioning system for all CMS workflows, analysis, MC production and detector data reprocessing activities. In Figure 3.5 (as taken from [138]) a schema of the SI is displayed, with the CMS Global Pool portrayed as one of their main components. A glideinWMS frontend manages the CMS Global Pool, reaching out to multiple glideinWMS factories to submit pilot jobs to several sites. The system functions in High-Availability, which involves submitting a large number of jobs simultaneously. Then, the HTCondor Negotiator, operating within the Central Manager of the pool, matches payload jobs to pilots. The key components of the CMS Global Pool comprise the glideinWMS frontend and factories, the HTCondor Central Manager, and the Condor Connection Broker (CCB). These components run on 24-core, 48GB (RAM) virtual machines (VMs) hosted on hypervisors with 10 Gbps ethernet connectivity. Additionally, there are approximately 30 job submission nodes, referred to as schedds, connected to the pool.

The CMS Global Pool is a unified HTCondor pool that encompasses all Grid computing resources dedicated to CMS, including substantial Cloud and opportunistic resources. It serves as the primary computing resource provisioning system for all CMS workflows, encompassing analysis, Monte Carlo production, and detector data reprocessing activities.

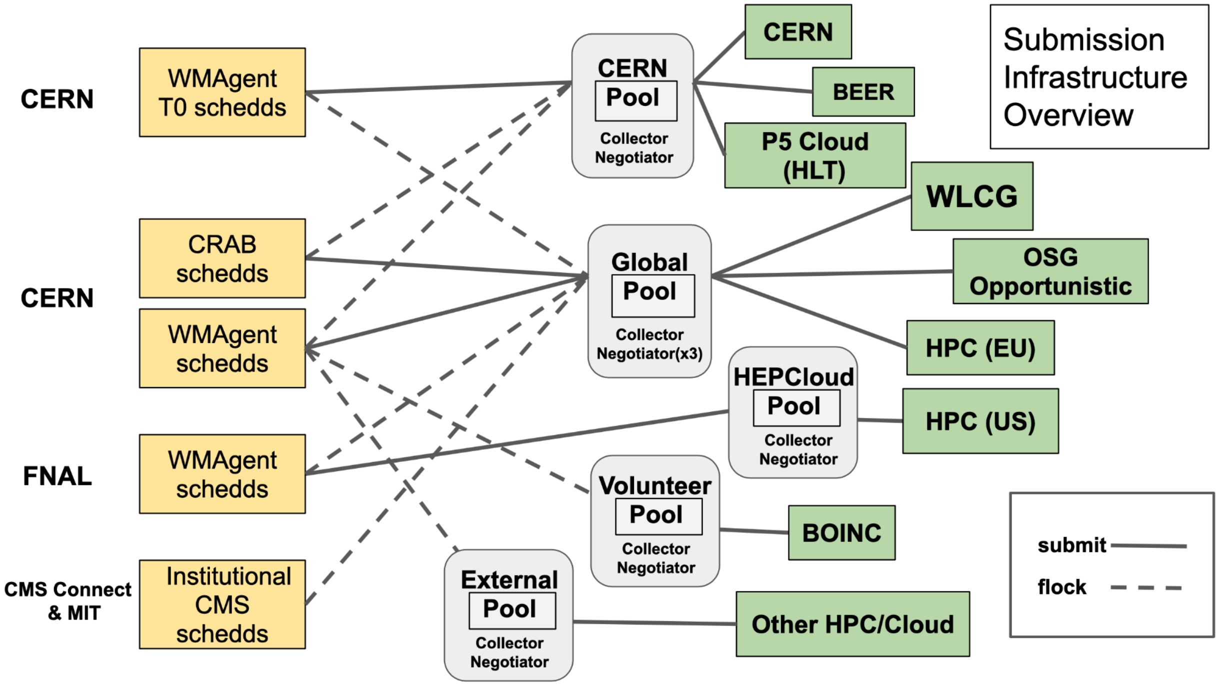


Figure 3.5: Schematic overview of the CMS Submission Infrastructure.

In addition, HTCondor is one of the most used batching systems at CMS sites, managing and distributing computational jobs across the distributed computing infrastructure. It optimally allocates tasks based on job characteristics, ensuring efficient utilization of Tier sites' resources. However, the CMS experiment allows remote reads using the XRootD redirector infrastructure. Streaming data from distant SEs negatively impacts the performance of the analysis tasks, hence data caches are seen as a very promising service to improve the CPU usage at the sites. The efficient operation of the CMS computing model relies on the symbiotic performance of SI, WMCore, WMAgent, and HTCondor. These components work together to convert high-level user tasks into actionable operations executed on Tier sites' resources following optimal and efficient distributions across the available CEs.

Figure 3.6 shows the total number of CPU cores handled by the SI, which include the CMS Global Pool, in a two week period of 2023, at the Tier-0, Tier-1, Tier-2 and Tier-3 computing sites, illustrating the volume of executed tasks running in the CMS computing infrastructure. A peak at about 500k CPU-cores can be seen at the beginning of October 2023.

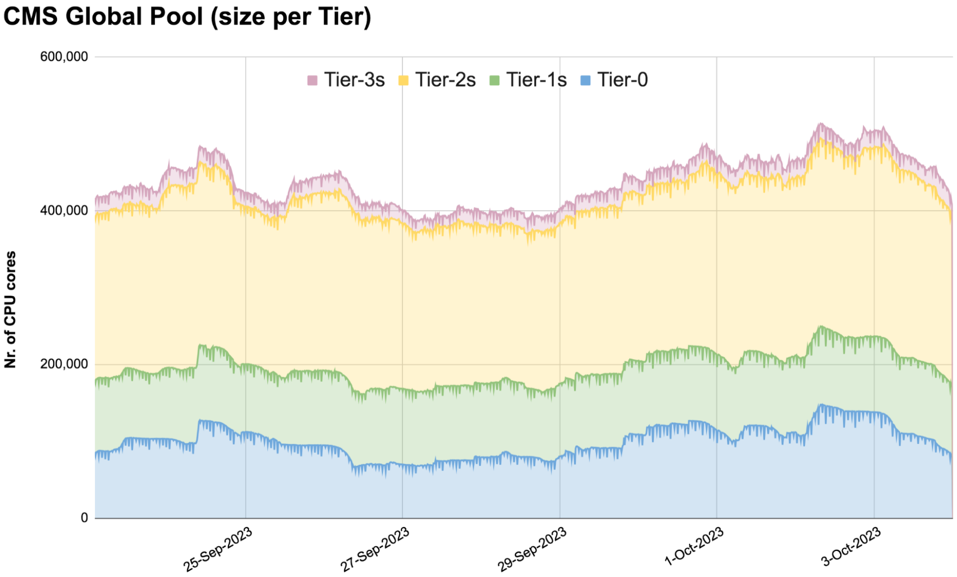


Figure 3.6: Total CPU cores handled for CMS through the SI during the period from September to October of 2023.

3.5. The CMSSW software framework and the Event Data Model

The CMSSW software framework [139] is structured around three components: a Framework, an Event Data Model (EDM), and Services. These components work together to support the creation and use of reconstruction and analysis software. Services play a crucial role by providing essential functions like accessing data and handling input/output (I/O). The EDM is centered on the concept of an "Event", which acts as a container for all the RAW and reconstructed data linked to a specific collision. As the data travels through various stages, it's moved from one module to another within the Event. This approach ensures that data are consistently and effectively managed throughout the entire process. Within the CMSSW event processing model there is an extended program involved, *cmsRun*, and many plug-in modules. These modules contain code that carries out tasks such as calibration, reconstruction algorithms, and analysis tools. What's noteworthy is that the same *cmsRun* program can be used for both actual detector data and simulated Monte Carlo data.

CMSSW is mostly written in C++ and is designed to be highly modular, allowing for efficient development and maintenance of the software. The CMSSW project on GitHub has gained significant popularity, with 4.1k forks and 982 stars [139]. The project has 1,098 contributors and has released 1,161 versions, with the latest release being CMSSW_13_2_5_patch33.

The *cmsRun* is customized for each task using a configuration file made by the user. This file tells *cmsRun* which modules to use, how to set them up, and the sequence in which they should run. It also defines the data to be processed, the resulting output files, and any other settings needed. The CMSSW EDM stands as a potent tool for crafting and applying reconstruction and analysis software. It offers a consistent and efficient approach to managing data, simplifying the creation of complex processing pipelines. Configurations in CMSSW are composed using Python, offering researchers the ability to validate Python syntax within the configuration and perform straightforward checks using the CMS Python module *FWCore.ParameterSet.Config*. Python's integration in CMSSW provides a flexible and easily adaptable development environment for researchers.

3.5.1. *Events in the collision data collection*

In the CMS experiment (and most of the HEP experiments), an event is the outcome of a single reading from the detector's electronics. It encapsulates the signals generated by particles, tracks, and energy deposits across various bunch crossings. It also encompasses the pile-up data, describing the resulting remaining collisions of a bunch crossing. In addition, there are various algorithms to be executed in order to comprehend all the physical processes that occur within an event.

In terms of software, an Event is a C++ type-safe container called in the EDM 'Event'. It starts a collection of the RAW data from a detector or MC event, that is read from a file (if needed) and stored in memory. In this space, any C++ class can be placed within an Event, and there is no requirement to trace back to a shared base class. As the event data keeps on processing, products are reconstructed into the Event as reconstructed RECO data objects. Consequently, the Event holds all data that was taken during a triggered Physics event as well as all data derived from the taken data. It also contains metadata describing the configuration of the software used for the reconstruction of each contained data object and the conditions and calibration data used for such reconstruction. The Event also includes metadata outlining the configuration of the software employed to reconstruct each embedded data entity, along with the conditions and calibration data enlisted for such reconstruction.

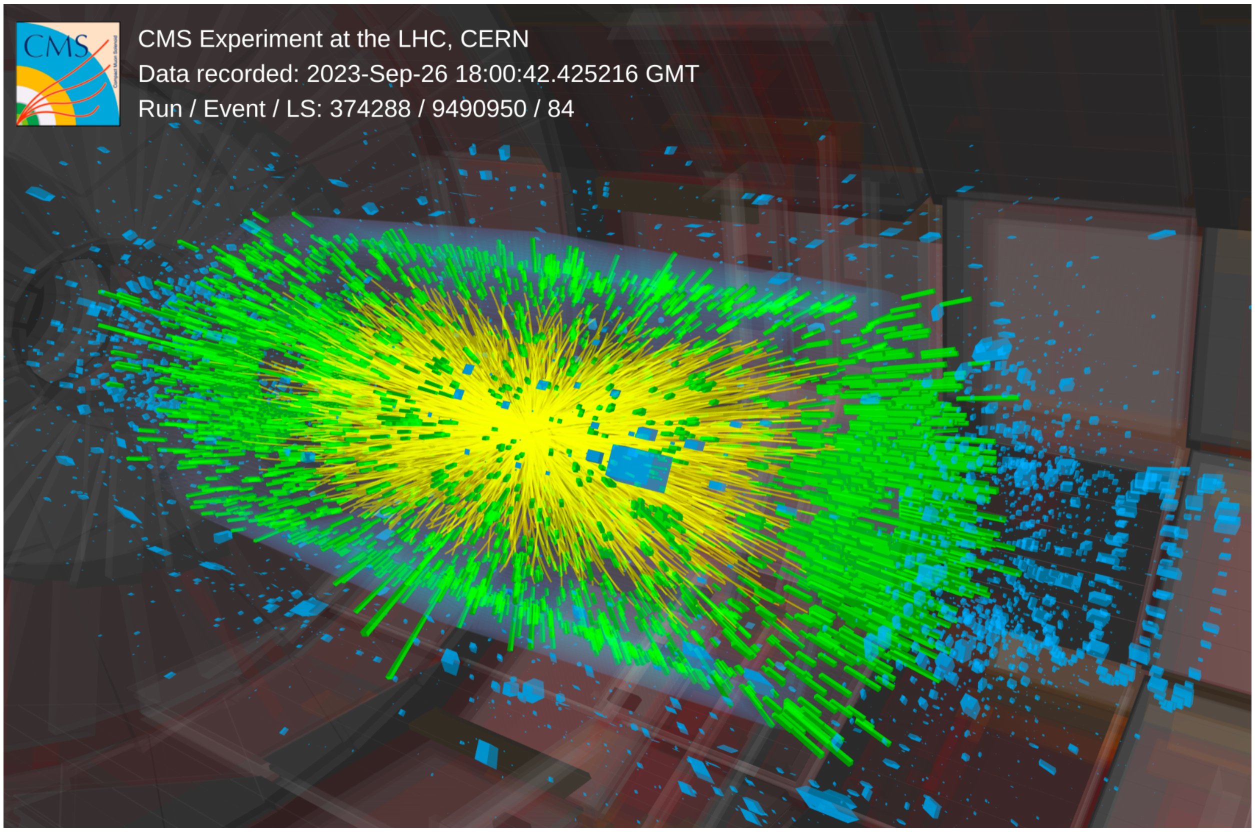

CMS Experiment at the LHC, CERN
Data recorded: 2023-Sep-26 18:00:42.425216 GMT
Run / Event / LS: 374288 / 9490950 / 84

Figure 3.7: Event display of lead-lead ion collisions in CMS detector during Run 3, 26 September 2023.

Figure 3.7 (as taken from [140]) shows an example of the so-called Event display, a graphical representation of the particles appearing in collisions within the detector. The output of Event data is directed towards binary files that can be perused using ROOT [141]. ROOT, developed by CERN, is a dedicated data analysis framework for HEP. It offers an object-oriented programming framework, diverse data structures, a GUI for visualization, and a scripting language, making it adaptable for various data types. Through this framework, this enables the Event to be dissected and used as an n-tuple, a data structure designed to store and organize multi-dimensional data, for final analysis. The CMS output files are written in *.root* binary format. Within an Event, products are organized into separate containers, distinct organizational entities crafted to assemble specific data types independently. These include particle containers (one per particle), hit containers (one per sub-detector), and service containers for aspects like the tracking source. Hence, these events are subsequently processed and reprocessed into the main data tiers included in the CMS computing model. On the other hand, Figure 3.8 (as taken from CMSSW framework TWiki [142]) shows how the event changes with the processing chain of all the data tiers. Furthermore, when it comes to classifying events, the model offers a dual approach, accommodating both abstract physicist notions (like datasets and event collections) and tangible packaging principles to the underlying computing and Grid systems, such as files.

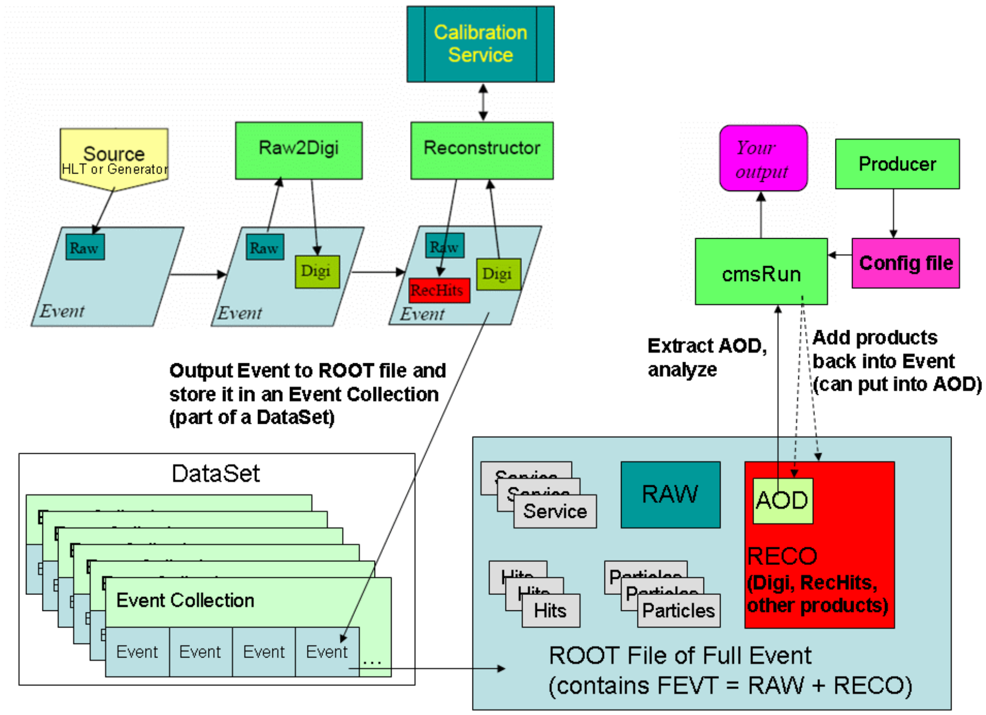


Figure 3.8: CMS event processing workflow across all the data tiers.

On top of that, to allow flexibility and customization in data processing, the events in CMSSW have a modular content. This modularity is characterized by the different data layers using different data formats that can be configured, and a given application can use any layer or layers. The modular architecture of the framework is related to identifying data in the event because each module produces a specific type of data, which is stored in a separate container within the event. This allows for easy identification and access to specific data within the event.

3.6. The CMS Remote Analysis Builder (CRAB)

Among the multiple tools that CMS uses to execute and manage user analysis jobs, the CMS Remote Analysis Builder [143] (CRAB) stands out. CRAB is the official CMS data analysis software that efficiently communicates the user with the Grid environment, being a friendly tool that facilitates its exploitation without having to manage the complexities of the system. To do this, CRAB is installed directly in the user interface (UI), having access to all available versions of CMSSW to be able to carry out the analyzes whose modules or libraries require depending on the type of data being analyzed or the version necessary for them.

In fact, CRAB allows users to transparently access and analyze CMS data stored on any Tier site making use of WLCG's underlying middleware, such as XRootD, without the user having to worry about where and how they are executing the tasks. The integration of CRAB with the middleware that allows remote reading of the data in a transparent and easy way for users is one of the most beneficial features for the user experience for the analysis of the experiment. CRAB also has a stageout plugin, which allows you to bring the data to the user specified endpoint. CRAB has a fallback mechanism that allows you to read remotely from sites that do actually have the desired input data. CRAB's architecture takes a modular software approach with independent components that are implemented as agents communicating via an asynchronous, persistent message service, based in part on GridFTP and a dedicated proxy service delegation. CRAB interacts both with the local user environment, with the CMS Data Management services, and with the Grid middleware. In this way, the client-server implementation is transparent from the user's perspective, so jobs are sent to the Grid in the transaction.

3.7. The CMS experiment variety of workflows

There is a wide variety of workflows and types of jobs intended for certain actions within the use given to the different data and CMS data tiers executed by the different submission tools (such as CRAB, WMCore,...). All of them have unique characteristics based on their use of CPU, I/O and possibilities of remote and/or local access to the data. For example, central data processing workflows and main MC simulations within the CMS experiment primarily occur at Tier-1 centers. These workflows include tasks like re-reconstructing and skimming collision data, as well as reprocessing simulated data. Proton-proton collision simulations mainly occur at Tier-2, where half of the sites are allocated for central MC production, while the remaining half serves user analysis needs. The following subsections briefly describe the most relevant execution tasks monitored by CMS.

CMS jobs can be categorized into three primary types: analysis, production, and processing. Analysis jobs are user-initiated tasks that typically access AOD-derived formats, such as MINIAOD and NANOAOD, for both real and simulated data. These analysis tasks are executed within the CMS analysis operations infrastructure, comprising the WMS and CRAB. Due to the specification of the desired data through its file LFN, CMSSW can open the desired files through XRootD protocol within the job executed by CRAB submission tool, but not always. This is because the TFC uses the selected local site rules to select the proper protocol. On the other hand, production jobs involve Monte Carlo simulations of proton-proton collision events, generating outputs in GEN-SIM-RAW, GEN-SIM-RECO, and AODSIM formats. Processing jobs handle CMS detector RAW data, producing reconstructed data in RECO and AOD formats, and play a crucial role in the data re-reconstruction workflow at Tier-1 sites. And, finally, merge jobs, which are linked to processing jobs, combine multiple smaller datasets into larger, more manageable files to improve the efficiency during later stages of analysis and data processing.

3.8. The CMS Monitoring services and tools

To ensure the efficient performance of the many tools involved in the CMS computing infrastructure, it is necessary to count with a robust and reliable monitoring system. CMS monitors all of its services and resources through a variety of tools and applications, collectively known as the CMS MONIT infrastructure. This infrastructure includes components for handling workloads, data, transfers, user submissions, and centrally managed production requests. The CMS monitoring architecture uses several data sources to provide insights into both current and historical performance. It is built on flexible and open source tools that fit into the specific needs of the experiment. The MONIT system helps to deploy CMS monitoring applications and many other sources of CMS monitoring data use ActiveMQ [144] to send data in JSON format. At CERN, MONIT ingests this data using a Kafka pipeline and sends it to different storage systems, such as Elasticsearch (ES) [145], InfluxDB [146], and Hadoop Distributed File System (HDFS [147]). CMS uses the MONIT system to track over 25 different aspects of its operations. These include things like how jobs are configured in HTCondor with unique JobIDs, how data is transferred between computing centers, how users interact with the CRAB, another CMS tool to submit and manage user jobs within the CMS Grid resources, and WMAgent tools; also, how CMS web services are performing. MONIT can also be used to troubleshoot problems. For example, if a workflow fails on the CMS computing cluster, MONIT can help identify the root cause of the issue. It can also be used to detect problems with specific computing centers or with the availability of data. Additionally, MONIT can be used to track how much data is produced by different workflows. Besides MONIT, CMS offers a range of APIs, such as DAS, Rucio views, Kibana [148], SWAN [149] and more, designed for efficient data discovery and utilization. These APIs equip researchers with improved tools and resources to access and analyze data gathered by the CMS experiment. For instance, Kibana stands out as a potent data visualization and exploration tool, empowering users to visualize, and analyze immediate information. It also features a REST API that provides JSON output, ensuring machine-readability and enabling automation1. Another notable API is DAS, offering a unified view of the CMS data environment, enabling users to search for and identify relevant data sources. For the HTCondor part of the CMS monitoring infrastructure, there is a special tool called a "spider" that collects data from the HTCondor computer pool every 12 minutes. This data is then converted into a format that MONIT understands and sent to MONIT in JSON format via ActiveMQ. The HTCondor data is then stored in OpenSearch [150], InfluxDB, and HDFS. CMS also uses a Spark platform[7] called CMSSpark that can be used for multi-purpose monitoring. This platform is used, for example, to gather and assemble data for views about data popularity (PopDB) used by CMSSW [151].

[7] In fact, part of the analysis computed in this Thesis has extensively exploited this technology.

Finally, CMS has a vast variety of services dedicated to ensure that the integration of resources of WLCG are intensively working. The Service Availability Monitor (SAM [152]) is a system that monitors the availability of these systems and services. SAM collects data from a variety of sources, including system logs, metrics, and alerts. This data is then used to generate reports and alerts that identify potential problems. HammerCloud [153] is a cloud-based service that provides a variety of tools for testing and monitoring the performance and reliability of systems and services. CMS uses HammerCloud to test the performance of its data transfer and storage systems. These two systems work together to ensure that the CMS experiment has a reliable and stable infrastructure.

Chapter 4

Data challenges towards HL-LHC

Back in 2020, the Community White Paper (CWP) was published in order to establish consensus among collaborating stakeholders regarding software objectives and priorities for the HEP community in the near future, specially for the HL-LHC. It also aimed to align efforts, promoting synergies, and identifying key areas for software research and development. These investments were intended to enhance software efficiency, scalability, and overall performance. By harnessing advancements in CPU, storage, and network technologies, the CWP looked forward to achieving the forthcoming challenges that will be faced by the experiments within the LHC's Run 3 and, eventually, the HL-LHC era in 2029.

These challenges cannot be met by simply scaling current solutions in a flat-budget model scenario, assuming that Moore's Law will continue indefinitely. It also acknowledges that I/O rates of modern disk drive systems may not indefinitely increase at similar rates as today, meaning that limitations in affordable storage and I/O rates of higher capacity hard disks also pose a major challenge for HEP computing. Additionally, the evolving landscape of computing hardware introduces novel paradigms, some of which may have positive impact, while others negative. To address these challenges, the HEP community must develop novel computational solutions that excel in efficiency, scalability, and adaptability. Furthermore, investing in advanced computing infrastructure is necessary to meet experiment demands, and cooperating with other HEP experiments and other sciences with similar needs is a must, since much of the underlying computing infrastructure is shared.

Current efforts are underway to establish standards and best practices for software development. The objective of software efficiency, scalability, and performance is following the development of innovative software architectures and algorithms, while existing software is being optimized for enhanced performance. Novel data analysis algorithms and methods are expanding the capabilities of particle identification and Physics information extraction. Also, the distributed computing systems are undergoing optimization in data and workload management, along with new data and job distribution strategies. On the other hand, R&D in computing hardware and architectures are leading to new solutions for processing and analyzing large datasets. In this direction, new hardware architectures are being explored to improve software performance and scalability. And, therefore, the creation of new software tools and frameworks for data processing and analysis is simplifying the development of software and enabling researchers to face complex data challenges more efficiently.

The LHC experiments will not be the only ones benefiting from the scheduled R&D programme covered by the community in the CWP. The document underscores that HEP experiments sharing a parallel timeline and challenges will also benefit from these computing advancements. The CWP highlights the Deep Underground Neutrino Experiment (DUNE [154]) as exemple, or notably some Astrophysics experiments, such as the Square Kilometer Array (SKA [155]), the Cherenkov Telescope Array (CTA) and the Large Synoptic Survey Telescope (LSST [156]).

4.1. Evolution of facilities and distributed computing

The main challenge for HL-LHC computing lies in optimizing the configuration of facilities and computing sites, considering regional funding disparities and the need to accommodate local considerations. The increasing demand for heterogeneous resources, including HPC infrastructures, volunteer computing, and Cloud computing, poses challenges for efficient utilization due to diverse interfaces and dynamics. In addition, the wider range of computing architectures will make the resource management more complex. In this context, the resource environment must be paired with reliable data storage systems and a robust network infrastructure for efficient data delivery. While CPU and disk capacity are projected to increase at steady rates, research network capacity is expected to grow exponentially. This trend suggests a shift towards network-centric computing models that rely on network-based data access, minimizing disk deployments across fewer sites. A notable part of this R&D involves developing a federated data center concept, a distributed network of interconnected data centers, namely CDNs, to achieve efficiency, reliability, and scalability. These federations aim to include efficient caching systems to bring data closer to compute nodes when not found in local storage. Storage system technology is evolving towards object stores, but R&D is needed to understand their role within HEP infrastructures. The challenge of creating an effective worldwide data management infrastructure from diverse and distributed systems still remains, particularly as HL-LHC needs multiple data replicas for redundancy and availability. Transitioning to HL-LHC requires changes and a comprehensive understanding of the costs and benefits associated with the proposed solution. For that purpose, a cost-model that evaluates these changes, accounting for hardware and human costs, software and workload performance, and their subsequent Physics impact, is necessary.

4.2. Evolution of the data management

Regarding the data organization, management and access (DOMA) program, there are several tasks that need to be addressed in order to demonstrate that the increased volume and complexity of data expected over the coming decade can be stored, accessed, and analyzed at an affordable cost. The design of experimental computing models for this era demands a multi-perspective approach. Of particular importance is the increasing availability of high-throughput networks, which could obviate the need to co-locate CPUs and data at the same site. These high-throughput networks could suggest a paradigm shift towards extensive utilization of data access over the WANs. Such a transition could conclude in global and federated data namespaces and improve data caching mechanisms. Adaptations in data presentation and analysis paradigms are equally essential. Integrating event-based data streaming alongside more conventional dataset-based or file-based data access methods will be necessary to optimize the utilization of opportunistic computing resources. This optimization includes HPC facilities, commercial Cloud resources and campus clusters that could help to alleviate the requirements in terms of computing and could benefit from advanced data management policies.

4.2.1. *The data organization, management and access (DOMA)*

Initial LHC computing models were based on simpler paradigms prior to the integration of distributed computing at the core of experiments. Throughout the years, the original LHC computing models have been adapted to address the demands of distributed computing and increasing data volumes. However, these models have always clearly divided data interaction into three distinct aspects: organization, management, and access, summarized under the general concept known as “DOMA”. Regarding each of these aspects, those are found within a context where several protocols, tools and services are required to orchestrate the amounts of data produced and served by the LHC. Data organization refers to how data is structured when it is written. This data is written in ROOT files in a column-wise organization, with the corresponding records to these columns compressed. The management of the data is delegated to the computing infrastructure of WLCG, with each experiment using their own data placement and FTS. The experiments usually use a system of catalogs to move this data between the sites and have control over them within the SEs, which is very dynamic, since data are moved and deleted constantly. In the past, the placement system was mainly static, based on placing the data at certain sites and the jobs sent where the data was. During the last few years, this model has been more flexible. For example, pre-placement of data is based on data popularity, so non-accessed or unused data is avoided from being inefficiently placed. On the other hand, applications now have the availability to interact with catalogs or directly with the WMS.

The experiments within WLCG are allowed to access data for direct reads using XRootD. This is traditionally done by staging-in or caching the data when jobs require it. As a consequence, XRootD has settled as the main protocol to access the data for direct reads. Through the years, a practice that has become more usual in the last years is the remote access to data without stage-in. Again, this feature is provided by XRootD or http protocols. During the last years, the need and symbiotic relationship of these areas involved in a common purpose have been scheduled to be optimized together, instead of separately as in the past.

4.2.2. *DOMA challenges*

The future of the LHC data management will require new storage technologies, data compression algorithms, and dynamic, flexible, and accessible systems to access the data. This is because they will effectively facilitate the integration of new computational paradigms crucial for meeting the requirements of the HL-LHC, such as Clouds and HPCs centers. Integration of these resources often lack the essential features for simplified data access and utilization through the commonly employed protocols of LHC experiments. Additionally, emerging applications like machine learning training and high-rate data queries require a reevaluation of how and where data is provided efficiently to perform these tasks.

The evolving storage landscape is becoming increasingly diverse, mirroring trends in computing resources. Adapting to and efficiently incorporating new storage technologies into existing data delivery models is a significant challenge. This includes the utilization of "tactical storage," which becomes cost-effective as it becomes accessible, such as from a Cloud provider. To meet this challenge, a flexible data management and provisioning system is necessary, capable of harnessing these resources on short notice. The presence of volatile data sources will have widespread impacts on various system aspects, including cataloging, job allocation, monitoring, alerting, accounting, and the applications themselves. R&D efforts are crucial to explore alternative data archiving approaches, considering cost and performance trade-offs.

Currently, tape storage is widely used for data that cannot be economically kept online, but it incurs high latency. It is recommended to investigate separate direct-access archives (e.g., disk or optical) or hierarchical models that combine online direct access with archival space, particularly when access latency correlates with storage density. Therefore, it is necessary to ensure that any changes made do not interfere with data accessibility compared to current computing models. Opening to substantial alterations in data management and analysis methods is crucial, as the present practices would be impulsed and benefited from them by the HL-LHC.

4.3. Data Challenges

Data Challenges are a series of tests that began in 2021 and are conducted every two years using the production services of the WLCG sites and experiments [157]. The aim of these challenges is to assess the readiness of the infrastructure in preparation for HL-LHC and to commission the capability to transfer data at incremental higher rate for special periods of Run-3. The first Data Challenge focused on evaluating two critical aspects for the experiments. Firstly, it centered on the export of RAW data from CERN to the Tier-1 sites. This challenge involved transferring the RAW data produced by the ATLAS and CMS experiments from CERN to the Tier-1 sites. The goal was to export the data in real time, which requires a network capacity of approximately 400 Gbps per experiment. Additional network capacity of around 100 Gbps per experiment would also be needed to account for other data formats. On the other hand, there was the data processing challenge, focused on the reprocessing of data stored at the Tier-1 sites. The main aim was to stage the data from tape and export it to the Tier-2 sites for processing. The estimates were based on the scenario where 100% of the data collected in a year at a specific Tier-1 site is reprocessed in less than three months. The network capacity required for this challenge aligned with the export of RAW data from CERN to the Tier-1s. The Data Challenge involved a group of data experts contributing to a common dashboard solution assessing the serial of tests. The DC21 challenge conducted in the early October of 2021 proved to be very useful to test the infrastructure and identify potential enhancements in the data management. The normal production traffic already reaches the minimal model threshold of 480 Gbps. The injection of data during the DC21 allowed the throughput to exceed this threshold, reaching a maximum throughput of 64.5 GB/s. The amount of traffic produced by the DC21 activity towards a Tier-1 was in testing the Tier-1s ability to handle the ingress and egress movements. The goal was to achieve a traffic rate of at least 480 Gbps (240 Gbps ingress + 240 Gbps egress) and move an estimated 2x 2.6 PB volume in any 24-hour period. Normal production traffic exceeds these thresholds, though not continuously, so 240 Gbps was decided for the DC21 in order to double the target [158]. Hence, the Data Challenge setup tested not only the network but also the limits of the production infrastructure. During DC21 disk space was identified as the most limiting factor. Recommendations were made to separate testing of the network and infrastructure in future challenges and to consider a load generator of data to ensure the uniqueness of data transfers. The target peak achieved of 1 Tb/s among the transfers of all major LHC experiments [159], including the extra data boost, can be observed in Figure 4.1 (as taken from [160]).

Foreseeing a new data challenge in 2024, namely DC24 [161], the WLCG DOMA group is organizing activities to provide visibility for activities proposed by groups, sites, or experiments, while defining targets and timelines. The DC24 is viewed as an opportunity to demonstrate and evaluate new capabilities at scale. Some of the new planned work areas include perfSONAR site network [162] debugging previous to DC24. It also encompasses enabling WLCG Site Network load monitoring and the deployment of traffic marking for data transfers, as well as using JSON tokens to authenticate to storage services. Finally, DC24 will also aim to demonstrate packet pacing at two or more sites, while keeping on performing network load tests to identify potential bottlenecks.

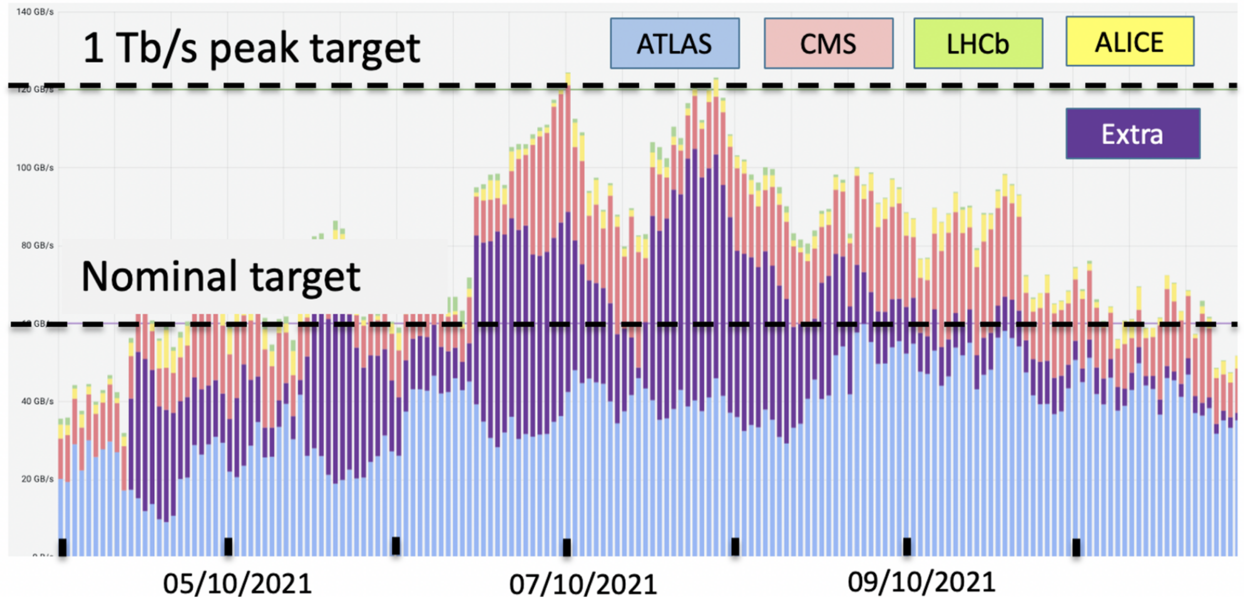


Figure 4.1: Data transfers in WLCG during the DC21 challenge.

4.4. DOMA ongoing R&D and proposals towards HL-LHC

During this chapter it has been explained that the community aims to evolve towards a more cost-effective and equally efficient model compared to the current one employed within WLCG. Some scheduled tasks are aimed to demonstrate the ability to effectively store, access, organize, and analyze the increasingly voluminous and intricate data that will be handled in the coming years. For example, employing exploratory analysis techniques and strategic data organization methodologies resemble those found in the domain of Big Data, extending their utility beyond scientific contexts to various industries. This involves refining file granularity, potentially down to the event-level or even finer and strategically determining the optimal data placement to maximize computing power utilization and efficiency. Furthermore, the smart placement of data is crucial in order to efficiently use the CPU resources to avoid excessive use of the network on remote reads and improve latency. In this sense, caching systems could help to get this data close to compute nodes. In order to achieve this objective, data studies over the storage systems in the sites must be carried out in order to identify which can be the benefits experienced by the different kinds of workflows. Beyond that, the DOMA research is aimed to explore alternative approaches to the current data delivery systems such as CDNs and

Named Data Networking (NDN). This research is essential to establish its feasibility for WLCG as a whole or for smaller WLCG regions, an exhaustive research about the implications of consolidating the data in less and larger locations, the ‘Data-Lake’ approach, is necessary to be carried out.

Among the several proposals to achieve the goals posed by the future data challenges in the DOMA area of the LHC experiments towards HL-LHC, this Thesis aims to investigate those related with the last two proposals exposed in this section. Given the significance of factors like data granularity in the context of flexible integration with big data techniques, delving deeper into the optimal data placement and alternative data delivery solutions proves to be an interesting field for exploration. Data-Lake models based on CDNs are defined through concepts such as the integration of computing and storage resources between small regions that have good connectivity. In the scope of this work, we have the opportunity to leverage the resources of both PIC Tier-1 and CIEMAT Tier-2, two highly reputable centers within the WLCG. These centers are well-suited for conducting testbeds related to the research components of this Thesis.

Chapter 5

CDN utilization in WLCG

Nowadays, the majority of users engaging with bandwidth-intensive multimedia content are more inclined to stream it through dedicated platforms like Netflix [163] or Spotify [164] or access it via Software as a Service (SaaS) portals, rather than relying on locally stored files or peer-to-peer (P2P [165]) downloads (only local and temporary caches are typically used). This shift in user-content interaction is primarily driven by increased demand, the growing complexity of data and, most significantly, advancements in the data-handling capabilities of home networks. In previous years, this scenario was considerably more challenging to achieve reliably due to technological limitations. Consequently, the efficient and widespread distribution of data has become feasible and profitable, prompting the rise of CDNs in various industries since the late 2000s, largely thanks to the inclusion of platforms like those mentioned above. Other leading players in the digital landscape, like Facebook [166] in social media, and Amazon [167] in e-commerce, heavily rely on CDNs to ensure larger content delivery to their end users. The CDNs offer suitable and cost-efficient solutions for businesses that attract large web traffic. The CDNs also play a central role in the Internet, with more than 74% of the Alexa Top 1K websites being served by CDN providers [168].

Historically, CDNs were meant to improve the QoS for IP video services. However, in response to the exponential increases in traffic of videos, CDN operators are now transitioning their infrastructure to the Cloud in order to benefit from its flexibility. IP video streaming was 82% of the total IP traffic worldwide in 2022 [169]. Large scale streaming deployments heavily rely on exhaustive analysis of smart caching, the consumption of energy, network traffic and resource availability and flexibility in order to improve its performance and reduce their costs. When considering WLCG as a distributed global computing and data storage infrastructure, providing data to users of experiments around the world, it is inevitable to see the similarities between the use cases of users of these platforms and those users who analyze the LHC data. The data production, distribution, and traffic growth in the HL-LHC era pose similar challenges to those faced by the industry in managing their large amount of streamed data. The industry has addressed these challenges by deploying CDNs models, and it is likely that similar solutions could effectively fit for the HL-LHC era, since the underlying network in WLCG is very reliable and scalable.

5.1. The early inception of CDNs

The earliest recorded mention of CDNs dates back to a 1995 paper authored by Paul Mockapetris [170], also known as the creator of Domain Name System (DNS) [171]. In this primordial work, Mockapetris introduced a concept called "distributed caching", designed to expedite content delivery to end-users in order to achieve a more efficient and better solution than reproducing content after downloading it. The first commercial CDN to be physically materialized was Akamai [172], established in 1998. Akamai's CDN ingeniously placed and managed a global network of servers to deliver content to end-users based on their geographical proximity using caches. This innovative approach achieved substantial reductions in content download times. Since then, CDNs have evolved into a fundamental component of the internet's infrastructure, and CDNs have widespread utility across a variety of websites and applications, guaranteeing fast and reliable content delivery to end-users. Some of the benefits provided by deploying CDNs are:

- **Reduction of latency:** CDNs can reduce the latency of content delivery by caching data closer to end-users. This is especially beneficial for users in remote locations or with slow internet connections.
- **Higher performance:** CDNs improve the performance of websites and applications by reducing the load on the origin servers by network traffic reduction. This leads to faster content loads and improved QoS and better user experience.
- **Scalability:** CDNs can scale to meet the needs of high-traffic websites and applications. This is beneficial for websites that suddenly experience peaks in traffic.
- **Cost-effective solution:** CDNs reduce the costs of content delivery by caching content closer to the end-users (even in their devices). This helps to reduce the amount of bandwidth that needs to be purchased from internet service providers.

In order to achieve these benefits, several CDNs approaches have been explored in the industry and academia, most of them based on the fundamental principles of smart data placement using caching mechanisms.

5.2. Various CDN approaches

When a web browser initiates a request for a specific resource, its first task is to ascertain the IP address of the server responsible for hosting that resource. This critical step is accomplished through a DNS request, similar to searching for a phone number in a directory: the browser furnishes the domain name of the desired resource and anticipates receiving the corresponding IP address in response. In the case of smaller websites, a single IP address may suffice to cover the entire domain. However, for larger websites, a more intricate approach is employed, involving multiple servers and multiple IP addresses, each linked to a distinct segment of the website. This strategy is adopted to optimize performance, permitting the browser to establish a connection with the server situated in closest proximity. The geographical distance between the browser and the server holds considerable influence over performance. For instance, if a user in the USA seeks access to a website hosted in China, the request will invariably experience longer delays compared to a scenario where the website is hosted within the United States.

To enhance performance and streamline operational costs, major corporations frequently deploy servers housing copies of their data in strategically selected global locations. This network, known as a CDN, boasts servers, often referred to as edge servers [173] that stand at the frontier closest to the end-user within the company's network infrastructure. In the classic definition, a CDN is then seen as a geographically distributed network of servers that work together to provide fast and efficient delivery of content to users. These CDNs are designed to reduce latency, improve load times, and handle high traffic loads by caching and serving content from servers that are closer to the end-users. Here are presented the usual approaches to content delivery and network infrastructure in the digital industry and literature:

- **Pull-based CDN:** the client requests and fetches the content from the CDN. The CDN retrieves the content from the origin server and delivers it to the client. The content is pulled from the origin server as needed, reducing the load on this server and improving delivery to end-users.
- **Push-based CDN:** proactive delivering content to the client cache without being asked. Done by monitoring the client's traffic patterns and predicting which content is likely to be requested. Typically, the content is pushed to edge servers in advance of user requests, reducing latency and improving content delivery to end-users.
- **Hybrid CDN:** it combines the features of pull-based and push-based CDNs, by combining or using one of the approaches when needed. This further optimizes content delivery and reduces latency.
- **Edge computing CDN:** a type of CDN that uses edge computing devices that are located closer to the end-user than traditional CDN servers. This can improve

performance by reducing the distance that the content needs to travel.

- **Distributed CDN:** a type of CDN that uses a network of servers to deliver content. This can improve performance and reliability by distributing the load across multiple servers.
- **In-house CDN:** a type of CDN that is owned and operated by the organization that uses it. This can give the organization more control over the CDN and its performance.
- **Peer-to-peer (P2P) CDN:** content is distributed among end-users, reducing the load on the origin server and improving content delivery to end-users.
- **Cloud-based CDN:** content is delivered from Cloud-based servers, reducing the load on the origin server and improving content delivery to end-users.

The best CDN model for a particular organization will depend on its specific needs and requirements. For example, a push-based CDN may be better suited for delivering streaming media than a pull-based CDN. On the other hand, larger organizations may need a more scalable CDN than smaller organizations. In terms of costs, the eligible CDNs must fit on the model and the features that are included. This election has consequences over the flexibility or the level of control some CDNs have.

5.3. The Data-Lake model

The WLCG Data-Lake model's architecture was introduced at the Computing for High Energy Physics (CHEP) Conference in 2018 [174] in Sofia (Bulgaria) as a proposal to evolve the underlying data infrastructure towards HL-LHC [51]. This Data-Lake constitutes a storage service that spans multiple geographically dispersed data centers, all interconnected by a high-speed, low-latency network. Its core purpose is to function as a content delivery and caching system, effectively serving data to processing centers via a distributed storage infrastructure. This requires modifications in the way the data is orchestrated and accessed. In particular, some of these concepts have been tested under the umbrella of the European funded ESCAPE project [175] (2019-2023). Many sciences with high (current or future) storage requirements participated in this project, recognising synergies between HEP, Nuclear Physics, Astronomy, Cosmology, and others, since many of these sciences use the same data centers as WLCG, and are funded by common funding agencies. The experiments involved are large volume generators (up to multi-Exabyte scale) and the project aimed to set the basis at the infrastructure, network and associated tools, to provide a large-scale distributed storage with efficient big data management, adopting and enhancing both FAIR and Open-science principles. As a result, an ESCAPE Collaboration has emerged as a long term sustainability mechanism for the cluster to be developed to fit the future needs from these big science data generators. In fact, the ESCAPE project envisioned a future based on Data-Lake models.

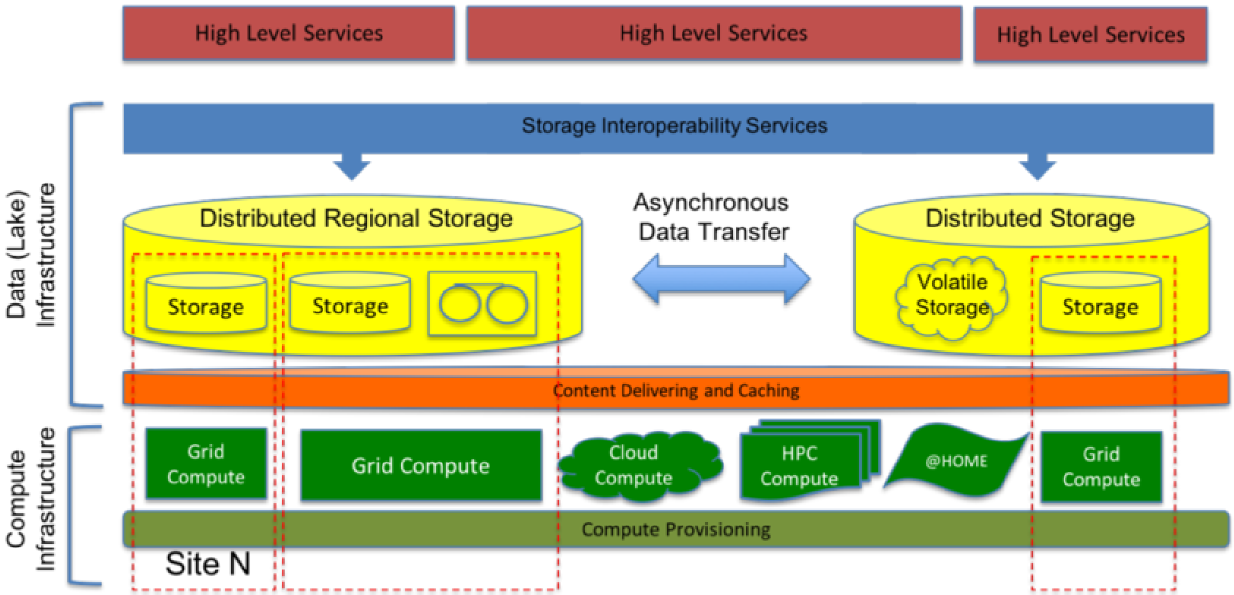


Figure 5.1: Diagram of the 2018 prototype of WLCG Data-Lake model based on 'eulake' nodes.

The WLCG Data-Lake inception looked forward to moving on towards a consolidated storage infrastructure, with networks in the Terabit order, and heterogeneous processing capacity, which may not be co-located with storage counterparts. A detailed diagram for the elements within the planned architecture can be seen in Figure 5.1 (as taken from [51]), which outlines heterogeneous compute infrastructure connected to a data storage system with varying QoS levels. Data transfer relies on efficient WAN protocols and is managed through storage interoperability and high-level services. To overcome latency challenges resulting from non-co-located data and computing resources, the system employs content delivery and caching techniques. In order to be fully operative, the Data-Lake has the ability to interconnect different SEs utilizing heterogeneous storage technologies, underscoring its versatility and compatibility within the broader storage ecosystem. This architecture proposal aims to provide flexibility to scale and reduce hardware and operational costs, by optimizing policy-driven decisions on data placement and retention.

Preliminary tests were carried out by submitting two different workflows from CERN compute nodes while reading data locally or distant sites, participating in the WLCG Data-Lake testbed. This model showed that the remote access scenario (data not at CERN) through the WLCG Data-Lake was surprisingly similar to the performance of a local access scenario (data at CERN). Many aspects are expected to produce a degradation on the CPU efficiency on the LHC experiment's jobs while reading data, such as distance, latency, WAN status, routing and the dominance of read overheads. Because of the limited setup, first WLCG Data-Lake tests did not make clear whether the impact of WAN was significant in this case, and what the impact of bandwidth was. However, successful use cases based on CDN solutions, ranging from caching proxies to analysis facilities, were tested and deployed within the US and Italian

LHC community. Some of these experiences proved that a deployment similar to that proposed by the WLCG Data-Lake would allow for less centralized management of local resources. Moreover, due to the distributed nature of the WLCG infrastructure, local improvements in regions with nearby centers and low-latency connections would contribute to the overall enhancement of the Grid. A Data-Lake model or similar based on the underlying WLCG distributed infrastructure including its own dedicated services, protocols, etc..., should rely on a CDN delivering the LHC data through strategically located data caches. This solution avoids the need to co-locate computing nodes with the storage systems handling the data, delving into a more cost-effective than the current one deployed. Caches can improve the efficiency and scalability of data access by reducing the latency of data retrieval, reducing as well the load on the network and the WLCG SEs. Also, caches can store frequently accessed data closer to the users or applications, reducing the need to retrieve the data from remote storage systems in a repetitive manner. However, the deployment of caches also introduces new challenges, such as cache consistency, cache validation, and cache replacement policies. These challenges need to be addressed through the development of new cache management techniques and algorithms that can optimize the use of caches across the distributed computing infrastructure.

5.4. Data cache services in the LHC experiments

WLCG has extensive experience in utilizing caches. The successful examples include the deployment of software through CVMFS of the utilization of Frontier and Squid caches [176] for accessing detectors' conditions data. LHC reduced data and reduced outcomes of simulations are typically accessed by many scientists worldwide (sometimes, data is re-accessed very often to improve the outcome of an analysis, or resubmission of previously failed and debugged analysis tasks). Since these data products are very popular and accessed frequently, in particular ahead of big scientific conferences, several R&D projects have emerged in the LHC collaborations to enable data caches for these scientific data products. This situation poses new challenges, since the amount of managed data in these data caches would be considerable as compared to CVMFS for software deployment or access to the WLCG Frontier and Squid servers for detectors' conditions. On the other hand, the controlled data processing and massive simulation campaigns tend to be well organized by the experiments, and the input data needed in these campaigns tend to be less re-accessed than the end-user data. This fact will be shown in detail in Chapter 7, where we have addressed the CMS data popularity based on the type of data that the CMS experiment produces.

Caching frequently accessed data close to end-users improves the user experience as it reduces the latency of accessing input data and improves analysis tasks performances (i.e. tasks end faster, which is well appreciated by the end-users). This is particularly relevant, since many users access the same datasets when performing a particular analysis, and some of these datasets can become really popular in terms of access before big conferences or close to a paper submission deadline. Typically, the CMS jobs are executed in sites that contain the input files in their SE but have the capability to read them remotely by using the CMS XRootD's Data Federation [177]. Data caches offer exceptional scalability, easily handling substantial data and high traffic loads, rendering them versatile for diverse use cases. Their compatibility with existing systems and protocols makes integration feasible and they outperform those in load balancing by distributing the workload across multiple servers, ultimately improving system performance. Additionally, data caches can be deployed on outdated hardware (old disk servers), as cached data is usually not cataloged or considered critical by the experiment. Hence, data caches might not need to adhere to pledges, and they would act as an opportunistic resource that would otherwise have no direct utility for the experiments. This also would impact on the costs of storage resources deployed elsewhere.

For effective integration of CDNs in WLCG, smart data placement based on caching is important to deliver content close to the compute nodes in which the end-users execute their analysis tasks. XCache has been developed and commissioned in WLCG to act as data caches, since it is a cache system designed to optimize data access based on the XRootD protocol. For CMS, the XCache can subscribe to the hierarchical structure of CMS XRootD redirectors, and use this infrastructure to populate the data cache. The use of XCache has expanded to other VOs, such as VIRGO [178] (gravitational waves), which use a XCache-based cache hierarchy based on HTTP protocol (the so-called StashCache [179]). XCache offers two deployment options: physical XCache or XCache proxy. Physical XCache involves caching data directly on each storage server (caching part of the data in memory), leading to improved performance and reduced network usage while utilizing storage space. This deployment is particularly suitable for small-scale implementations due to its ability to enhance direct data access. It can serve compute nodes without requiring a persistent storage system if the cache is adequately sized and configured to manage the input data loads. On the other hand, XCache proxy functions as an intermediary server positioned between XRootD clients and storage servers. It intercepts data requests, checks its cache for the requested data, and serves cached data directly to the client. If the data is not cached, the proxy server retrieves it from the storage server, caching it for subsequent requests. XCache proxy centralizes caching on a proxy server, promoting scalability while introducing a minor performance overhead. The XCache service architecture will be explained in detail further.

While other alternative data cache solutions have been investigated within the WLCG, such as expanding the utilization of CVMFS or Squid proxies, deploying Varnish [180] cache with reverse proxies, or implementing Apache Traffic [181] servers with HTTPS, none of these options has demonstrated superior effectiveness compared to XCache. Several successful experiences have been demonstrated so far with the use of XCache, and in particular by means of the results that are presented in this Thesis. The USA-ATLAS collaborators have tested a computing environment based on distributed XRootD caches [182] cataloged and managed by Rucio. The cache hit-rates increased and reduced the volume of requests to the ATLAS data management system, and in particular these caches have proven to be essential to bring input data files to HPC centers (that typically do not offer large storage capabilities), and even reducing the network input traffic to large HPC facilities, such as the National Energy Research Scientific Computing Center (NERSC [183]). The USA-CMS collaborators deployed a rather big XCache with 20 disk servers adding up ∼1.2 PB total capacity, deployed on Kubernetes, to serve data to both University of California San Diego (UCSD [184]) and Caltech [185] Tier-2 sites [186]. The network was dimensioned to facilitate a good connection with the sites (10 Gbps for UCSD and 40 Gbps for Caltech). This system pre-placed popular datasets that are used for analysis, and pre-placement decisions are taken in a timely manner. The authors reported that the data analytics that drive these decisions saved up to six petabytes of disk space that would otherwise have had to be purchased.

5.5. The XRootD's cache: XCache

XCache, based on XRootD, is an efficient caching mechanism used in scientific computing environments. Operating within the high-performance and scalable XRootD storage system, XCache specializes in caching static scientific data files. Its versatile configuration and plugin options allow for customization, enabling its deployment either as a single instance or as a cluster. The functionality of XCache revolves around asynchronous fetching and caching of file segments or entire files. The XRootD's cache was designed to optimize data access, enhance efficiency and minimize network load in distributed computing setups. In the beginning, back in the 2019, XCache was firstly used with Rucio to improve cache hit rates and replace Squid for better distribution of large files in CVMFS for ATLAS. Furthermore, in the last few years, it has also been adapted for extensive delivery of scientific data, including HPC environments and Data-Lake models.

5.5.1. *The XRootD's protocol*

XRootD is a high-performance, scalable, and fault-tolerant software framework for accessing and managing data repositories of various types, developed at SLAC [187]. It is typically used to provide fast, low-latency access to file-based data organized in a hierarchical file system-like namespace. The architecture of XRootD comprises a server-side and a client-side framework. The server-side framework oversees data storage and access, while the client-side framework offers a user interface to interact with the server-stored data. An outstanding feature of the client-side framework is the declarative API, which aligns with modern C++ paradigms and provides an asynchronous interface. XRootD has a modular architecture based on plugins, which extend and improve the functionality of the framework allowing users to customize and adapt the system to their specific needs and use-cases. These plugins can be easily integrated into the XRootD framework to provide added features, improve performance, and enable compatibility with various systems and technologies. By choosing and configuring the appropriate plugins, developers and administrators can adapt XRootD to a wide range of use cases and environments. For example, one of these plugins is XCache.

XRootD is also considered a fault-tolerance protocol due its redirectors system. Redirectors are a key component of the XRootD framework that manages data location services. It provides location-transparent data access and simplifies finding and accessing data in the distributed XRootD storage system. When a client requests a file, the redirector determines the appropriate data server that holds the file and redirects the client to that server. This allows clients to access data without knowing its specific location. The redirector system shown in Figure 5.2 (as taken from [188]) is customizable and extensible through plugins, which can be developed to meet site-specific requirements and improve overall XRootD performance.

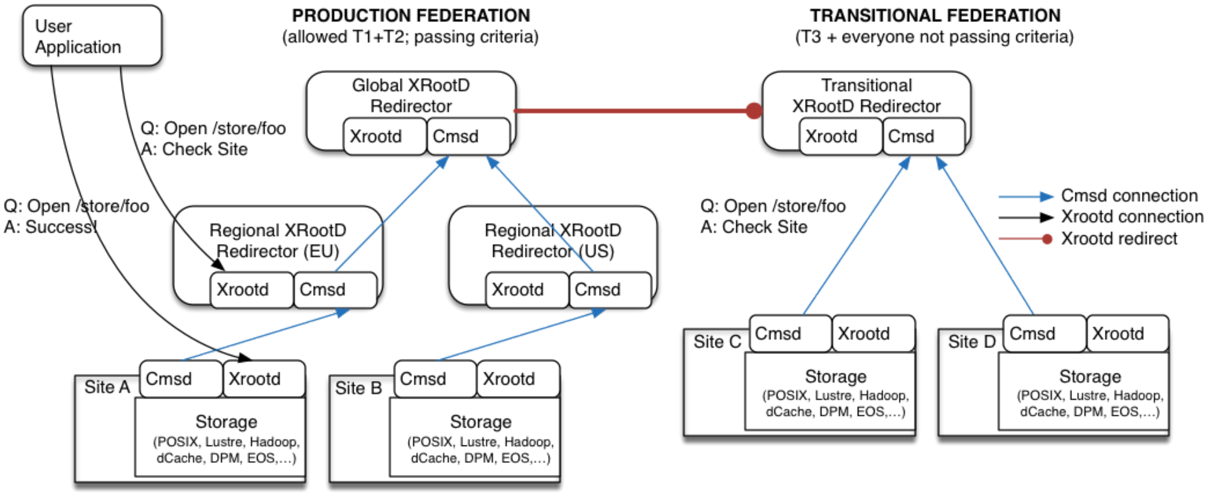
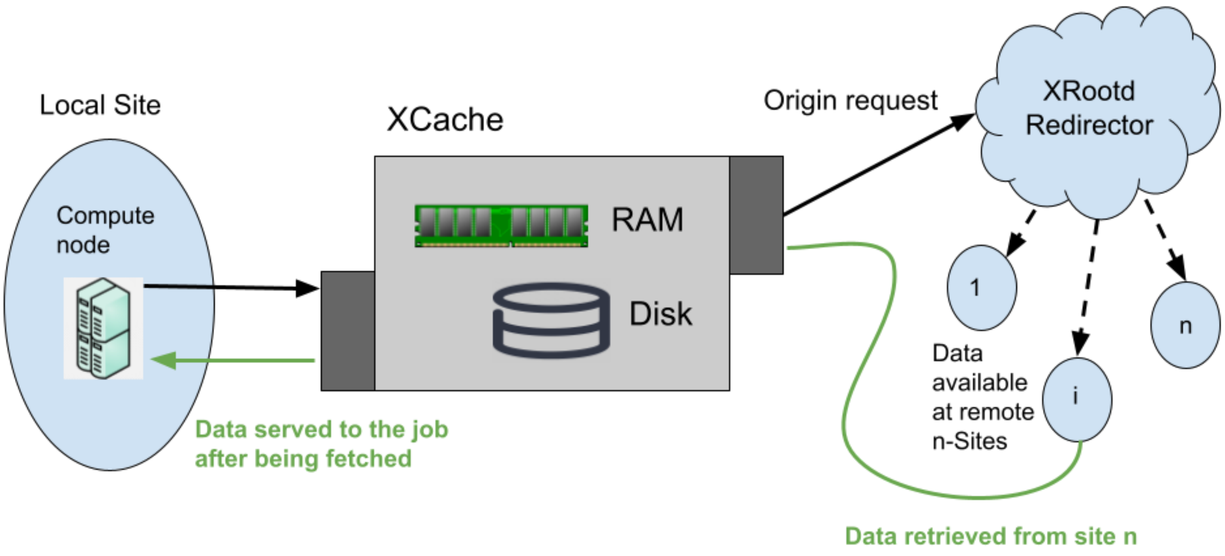


Figure 5.2: Schematic organization of XRootD redirectors across the WLCG.

5.5.2. *The XCache architecture*

The proposed architecture for XRootD's caching proxy is a disk-based solution that optimizes data access, data placement, and data replication. In its main definition it is a Squid-like cache that supports the XRootD protocol and the http protocol, making it suitable for a wide range of applications. With its focus on science data, XCache is capable of efficiently handling both large and small static data files. The XRootD proxy service can be used as a local cache, minimizing WAN traffic depending on the specific workload. The proxy server can be configured to specify an alternate cache server to use in case the caching proxy becomes overloaded. XCache usually is deployed choosing between two configurations. The complete-file auto-prefetching proxy is a configuration where the file retrieval starts as soon as a file-open request is received. This approach is well-suited for scenarios involving entirely random file access or cases where it's predetermined that an entire file will be read. In contrast, the partial-file, block-based on-demand proxy selectively downloads specific fixed-size blocks of a file in response to requests, optimizing data retrieval and transmission.



Local Site
Compute node
XCache
RAM
Disk
Origin request
XRootd Redirector
1
i
n
Data available at remote n-Sites
Data served to the job after being fetched
Data retrieved from site n

Figure 5.3: Diagram of the architecture of XCache service, including the interaction between the local cluster with the tunable caching proxy with the XRootD's federation.

Figure 5.3 presents a high-level diagram outlining the operational flow of the caching proxy for both usual XCache implementations. This diagram explains the sequence of events that occur when attempting to read from local storage proves unsuccessful, prompting the client to engage with the local proxy. The proxy initiates communication with a redirector to locate a file replica on an alternate site. After successfully acquiring the data, the proxy serves it to local clients, concurrently creating a copy on disk. If XCache is configured to retrieve data from an XRootD(s) or HTTP(s) data source, the format for accessing the data can be either "root(s)://Xcache//file" or "http(s)://Xcache/file". Conversely, when XCache is set to fetch data from any source, concatenated URLs come into play. For instance:

- Fixed root(s) data source:
  - `XRootD protocol: "root(s)://Xcache//file"`
  - `HTTP protocol: "http(s)://Xcache/file"`
- Fixed HTTP(s) data source:
  - `XRootD protocol: "root(s)://Xcache//http(s)://cern.ch/eos/file"`
  - `HTTP protocol: "http(s)://Xcache/http(s)://cern.ch/eos/file"`

- Flexible data source:
    - `XRootD protocol: "root(s)://Xcache//root(s)://cern.ch//eos/file"`
    - `HTTP protocol: "http(s)://Xcache/root(s)://cern.ch//eos/file"`

Several strategies can be employed to optimize the performance of XCache. The system offers adjustable RAM buffers, allowing for data caching before eventual storage commitment. Tunable write-queues are available to fine-tune write operations, optimizing storage performance. Flexible policies can be configured to govern cache storage, encompassing aspects like low/high watermarks, cache replacement strategies for data retention, and unconditional purging of less-accessed ("cold") files.

Chapter 6

A CDN strategy for CMS Spanish sites

This chapter delves into the explanation of the data cache deployment strategy proposed for the Spanish region of CMS. The primary objective is to establish a CDN that not only optimizes data management in the region but also alleviates the associated costs. This initiative aims to demonstrate the potential benefits outlined in Chapters 4 and 5. Initially, the chapter starts with an overview of the CMS Tier-1 and Tier-2 sites in Spain, specially delving into the selected sites for the CDN testbed, PIC and CIEMAT. Following this, the practical implementation of the caching system, while offering insights into the CDN's prototyping process will also be explained. This Chapter will conclude by elucidating the criteria underlying the technology choices, also providing a comprehensive understanding of the decisions made in the context of the physical deployment of the physical cache. The experience of deploying this service is part of the Thesis contribution and the initiative has been actively supported by PIC and CIEMAT members.

6.1. The CMS Tier sites in Spain

Spain has participated in the development, deployment, and operation of LHC computing since its first inception. Spain contributes with a Tier-1 center (PIC) and three federated Tier-2 centers for the ATLAS, CMS, and LHCb experiments. The Spanish commitment in the WLCG MoU is to provide 4% of the resources at the Tier-1 and Tier-2 levels. RedIris, the Spanish National Research and Education Network and ICTS [189], plays a central role in providing 100 Gbps connectivity between the Spanish WLCG sites and international connectivity to the rest of the WLCG infrastructure. The CMS resources in Spain are deployed at the Tier-1 center, which is situated at the PIC in Barcelona, and in two Tier-2 sites, deployed at CIEMAT in Madrid and IFCA in Santander. About 75% of CMS Tier-2 resources are deployed in CIEMAT, while the remaining 25% in IFCA. As compared to the global resources deployed for WLCG in Spain, ∼34% of CPU resources, ∼37% of disk resources, and ∼38% of tape resources are assigned to the CMS experiment. The computing centers providing resources for WLCG in Spain are displayed in Figure 6.1. As a key facility within WLCG, the PIC plays a vital role in supporting the computing needs for the LHC experiments in Spain. About 50% of the total resources for WLCG in Spain are deployed in the Tier-1 center. These pledged distributions can be observed in Figure 6.2.

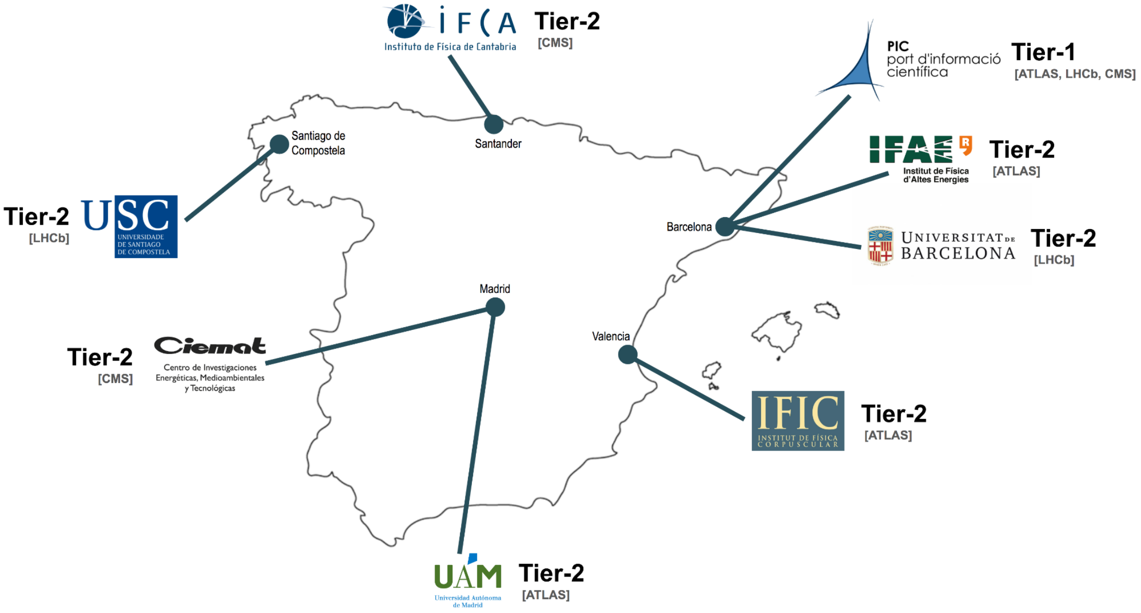

IFCA Instituto de Física de Cantabria
Tier-2
[CMS]
Santander
PIC port d'informació científica
Tier-1
[ATLAS, LHCb, CMS]
IFAE Institut de Física d'Altes Energies
Tier-2
[ATLAS]
Santiago de Compostela
Tier-2
USC Universidade de Santiago de Compostela
[LHCb]
Barcelona
Universitat de Barcelona
Tier-2
[LHCb]
Madrid
Tier-2
[CMS]
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
Valencia
IFIC Institut de Física Corpuscular
Tier-2
[ATLAS]
UAM Universidad Autónoma de Madrid
Tier-2
[ATLAS]

Figure 6.1: Spanish WLCG sites, including their supporting experiment and Tier level.

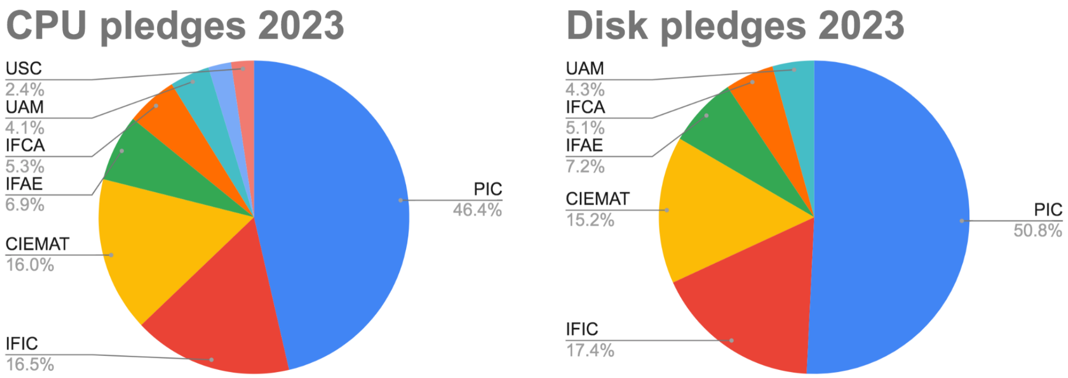


Figure 6.2: Distribution of CPU and Disk pledges in WLCG Spanish sites (all of the Tape is placed at PIC Tier-1 site). Approximately, half of CPU and Disk resources are deployed at the PIC Tier-1.

Figure 6.3 shows the CPU, Disk and Tape resource pledges deployed in PIC Tier-1 since 2009, for the ATLAS, CMS and LHCb experiments. At the Tier-1 the CMS experiment has currently access to ∼2650 CPU cores (the compute power is about 12 HS06 per CPU core as explained in Section 2.8), ∼4 PB of Disk, and ∼12.5 PB of Tape storage. The Figure 6.4 shows the CPU and Disk resources deployed at both CIEMAT and IFCA computing centers. The CMS experiment has access to ∼3400 and ∼1100 CPU cores, and ∼3.5 PB and ∼1.2 PB of Disk storage at CIEMAT and IFCA, respectively.

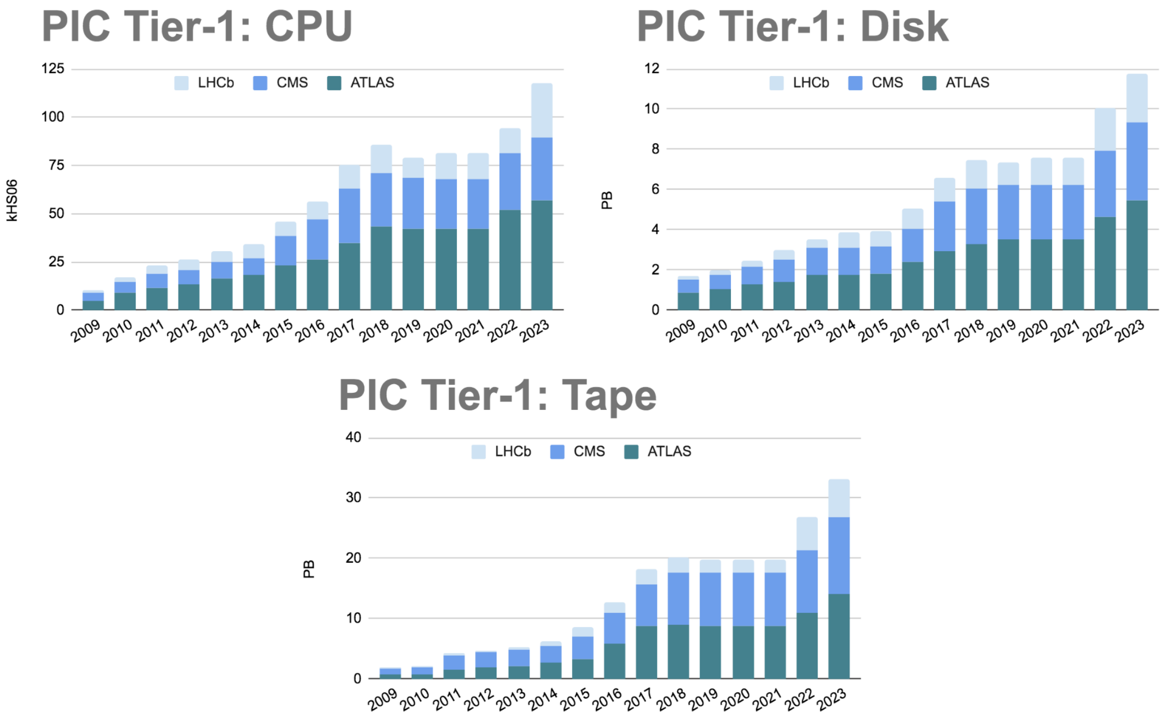


Figure 6.3: Spanish Tier-1 pledges for LHCb, CMS and ATLAS since 2009.

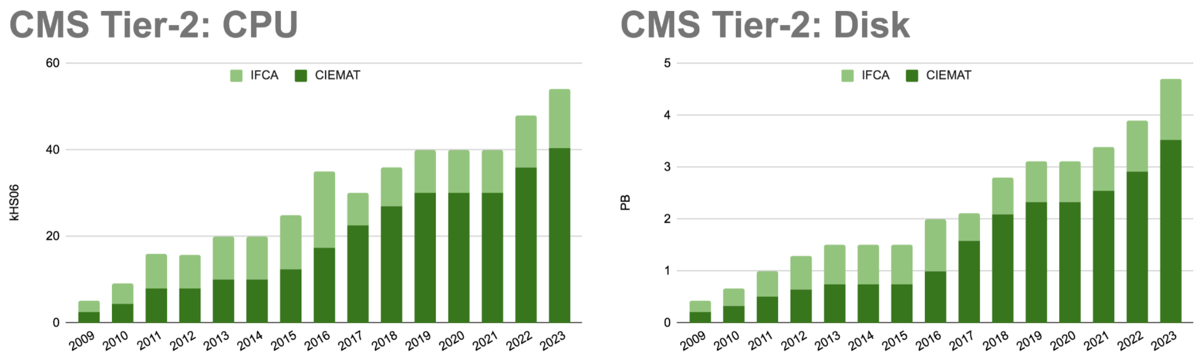


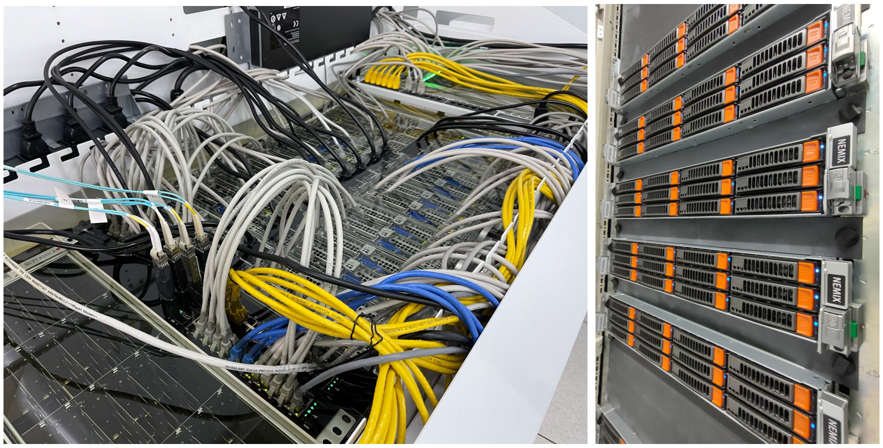
Figure 6.4: Spanish Tier-2 pledges for CMS since 2009.

6.1.1. *The PIC Tier-1*

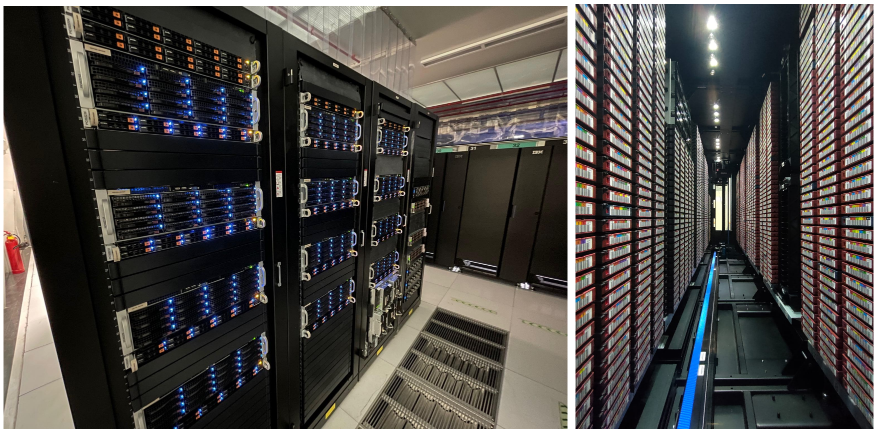
The CIEMAT and IFAE institutions maintain since 2003 a collaboration agreement for the maintenance and operation of the PIC scientific-technological center, located on the Bellaterra campus of International Excellence of the UAB [190]. This collaboration is supported by the Spanish Ministry of Economy and Competitiveness [191] and the Catalan Department of Economy and Knowledge [192].

The team of PIC researchers, engineers, and technicians develop and operate the Spanish Tier-1 center for the WLCG. Conceived originally as a WLCG Tier-1, PIC has transferred the LHC data processing techniques to other fields, particularly Astronomy and Astrophysics. PIC is the primary data center for the MAGIC telescope [193] and the PAU-Survey [194]. It also deploys one of the reference centers for the CTA telescope, and hosts Cosmology simulation repositories for DES [195], PAUSurvey, and the ESA EUCLID mission [196]. In 2019 the Spanish ministry broadened the RES scope to include data services, and PIC was included in the RES as a service provider. In 2022 PIC was promoted to the portfolio of Spanish ICTS, a designation granted by the Spanish government to research infrastructures that are considered to be of strategic importance for the advancement of science and technology in the country. Since then, a new Artificial Intelligence group has been created and a collaboration with ALBA [197] for material sciences has been formalized. Through its participation in the WLCG and other scientific endeavors, PIC plays a vital role in advancing scientific research and enabling groundbreaking discoveries in the realms of Particle Physics, Astrophysics, and Cosmology. The current infrastructure at PIC boasts approximately 153k HS06 of compute power, which translates to around 12,000 CPU cores, deployed in 160 compute nodes managed through HTCondor. With respect to storage resources, PIC deploys a total of 19 PB of disk storage managed by dCache (in ~60 disk servers), and around 64 PB of tape storage managed by Enstore (on LT08 technology). A Ceph Quincy platform of 2 PB (net) was deployed in June 2023, providing CephFS, RBD [198] and S3 storage [199]. PIC also counts with a Haddoop Big Data platform (2.5 PB) supporting Cosmology and Astrophysics projects based on Spark, Hive [200] and Dask [201] frameworks. The developed CosmoHub portal provides access to this cluster, and it has ~500 worldwide users registered that access the cosmological catalogs through this service. Additionally, there are 18 GPU units for Machine Learning and Artificial intelligence, some of them exposed to the Grid and used by Gravitational Waves experiments. Finally, PIC has deployed a JupyterHub [202] instance, with access to GPUs and the possibility of scaling starting a Dask cluster. Access to existing compute resources beyond those owned by WLCG has been explored as a way to increase the available capacity for LHC computing. PIC led the negotiations with the RES for the declaration of LHC computing as a strategic project. As a result of the agreement, preferential access to a fraction of the CPU resources at the BSC has been granted since 2020. Significant investments have been made to integrate and exploit these resources. The lack of outbound network connectivity from the BSC compute nodes, necessary to interface them to the workload management systems of the LHC experiments, required intensive R&D work to circumvent the limitations. New services were developed and deployed at PIC to interface the BSC with WLCG and handle the submission of jobs to BSC and the data flow from BSC to the PIC storage system. Monte Carlo simulation workflows from ATLAS, CMS and LHCb experiments are now routinely executed at BSC. Since 2020, a total of 100M hours have been allocated and consumed in BSC. This corresponds to an average installed capacity of approximately 50 kHS06, representing around 30% of the current Grid resources deployed at PIC for the LHC experiments. This places LHC computing at the top-user list in BSC.

PIC's facility features a machine room with high-density, air-cooled equipment, as well as an ultra-efficient compact module for liquid cooled equipment. This investment was made to improve energy efficiency and reliability, whose cooling costs have been reduced by nearly 50%, making PIC one of Spain's most advanced data center infrastructures for Science. The air-cooled room features a full hot-air containment system with an innovative above-ceiling scheme that improves cooling efficiency. The hot air is cooled through a dynamic system that combines air-side economizer (free-cooling) using air-to-air heat exchange with outside air, and traditional air-to-water heat exchange with chilled water. The air-side economizer can completely cool the installation in the winter and contributes to cooling through the dynamic system up to 75% of the year. The annual average Power Usage Effectiveness (PUE) is estimated at 1.4. PIC's liquid-cooled container occupies a 25 square meter fireproof enclosure in the basement of PIC data center, which was previously used for an energy-efficient air-cooled room. Four Green Revolution Cooling [203] oil immersion tanks with specially designed oil-to-water heat exchangers were installed in the enclosure at the end of 2015. A water loop runs to the building's roof where a redundant, modular dry cooler is located. Each tank can hold 45U of IT equipment and can dissipate up to 45 kW of heat, an unprecedented 1000 Watts per 1U. All new CPU purchases are installed in this facility, with an estimated Power Usage Effectiveness (PUE) of 1.1. PIC's installation occupies a significant portion of UAB's Building D, which also houses UAB's university computer center and the Regional Research and Education Network Point of Presence (PoP). The building's power supply is provided by two 1 MVA transformers from 25 kV, with three 500 kVA motor generators serving as backup power. The generators are directly linked to the cooling system while Uninterruptible Power Supplies (UPS) devices are used to safeguard the IT equipment. PIC's UPS system is a modular device with a maximum power of 550 kVA, utilizing Insulated-Gate Bipolar Transistor (IGBT) technology that ensures efficiency levels ranging from 97% to 99%. This UPS system covers both the air and liquid cooled installations. Picture 6.1 and Picture 6.2 display some of the hardware elements deployed at PIC, such as the oil-immersed WNs, disk servers and the ceph and Tape machines.



Picture 6.1: PIC oil-immersed compute nodes (left) and disk servers (right).



Picture 6.2: PIC Ceph storage (left) and IBM TS4500 tape library (right).

The building has recently undergone renovation, resulting in the installation of two redundant high-capacity chilled water plants. While the IT equipment consumes electricity, the University is committed to reducing its carbon footprint, by using 100% renewable energy and by installing solar panels on the building's rooftop in mid-2023. This will significantly reduce electricity costs by producing in-house renewable energy and aligning with the University's commitment to sustainable energy practices. The installation of solar panels is an important step towards reducing dependence on fossil fuels, and promoting an environmentally friendly approach

PIC's internal network connects CPU and Disk using high-speed fibers and advanced routers. The external connectivity is based on a high-speed 200 Gbps lambda and several 10 Gbps lambdas, managed using VLANs through the two private networks of LHCOPN and LHCONE. PIC is a major contributor of network traffic for the Spanish National Research and Education Network, transferring ~100 PB in/out per year through its WAN.

6.1.2. *The CIEMAT and IFCA CMS Tier-2s*

The CIEMAT computing center offers computing resources to several scientific disciplines (High Energy Physics, Astroparticle Physics, nuclear fusion, renewable energies, environmental sciences, biomedicine, etc.), which require the storage and processing of vast amounts of data, being the CMS Tier-2 the main deployment, since 2003. The center has a surface of 200 m^2 available for equipment and is fed with a total power of 1 MVA backed up with a diesel generator of the same power, and a UPS with a maximum power of 600 KVA. The CIEMAT cluster has a total of 4680 CPU cores managed by HTCondor, with GPUs. The cluster's disk capacity is ~4.5 PB, managed by dCache. For connectivity, the cluster has two links with RedIris: one utilizing LHCOne with a capacity of 20 Gbps, and the other connecting to the internet also with 20 Gbps. RedIris has already installed a PoP at the center to connect to WLCG via 100 Gbps, and the core networking will be soon updated to reach the maximum network capacity allowed. The IFCA cluster was established in 2005 in a specially conditioned room of over 100 m^2 in the Juan Jordá building of the University of Cantabria. Currently, it hosts the Altamira supercomputer [204], a node in the RES, along with several computing clusters integrated into the Grid and EGI federated Cloud e-infrastructure. The IFCA Grid cluster has 108 computing nodes offering a total of ~3300 CPU cores, managed by SLURM [205]. It also provides access to several tens of GPUs. In addition, IFCA has 1.6 PB storage space through IBM's General Parallel File System (GPFS), and 400 TB storage space through Ceph. RedIris provides 100 Gbps connectivity to LHCONE.

6.2. CDN prototype in the Spanish region for CMS experiment

The Spanish CMS sites are interconnected with high-speed and low latency networks provided by RedIris. For example, CIEMAT and PIC are separated by ~600 Km, and the average RTT yields a low value of~9 ms. Leveraging from these low latency measures, a regional prototype of an XRootD redirector was introduced in ~2018, so if any job executed in Spain fails to open the input file, the input files fallback mechanism asks first the regional XRootD redirector to check if the file is present at any of the rest of Spanish CMS sites. If data is not available regionally, an upper level XRootD redirector is contacted to find the input data elsewhere worldwide. A diagram for this XRootD's implementation is shown in Figure 6.5 (as taken from [206]). This has improved the performance of the executed tasks in the region, however many worldwide distributed remote accesses are present for end-user jobs that are executed in Spain, hence the deployment of a regional data cache in Spain might improve even more the CPU performance for these tasks. This data cache would be first contacted by the fallback mechanism, and its cache uncataloged contents would be populated by the CMS XRootD redirector's infrastructure. Leveraging from the efficient collaboration of PIC and CIEMAT CMS members, a data cache has been prototyped and tested for both PIC Tier-1 and CIEMAT Tier-2 sites.

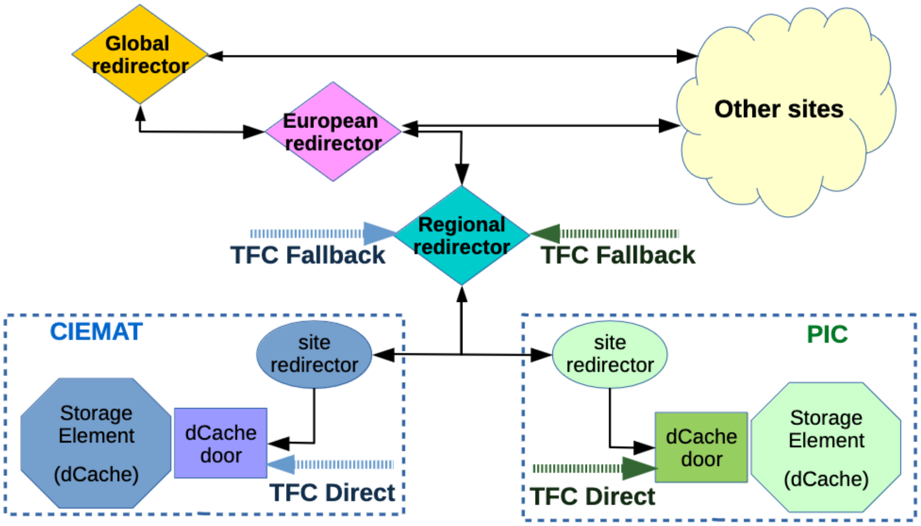


Figure 6.5: Diagram for XRootD's Spanish redirector retrieving requests between PIC and CIEMAT.

Picture 6.3: XCache server deployed at PIC Tier-1 center.

After some initial testing phase with old disk servers, the main XCache has been deployed in PIC Tier-1 with 175 TiB capacity (see Picture 6.3). It consists of a disk server with aggregated 6TB HDDs in RAID6, with 2xCPUs E5-2650L v3 (HT enabled - 48 cores), 128 GB RAM, and a bonding active-active 10 Gbps NIC. The XCache is running one of the latest XRootD versions, namely XRootD 5.5.1. The deployment of this service has been identified as a strategic R&D project and it is funded through the Red Española de Supercomputación (RES-DATA), with the project reference DATA-2020-1-0039. An additional smaller XCache is deployed in CIEMAT Tier-2, aimed to serve data to a new Analysis Facility that is deployed in Madrid. This XCache in CIEMAT is fully based on SSDs, and it has a capacity of 20 TiB. This service has been as well tested through the work developed in this Thesis, but the main studies have used and focused on the XCache deployed in the PIC Tier-1.

Figure 6.6 illustrates how the XCache deployed at both PIC Tier-1 and CIEMAT Tier-2 is embedded with the regional and the rest of CMS XRootD re-directors. During job execution on compute nodes, if the required input data is not found in the local SE, the fallback mechanism queries the cache systems. If data is already in the cache it is served, otherwise it is fetched from a remote site using the CMS XRootD redirector infrastructure, then served to the compute node and stored in the cache. The XCache can be configured to retrieve whole files or blocks of data within the file, with read-ahead. The current configuration fetches data in 10x blocks of 50 kB.

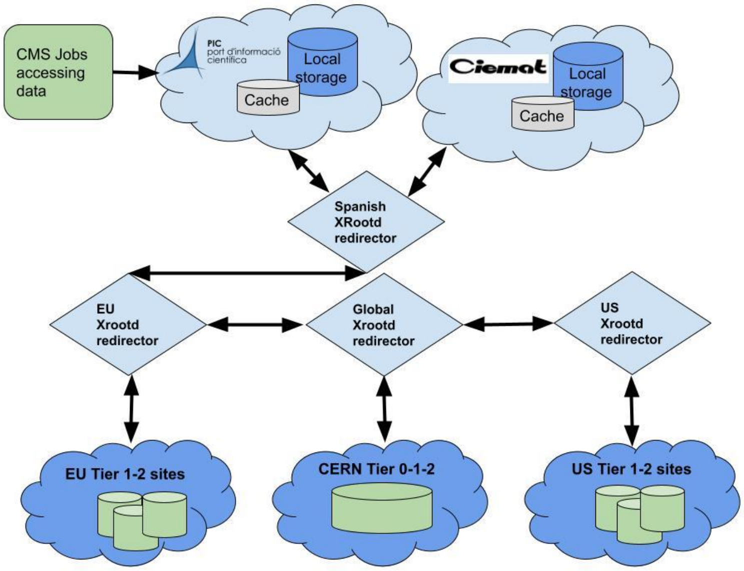


Figure 6.6: XCache configuration for remote input data reads of CMS jobs in PIC Tier-1 and CIEMAT CMS Tier-2.

The XCache service deployed uses the Least Recently Used (LRU) deletion algorithm to get rid of obsolete old unused data, an approach deemed highly effective for XCache services handling CMS data with larger cache sizes (up to 50 TB) [207]. This algorithm sorts the cached files based on usage and date, identifying files that haven't been accessed for an extended period as candidates for deletion. Deletion is triggered by watermarks representing specific occupancy thresholds. When the occupancy exceeds the High-Watermark (HW) of 95%, the algorithm initiates file deletion until reaching the lowest occupancy range, the Low-Watermark (LW) of 90%. Figure 6.7 shows the PIC XCache usage, and how the water marks act to keep the cached data below the maximum allowed usage. Not all of the data is cached. By means of the TFC we can select which data goes through the cache, and which data goes through the global CMS XRootD redirector mechanism. Frequently read data is worth being cached, hence some rules are introduced to achieve this, which are explained in Chapter 7. Additionally, in Figure 6.8 the evolution of hits and misses can be appreciated. As long as popular files keep on populating the cache, the number of hits increases, as the trend in Figure

6.8 clearly shows. Also, this popularity distribution is shown in Figure 6.9, with an early October 2023 daily snapshot. In particular, the XCache at PIC has been configured to serve data to all of the CMS jobs executed in PIC Tier-1, and in half of the computing farm in CIEMAT Tier-2. It serves data to 4500 CPU-cores in use by CMS at both sites. About ~5000 files are accessed daily through the XCache service (serving about 15 TB of data per day, on average).

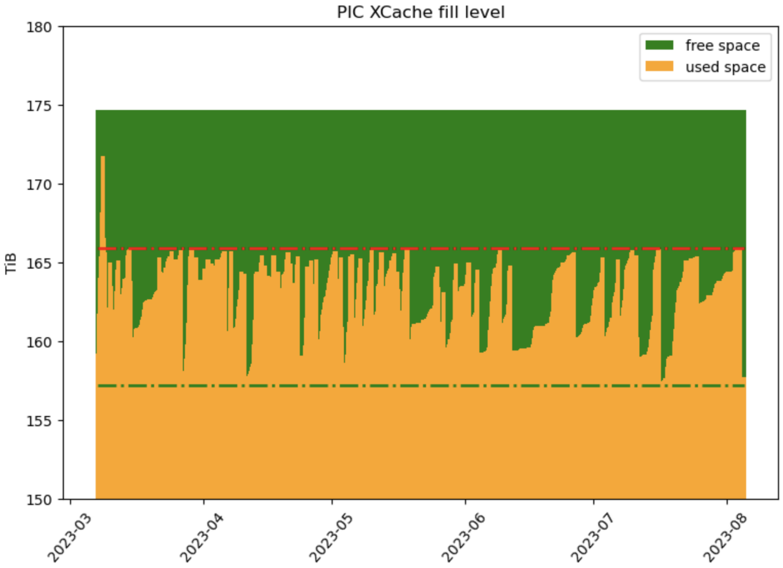


Figure 6.7: PIC XCache fill level over five months. The used and free space is seen, as well as the low and high water marks (green and red dashdot lines, respectively).

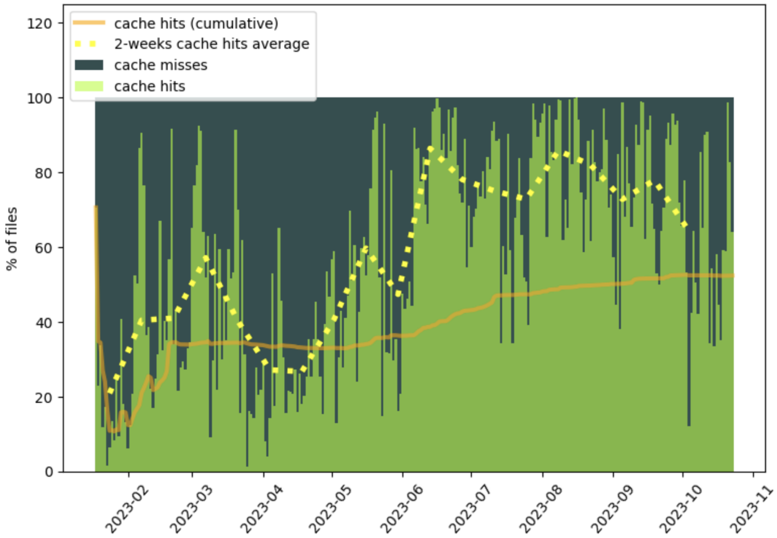


Figure 6.8: PIC XCache hits and misses over five months. The number of hits increases with popular files remaining within the cache.

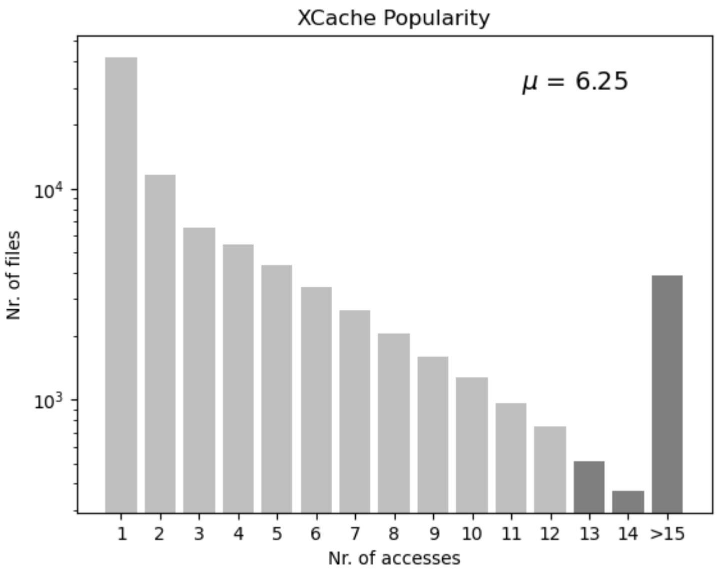


Figure 6.9: PIC XCache's popularity distribution. This distribution has been computed from a daily snapshot taken in early October of 2023.

The PIC instance is pioneering these types of services for CMS computing in production, serving data at scale. The XCache is populated with files coming from several remote CMS sites, hiding the latency of the accesses that, otherwise, would use the network to be opened. During the period of time shown in Figure 6.7 and Figure 6.8, the XCache was populated with files coming from several remote CMS sites as shown in Figure 6.10, with ~65.1% of files from sites in central Europe (excluding Spain), ~15.6% from America, ~8.1% from Spain, ~7.7% from North-Europe (UK and Finland), ~2% from Russia and ~1.5% from Asia.

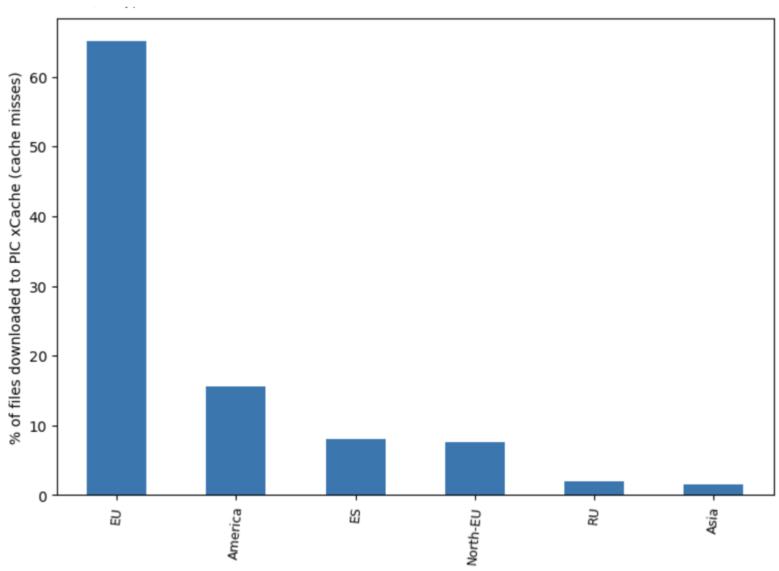
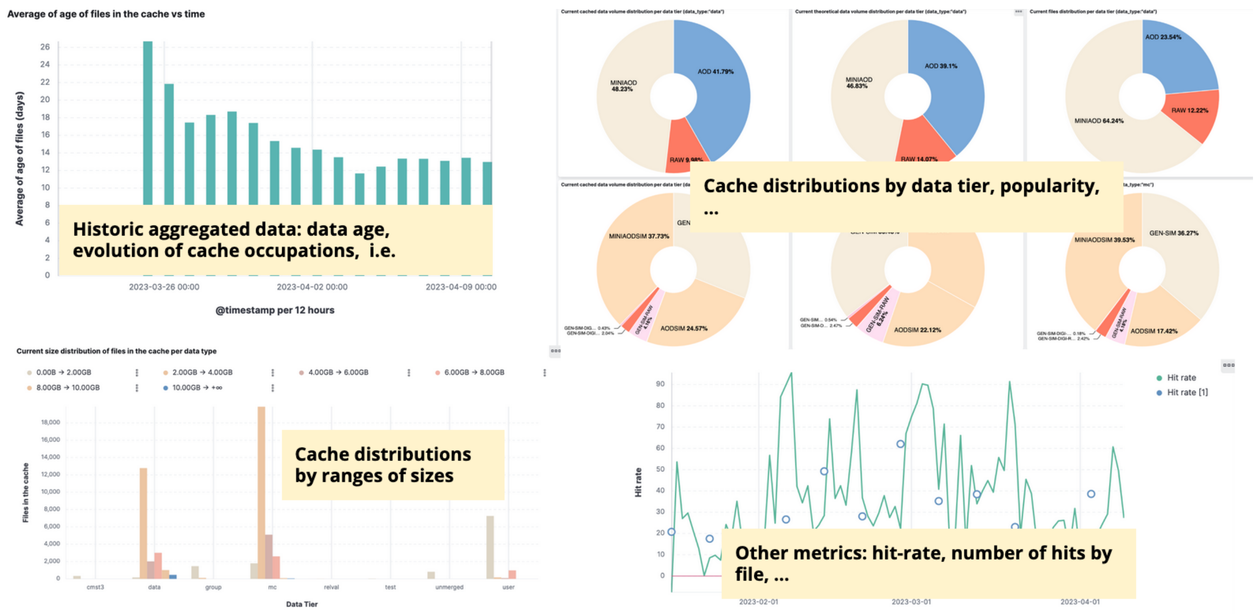


Figure 6.10: Percentage of files downloaded in PIC's XCache from external CMS regions.

For the successful deployment of the XCache service, monitoring tools have been developed as an outcome of a task effort between PIC and CIEMAT CMS collaborators, retrieving relevant information from the XCache *.cinfo* files. These binary files keep historical records of cached files accessed, information that is parsed and fed into an ES server. This data is then used to build dedicated Kibana graphs. In Figure 6.11 a snapshot showcasing some examples of the dashboards is displayed indicating the metrics that can be visualized.



Historic aggregated data: data age, evolution of cache occupations, i.e.

Cache distributions by data tier, popularity, ...

Cache distributions by ranges of sizes

Other metrics: hit-rate, number of hits by file, ...

Figure 6.11: Kibana dashboards generated through ES for XCache monitoring purposes.

The XCache monitoring process tracks three distinct events within the cache: 'creations,' 'accesses,' and 'deletions.' Consequently, it presents various dashboard categories in ES, each offering unique metrics that showcase updated cache information. First, the monitoring process generates daily snapshots of the cache's status and occupancy. These snapshots allow for the analysis of data distribution types, popularity, and file size distributions. Additionally, historical aggregated data plots illustrate how cache accesses, deletions, and creations evolve over time, offering insights into cache occupancy fluctuations with time. Furthermore, the monitoring service provides real-time values of critical metrics to manage XCache performance effectively. This includes information on the hit-rate, file distributions, and the duration of the files retention period within the cache.

Chapter 7

Data popularity studies

The XCache service enables fine-tuning some aspects of its configuration to improve its performance based on the specific use-case and implementation. The service has the capability to fetch remote files using the XRootD Data Federation when these are not found in the local SE where the jobs are executed. Each site counts with a JSON (the TFC) file that defines the rules for mapping LFNs to PFNs within a storage configuration, specifically for the XRootD protocol and, hence, for the Xcache server as well. These rules are used to translate requests for files from LFN to a PFN based on certain criteria. Modifying the rules can define which XRootD requests pass through the cache or not. This also includes the ability to define specifically which CMS types of data are desired to be cached. The choice of caching some types of data or some other should rely on such aspects as the repeatability of accesses (popularity) or frequency of access, that will eventually enhance the performance of jobs by keeping data locally in the data cache as much as possible. Studying CMS datasets popularity and files access patterns provides insights into how storage resources are being used and accessed by the LHC experiments.

Previous to the research that has been performed in this Thesis and presented in this Chapter, the ATLAS [208] and CMS [209] collaborations conducted similar, but incomplete, studies based on dataset access information, but not file-based studies. File-based studies can be performed both from the SEs perspective and from the execution jobs perspective. From the SEs perspective, the files are typically accessed by running tasks and also by transfers that are initiated by the experiment's data management systems, and can spot popular and hot datasets in the whole storage context. On the other hand, the execution jobs information is the ultimate relevant measure to understand how the data caches would populate, and yields information on how they must be configured to maximize cache hit rates and hence improve execution tasks CPU efficiencies. The research done in this Chapter focuses on the understanding of all the basic data characteristics (at file-level) and identifying popular CMS data from both SEs at PIC Tier-1 and CIEMAT Tier-2, as well as understanding the most frequently accessed data for the executed jobs at both sites. The two sites run the same storage manager (dCache), and same analysis techniques were applied to both sites, when focusing on the SE perspective.

7.1. PIC SE utilization

The dCache storage manager has built-in monitoring capabilities which provide an overview of the activity and performance of the SE by means of the dCache billing database (BillingDB [210]). Its primary function revolves around the meticulous tracking and recording of storage resource usage by distinct users or projects, facilitating precise file-based accounting. Activities like file creations, transfers, deletions, and access requests are recorded into this database. The dCache BillingDB records are stored in a PostgreSQL [211] database, and contains several tables in which the information is stored. The chosen timeframe for the analytical exploration of file studies spans from September 2017 to September 2018, aligning with the conclusion of LHC Run2. During the analyzed period, the CMS experiment at PIC Tier-1 managed a disk utilization of approximately 2.3 PB, in average. During the period, around 9 PB of data were written, involving 10.5 million files. Also, 24 PB were accessed from read operations, encompassing 3.5 million distinct files, while 9 PB were accounted for removal, accounting for 11.0 million files.

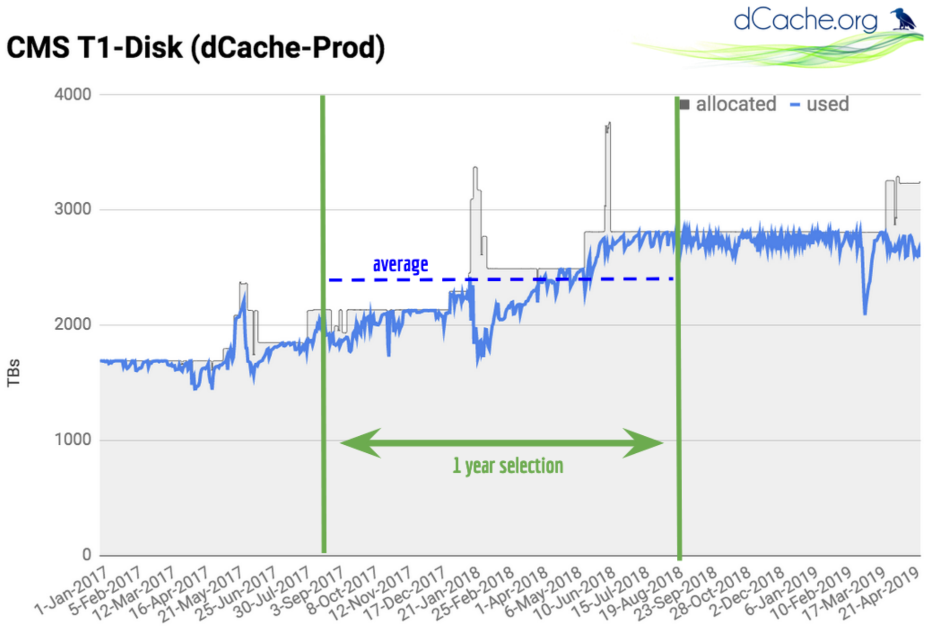


Figure 7.1: Average disk utilization for CMS experiment at PIC Tier-1 between 2017 and 2019.

Actions over the files come from jobs executed at PIC compute nodes, remote data accesses, data transfers from or to other CMS sites and activity in the fraction of disk at PIC dedicated to interface the PIC tape system. As seen in Figure 7.1, the disk at PIC Tier-1 is always working at saturation, hence the data management system from CMS has to delete files to allow for free space to be used by new written files. This is continuously performed along the year (indeed, the files written in a year are a factor ∼4 higher than the total available space). Note that "*allocated*" stands for space designated for potential use, but it may not yet be fully used or occupied by data or files, contrary to "*used*" space. To make room for new datasets slated for processing or to prepare space for upcoming processing campaigns, CMS Dynamo service [212] actively managed disk space by removing files from the disk areas. CMS Dynamo refers to a component of PhEDEx [213], which was developed to handle the distribution and replication of data in all of the WLCG sites used by CMS, that was later replaced by Rucio. It was responsible for managing the movement of data between various SEs and ensuring efficient access to the data for analysis purposes.

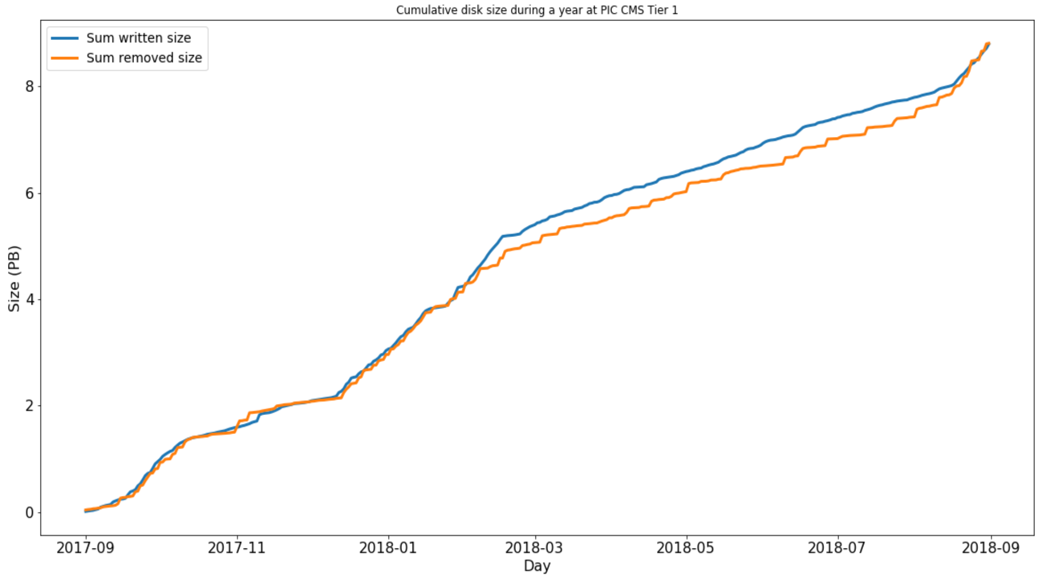


Figure 7.2: Written and removed size rate in disk at PIC for CMS experiment during the period 2017-2018

In 2020, the PhEDEx service was deprecated, with Rucio integrating the functions that Dynamo used to carry out for CMS. The continuous data recycle in PIC Tier-1 is shown in Figure 7.2. From the dCache perspective, Dynamo performed this data recycling of writing-deletion at a quite constant rate of ∼25 TB/day. This strategic cleanup involved either outright deletions or triggering transfers to tape storage, ensuring a continuous flow and optimization of storage resources in the different storage levels. Similar effects were present in other CMS SEs; in particular, the CIEMAT SE was also used at saturation during this period of study, with similar recycling volumes (as compared to their deployed storage). Usually, the most efficient way to operate storage systems in WLCG is to keep the most popular files on disk, so they are not deleted and replaced in an unnecessary manner. Ideally, one seeks swift access immediately upon the creation of a new file on disk, with minimal time intervals between the file's last access and its eventual deletion. Proceeding in this manner, files which are rarely used in disk should be kept on tape systems, and be re-called when needed.

7.2. Data popularity and access studies from PIC and CIEMAT SEs

To facilitate the study of data popularity and access specifically for the CMS experiment, a new and smaller PostgreSQL relational database was created, containing relevant CMS records from both PIC and CIEMAT full production BillingDBs. This smaller database was designed to efficiently store and manage the relevant CMS information, allowing access and analysis of the data in a straightforward and fast manner. Figure 7.3. shows a sketch of the workflow employed to generate the reduced BillingDB database, containing the relevant information from both sites. The reduced BillingDB database was installed in an independent server and contained 4 PostgreSQL tables specifically dedicated to specific actions in the file life cycle phases at both SEs: writes, reads, deletes and CMS data types. The resulting database contained 50 GB/year of information, which implied a relevant reduction as compared to the original 4 TB/year of data in the production BillingDBs. The SQL actions in the reduced database executed much faster and the studies were performed in a more efficient fashion. All the results provided within these studies were presented at the CHEP 2019 conference in Adelaide, Australia [214].

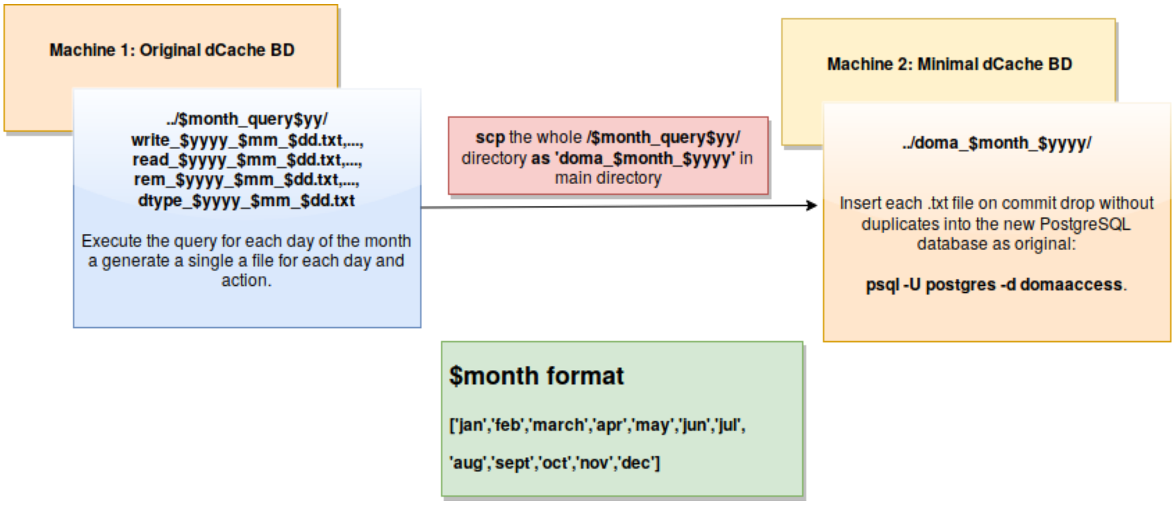

Machine 1: Original dCache BD
../$month_query$yy/
write_$yyyy_$mm_$dd.txt,...,
read_$yyyy_$mm_$dd.txt,...,
rem_$yyyy_$mm_$dd.txt,...,
dtype_$yyyy_$mm_$dd.txt
Execute the query for each day of the month a generate a single a file for each day and action.
scp the whole /$month_query$yy/ directory as 'doma_$month_$yyyy' in main directory
Machine 2: Minimal dCache BD
../doma_$month_$yyyy/
Insert each .txt file on commit drop without duplicates into the new PostgreSQL database as original:
psql -U postgres -d domaaccess.
$month format
['jan','feb','march','apr','may','jun','jul',
'aug','sept','oct','nov','dec']

Figure 7.3. Flow diagram for data transferred data from production billingDBs to the reduced one.

Over the selected period of study, the files were categorized into Collision data (named DATA in the following sections) and Simulation data (referred to as Monte-Carlo, or MC). By means of the number of file access during its lifetime, the so-called data popularity can be derived, as the mean value of file accesses for different CMS data types, as shown in (1).

$$Popularity = \frac{\sum_{i-file} Num._{acc,\,i}}{Total\ files} \qquad (1)$$

Figure 7.4 shows an example of the file accesses observed for all of the MC files created during the selected period at PIC Tier-1. Around 750k MC files were accessed 4M times, which implies an average file re-access of ~5.5 times. This average is a good estimator for data popularity. The Figure is displayed in log scale in y-axis, and the 95% percentile is shown for those bins with dark color, i.e. only 5% of MC files have experienced more than 8 accesses. These plots do not show the 0-bin, since newly created data can be created at the SEs but not being yet accessed by anyone at the CMS experiment. The number of files measured in this regime was <5% of the total considered.

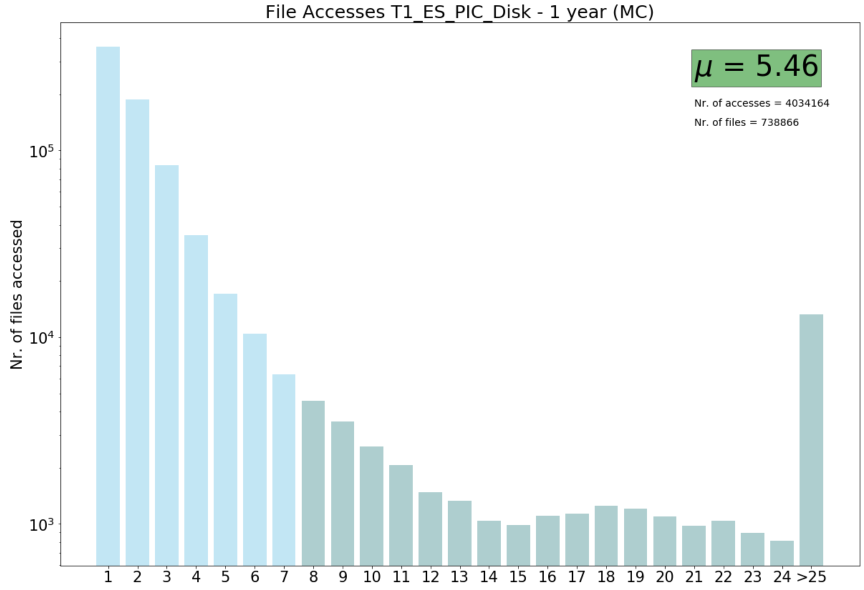
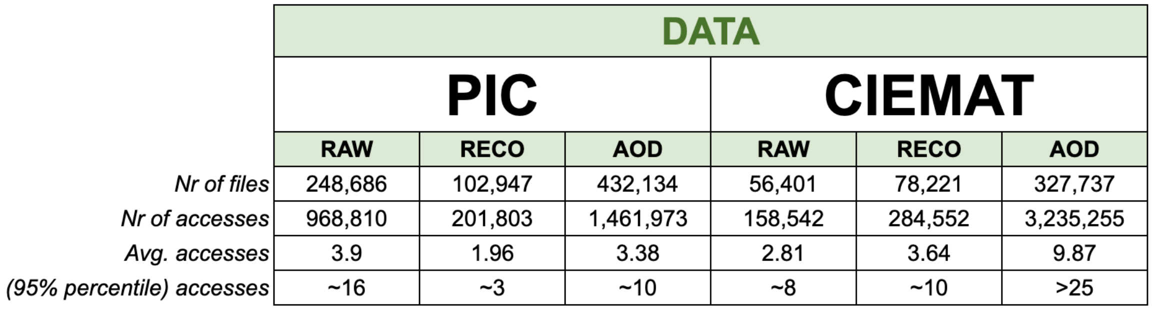


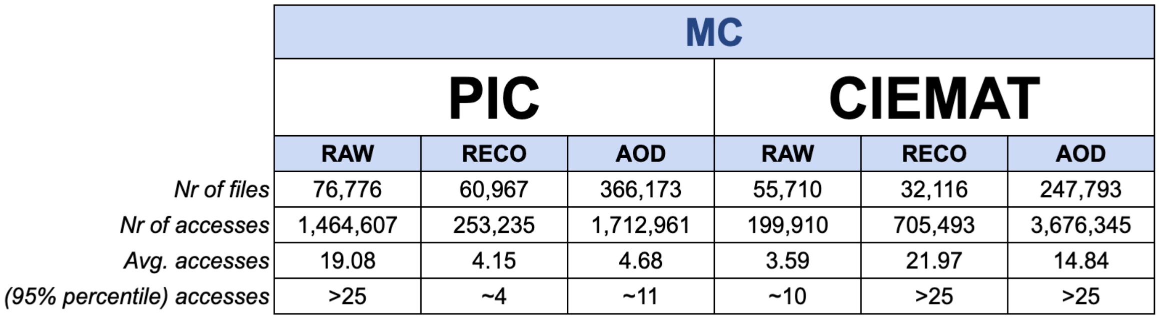
Figure 7.4: File accesses distribution (popularity) for MC files at PIC Tier-1 site (1-year period).

Tables 7.1 and 7.2 summarize the popularity results for the diverse data types at PIC and CIEMAT, for both DATA and MC categories. A total of 2M files and ~12.5M file accesses were considered to provide the inputs for these summary tables. Given the global scope of CMS operations it is quite safe to assume that data access patterns observed in this study would be similar as those obtained from other Tier-1 and Tier-2 SEs. Additionally, the file accesses are dominated by job executions (rather than file transfers across sites), so the popularity of data types identified through this SE study can be assumed as a general trend elsewhere across the Grid. Looking at the average accesses per file we can identify the data types that are popular and are likely ideal candidates to be kept in disk storages for longer periods (if needed). The results revealed that the AOD[8] files were really popular from the Tier-2 SE perspective, and that typically the file re-accesses are much higher at a Tier-2 SE than at the Tier-1 SE, which was expected since those datasets are much closer to the end-users. Many users run over the same datasets, and many analyses are repeated and refined, which explains why these files are popular. Additionally, the end-users have more CPU resource shares at the Tier-2 sites, since the Tier-1 sites are devoted mostly to central processing campaigns, and only a fraction of their resources are used by end-users. However, many campaigns run at Tier-1 to derive datasets for the final users, and this explains why the MC RAW files are re-accessed at the Tier-1 SE, since many reduced datasets are created from the same input MC RAW files. Differences between a Tier-1 and a Tier-2 are can be clearly appreciated in terms of storage composition and relative popularities of the respective samples.



	DATA					
	PIC			CIEMAT		
	RAW	RECO	AOD	RAW	RECO	AOD
Nr of files	248,686	102,947	432,134	56,401	78,221	327,737
Nr of accesses	968,810	201,803	1,461,973	158,542	284,552	3,235,255
Avg. accesses	3.9	1.96	3.38	2.81	3.64	9.87
(95% percentile) accesses	~16	~3	~10	~8	~10	>25

Table 7.1: Data popularity measurements for DATA type both at PIC and CIEMAT sites (1-year period).



	MC					
	PIC			CIEMAT		
	RAW	RECO	AOD	RAW	RECO	AOD
Nr of files	76,776	60,967	366,173	55,710	32,116	247,793
Nr of accesses	1,464,607	253,235	1,712,961	199,910	705,493	3,676,345
Avg. accesses	19.08	4.15	4.68	3.59	21.97	14.84
(95% percentile) accesses	>25	~4	~11	~10	>25	>25

Table 7.2: Data popularity measurements for MC data type both at PIC and CIEMAT sites (1-year period).

[8] In this Chapter, referring to RAW, RECO and AOD data tiers include their respective derived formats form MC and DATA according with the following: **MC RAW** → GEN-SIM-RAW, GEN-SIM-DIGI-RAW; **MC RECO** → GEN-SIM-RECO, ALCARECO, GEN-SIM-RAW-RECO; **MC AOD** → NANOAODSIM, AODSIM, MINIAODSIM; **DATA RAW** → RAW; **DATA RECO** → ALCARECO, RAW-RECO, RECO; **DATA AOD** → AOD, MINIAOD, NANOAOD.

7.2.1. *Data lifetime*

The CMS files are temporarily placed in disk areas at SEs. The data lifetime metric is characterized as,

$$\Delta t_{lifetime} = t_{deletion} - t_{creation} \tag{2}$$

where $\Delta t_{lifetime}$ is defined as the total file lifetime and $t_{deletion}$ and $t_{creation}$ the respective deletion and creation times of the file. In this study the survival percentiles of the total files analyzed and data tiers are computed. The survival percentiles of data lifetimes for DATA at PIC and CIEMAT respectively are shown in Figure 7.5. In the selected period at CIEMAT, approximately half of the DATA files of all types (RAW, RECO or AOD) were deleted within a month. At PIC, half of AOD DATA files are deleted within 10 days, and RAW DATA files are typically left on disk longer than the rest of data types. These RAW DATA files are usually kept on disk longer on purpose, since processing campaigns might need to re-process data, and it is more convenient to keep files on disk rather than restoring them from the Tier-1 tape systems. Regarding the MC files at PIC and CIEMAT, the results showed that this type of data lives longer at the Tier-2 than the Tier-1, where the space is constantly renewed to accommodate new processing campaigns. Average lifetimes are about 50% longer at CIEMAT than PIC for all MC datasets. Again, in both cases, half of MC AOD files are deleted within 10 days. The AODs are typically produced in Tier-1s and transferred to Tier-2 sites where most of the analysis are run, and they remain longer in those storages while they are popularly accessed.

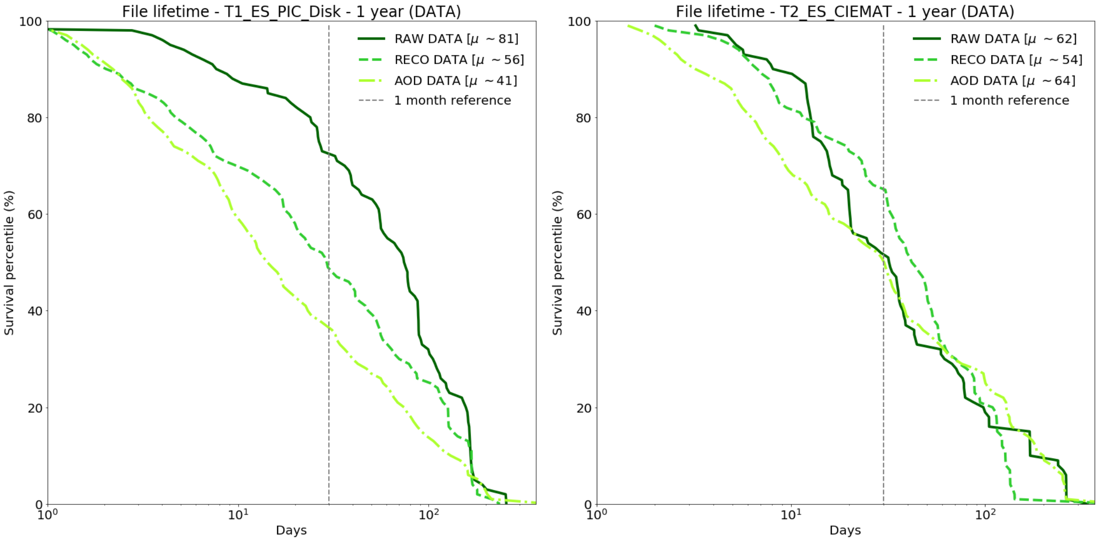


Figure 7.5: Survival percentiles of files lifetime for all DATA data tiers at PIC Tier-1 and CIEMAT Tier-2 (1-year period).

7.2.2. *File accesses*

During the adventurous lifetime of a file in a storage system, from its creation to its deletion, several accesses might occur. We can study the time interval between file creation and file first access $\Delta t_{creation \rightarrow 1st\ acc.}$ defined as:

$$\Delta t_{creation \rightarrow 1st\ acc.} = t_{1st\ acc.} - t_{creation} \tag{3}$$

The results for the selected period and file types are shown in Table 7.3. Approximately 20% of the DATA files are first accessed in their first day since creation, in both the Tier-1 and Tier-2 sites, while around half of them are first accessed within 10 days from their creation. The observed pattern was very similar at both sites.

				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
PIC Tier-1	DATA	RAW	201,661	113	71	34	6	2	0	0	0	21	29
		RECO	105,821	180	90	31	14	2	0	0	0	27	37
		AOD	434,657	155	104	42	8	1	0	0	0	31	47
	MC	RAW	63,264	87	72	20	1	0	0	0	0	16	30
		RECO	29,227	78	31	18	12	2	0	0	0	14	16
		AOD	310,521	178	61	24	6	2	0	0	0	21	35

				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
CIEMAT Tier-2	DATA	RAW	59,856	101	35	32	6	3	1	0	0	18	25
		RECO	92,865	112	79	40	10	4	1	0	0	25	32
		AOD	291,853	145	47	22	7	2	0	0	0	18	29
	MC	RAW	57,351	138	97	75	24	4	1	0	0	38	39
		RECO	28,654	77	28	16	3	1	1	0	0	12	18
		AOD	176,430	191	54	28	7	1	0	0	0	20	35

Table 7.3: Files creation to first access measurements for MC data type both at PIC and CIEMAT sites (1-year period).

As a reference, Figure 7.6. shows the time between file creation and first access for MC datasets at both PIC and CIEMAT. Half of RAW MC files are accessed within a day in PIC, since they are popular samples that have been either reprocessed quickly once created, or transferred to another site for a subsequent reconstruction campaign. The popular files have a prompt access, and the results showed that RAW MC files stored at CIEMAT are not very popular, since half of them are created and not accessed at all within the first month. On the contrary, for the other data types at both sites, approximately 80%-90% of files have been all first accessed within a month since file creation.

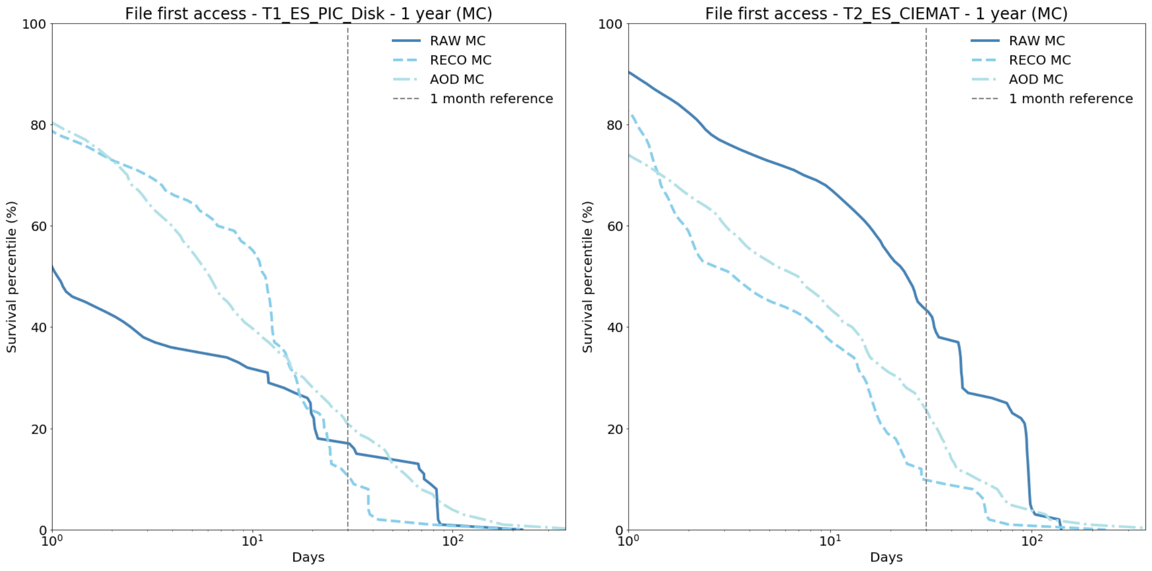


Figure 7.6: Survival percentiles of creation to first access for all MC data tiers at PIC Tier-1 and CIEMAT Tier-2 (1-year period).

Popular files are those which are read often, and the type of popular datasets might be different at Tier-1 or Tier-2 sites. For the purpose of evaluating these differences, time intervals between file accesses are computed. The time interval Δt_i between access *i* and *i-1* is:

$$\Delta t_i = t_i - t_{i-1} \tag{4}$$

Note that the time interval from creation to first access and the time interval from last access to deletion time is not considered on this metric. The summarized results for these times are shown in Table 7.4. Figure 7.7 shows the time intervals between file accesses for DATA types, for both PIC and CIEMAT, and in the period analyzed. Approximately 90% of RAW DATA re-reads happened within a day, since many jobs read the same input file in a processing campaign (job splitting). It can be observed that most of the files are re-read within a month, and there is a clear pattern for different data types. Popular files are read much more often, hence the time between re-reads is small. Similar distinctive patterns are seen for the three MC dataset types at both sites.

				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
PIC Tier-1	DATA	RAW	232,532	168	84	46	29	17	10	2	0	41	35
		RECO	91,636	145	101	34	13	4	2	0	0	32	39
		AOD	353,787	182	57	31	10	3	1	0	0	24	35
	MC	RAW	41,846	151	60	34	9	6	3	1	0	24	34
		RECO	66,226	168	128	76	11	7	1	0	0	41	47
		AOD	304,057	174	75	28	8	2	1	0	0	25	39

				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
CIEMAT Tier-2	DATA	RAW	44,030	162	40	31	20	12	5	2	0	27	31
		RECO	69,964	116	65	27	15	4	2	1	0	22	27
		AOD	267,940	185	86	33	12	3	1	0	0	30	47
	MC	RAW	56,302	173	103	101	30	16	11	1	0	47	44
		RECO	34,640	129	108	107	35	6	4	0	0	49	44
		AOD	189,719	204	121	45	8	2	1	0	0	36	52

Table 7.4: Intervals between access measurements for MC data type both at PIC and CIEMAT sites (1-year period).

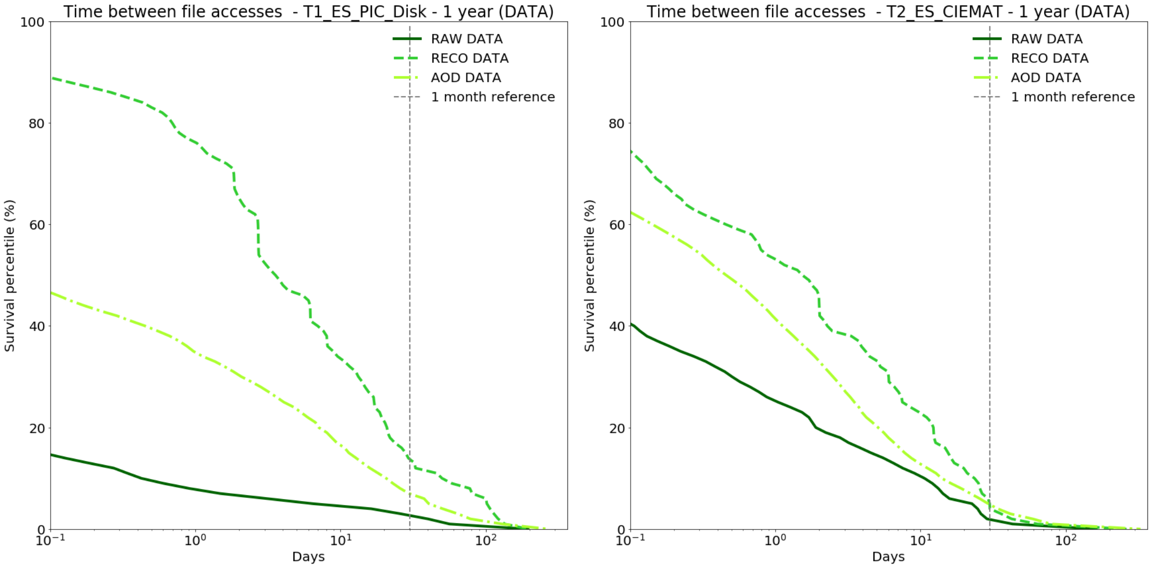
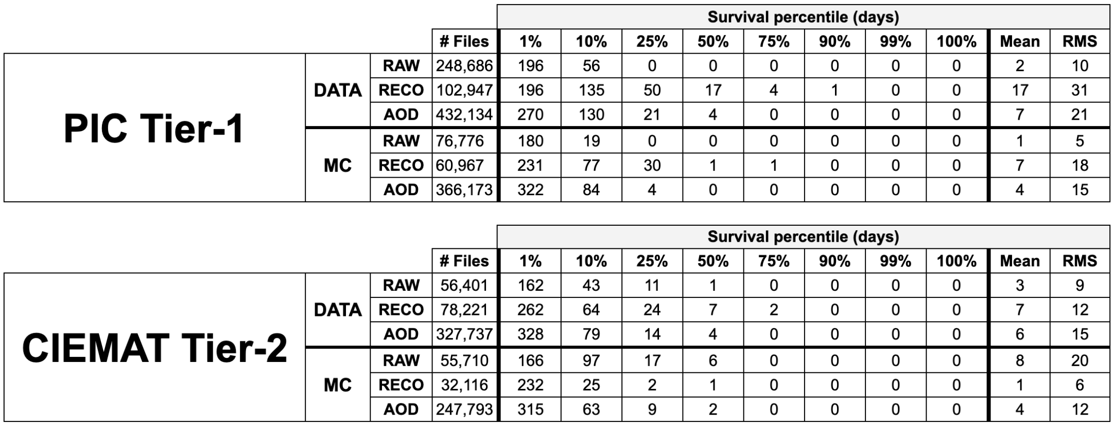


Figure 7.7: Survival percentiles of intervals between accesses for all DATA data tiers at PIC Tier-1 and CIEMAT Tier-2 (1-year period).

To complete the study, the time interval between last access and deletion from storage was evaluated as well. This is an important metric to monitor, since files that are not accessed anymore should be deleted from disk storages quickly. This metric is defined as the previous ones as follows in (5),

$$\Delta t_{deletion \rightarrow lastaccess} = t_{deletion.} - t_{last\ access} \qquad (5).$$

The results computed for all types of data can be observed in Table 7.5. Figure 7.8 shows this metric for DATA files in both PIC and CIEMAT, within the same period. As can be observed, RAW DATA at PIC is kept on disk longer than other data types, since last access. Half of the RAW DATA has last access to deletion time >1 month. For the rest of data types (RECO and AOD), ~25% of files have last access to deletion >1 month. At CIEMAT, it is measured that popular MC AOD samples are deleted promptly from last access, since 50% of the files have a last access to deletion time interval shorter than 10 days.



				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
PIC Tier-1	DATA	RAW	248,686	196	56	0	0	0	0	0	0	2	10
		RECO	102,947	196	135	50	17	4	1	0	0	17	31
		AOD	432,134	270	130	21	4	0	0	0	0	7	21
	MC	RAW	76,776	180	19	0	0	0	0	0	0	1	5
		RECO	60,967	231	77	30	1	1	0	0	0	7	18
		AOD	366,173	322	84	4	0	0	0	0	0	4	15

				Survival percentile (days)									
			# Files	1%	10%	25%	50%	75%	90%	99%	100%	Mean	RMS
CIEMAT Tier-2	DATA	RAW	56,401	162	43	11	1	0	0	0	0	3	9
		RECO	78,221	262	64	24	7	2	0	0	0	7	12
		AOD	327,737	328	79	14	4	0	0	0	0	6	15
	MC	RAW	55,710	166	97	17	6	0	0	0	0	8	20
		RECO	32,116	232	25	2	1	0	0	0	0	1	6
		AOD	247,793	315	63	9	2	0	0	0	0	4	12

Table 7.5: intervals between last access and deletion measurements for MC data type both at PIC and CIEMAT sites (1-year period).

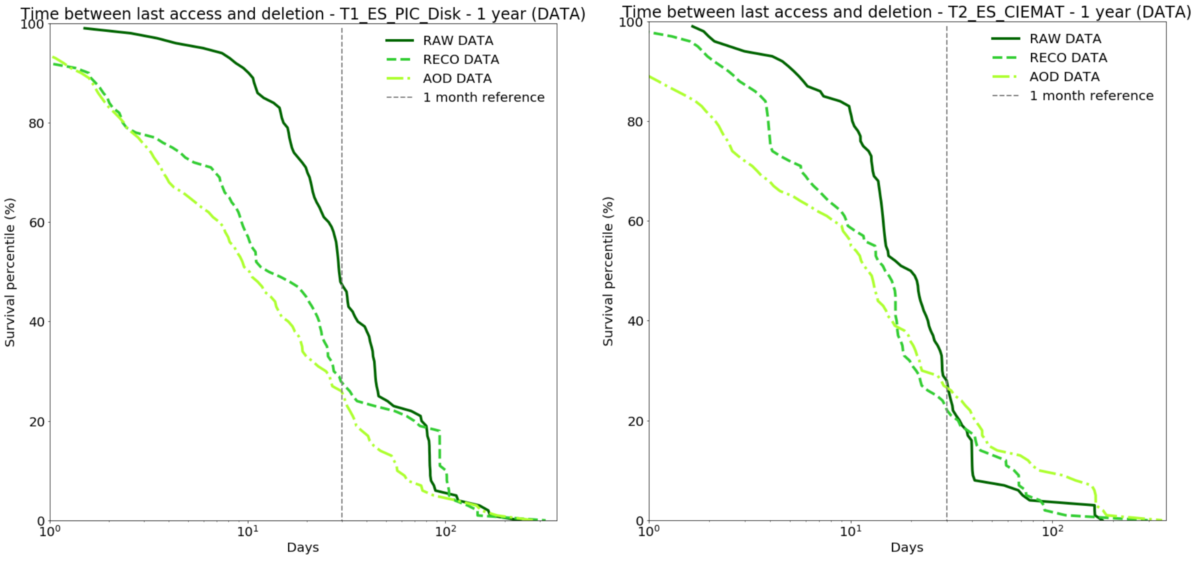


Figure 7.8: Survival percentiles of intervals between last access and deletion for all DATA data tiers at PIC Tier-1 and CIEMAT Tier-2 (1-year period).

7.2.3. *Data redundancy between PIC and CIEMAT as accounted by dCache*

The data redundancy in the region (for Tier-1 and Tier-2 levels) is constantly checked for storage efficient usage reasons. This redundancy in the region is very low, limiting the potential gains in disk volume that could be derived from an eventual storage consolidation. For example, CMS dataset blocks comparisons for PIC and CIEMAT on 30th October 2019 showed that out of 122.58k blocks (1.96PB) at PIC, and 56.17k blocks (1.84PB) at CIEMAT, only 2.23k blocks were present at both PIC and CIEMAT, with a total size of ~10TB, which is <0.5% of the storage used at the sites. Typically, the data redundancy in the region is measured to be always below 5%.

7.3. Data popularity from PIC and CIEMAT executed jobs

Understanding data usage and patterns is imperative for enhancing the efficiency of computing resources, particularly when dealing with disk-stored data, which can often act as a bottleneck for data retrieval. Gaining insights into how CMS jobs on compute nodes interact with data aids in pinpointing the most frequently accessed data, allowing the formulation of strategies for local caching of such data when it's read remotely. This approach effectively minimizes access latency for remotely accessed data, resulting in an overall enhancement of computing resource performance. The CMSSW package sends file access information for all of the running jobs to the CERN Hadoop (HDFS) infrastructure. The cmssw-popularity plugin sends small UDP packets to central servers at CERN, with details on the files accessed via the cmsRun application. This plugin relies on a UDP server *udp_server* (which runs as part of the execution tasks in the sites' compute nodes) and *udp_server_monitor* monitor application (which runs at CERN). Since the focus of the XCache usage is to improve applications performance by caching frequently accessed files, this data must be analyzed to develop the best strategies for caching data for the Spanish tier sites. All of the collected data from PIC and CIEMAT sites for the whole 2021 year was considered in this study [215]. Access to test files, such as HammerCloud or SAM tests, and other accesses to intermediate data (unmerged data) or PREMIX (pile-up samples placed at CERN and FNAL), as well as access to local files were filtered out in the study. The focus was primarily on understanding the access to remote files, in order to identify popular datasets that could be cached, potentially increasing the performance of the execution tasks at both sites. It is important to notice that since the monitoring system relies on UDP packets, some information might be incomplete due to lost UDP packets. However, checks were performed to confirm that the level of completion was high enough to trust the derived conclusions. Figure 7.9 and Figure 7.10 show the data popularity for DATA and MC data types, for all of the jobs run in PIC Tier-1 and CIEMAT Tier-2, respectively, reading data from remote sites and for the whole 2021 year. This overall picture shows that the jobs running in CIEMAT Tier-2 site have more file access repeatability, and could further benefit from data caching.

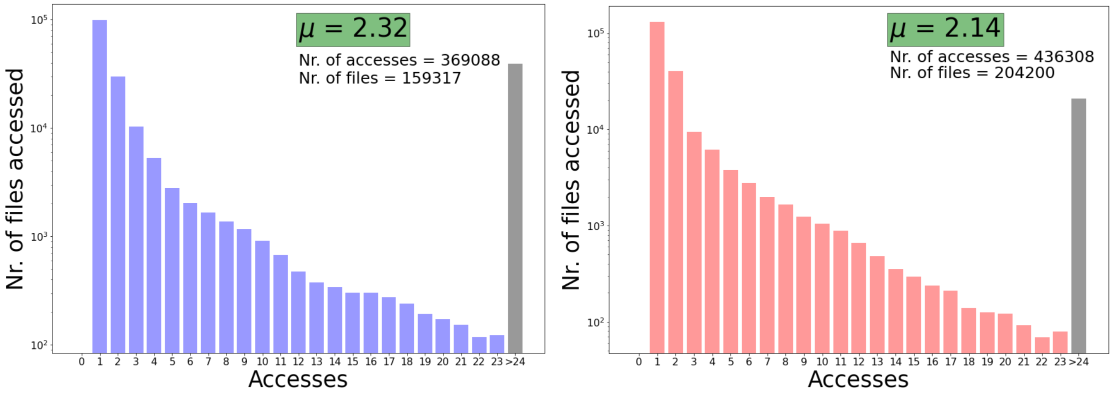


Figure 7.9: File accesses distribution (data popularity) for DATA files (blue) and MC files (red) at PIC Tier-1 site (2021 year period).

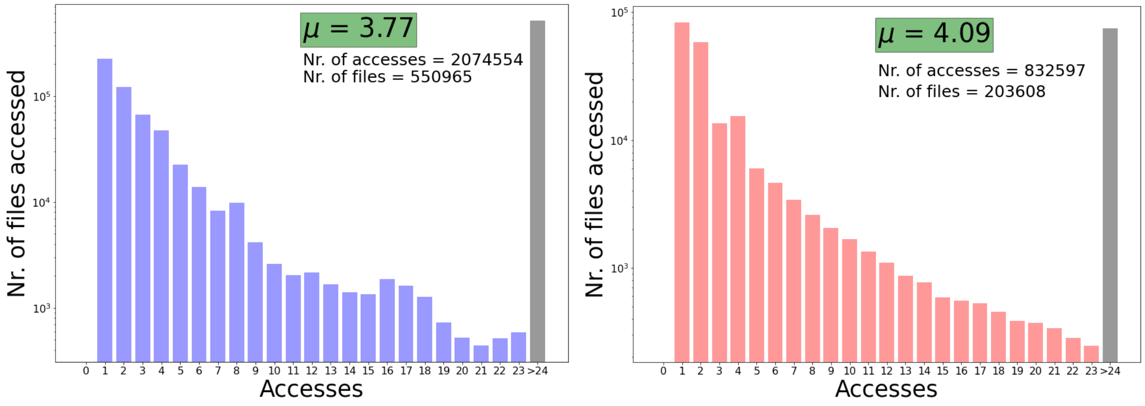


Figure 7.10: File accesses distribution (data popularity) for DATA files (blue) and MC files (red) at CIEMAT Tier-2 site (2021 year period).

The average number of accesses per file can be compared with the total volume of unique data accessed from each of the considered data types and data tiers. By means of this classification, one can identify the CMS data tiers in which the data is re-read often, and which should be the suitable size for being cached by a system deployed at each of the sites or a single cache in the region. Figure 7.11 shows the mean number of accesses per file computed over the total accessed files in the first ten months of 2021. The results show that files of type AOD are good candidates to be placed in cache systems, since MINIAOD and MINIAODSIM have the higher maximum values of mean access. Many of these files could not be completely read by the applications at run time, hence the measured sizes seen in this Figure must be treated as a maximum used space, assuming that all the files could have been read completely. The XCache allows caching files by blocks, and typically this option is enabled to optimize the use of the network and read performance. The total unique volume can reach up to 700 TB, which does not mean that data caches should be of that size (further discussed in Chapter 10), since data products lifetimes are at the order of months, and popular datasets can vary along the year. The deployed data cache system deals with these effects, in a dynamic way.

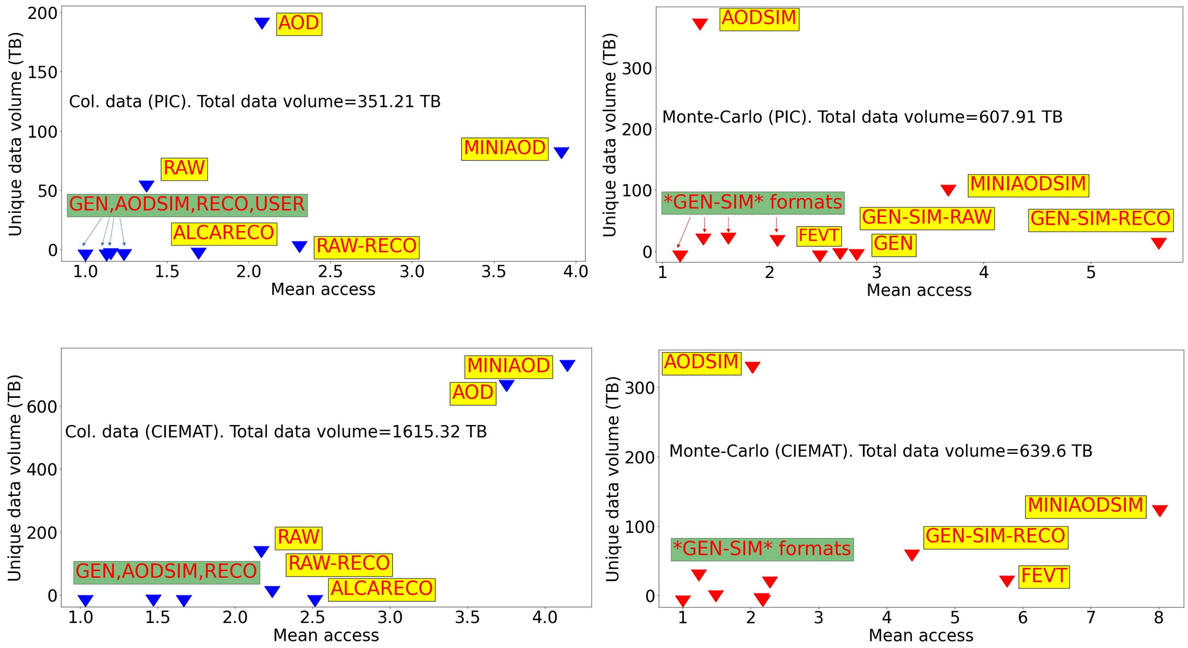


Figure 7.11: File accesses distribution (data popularity) for the PIC Tier-1 and CIEMAT Tier-2 for both MC and DATA, separated by data tiers and displaying the unique data volume accessed (for most of 2021).

As a clarification note, the data popularity results of this study yields mean access values smaller than those observed in the previously shown results of the similar study from the SE perspective. This difference can be explained considering that the dCache billingDB study includes an extensive record of all types of interactions and operations that a file undergoes within a storage infrastructure, and in particular, it includes local file accesses from jobs executed in local compute nodes, file accesses from jobs executed on remote compute nodes, and data transfers between local SE and other SEs. Hence, the SE perspective shows a complete overview of the file's activity and usage, but in order to cache files, the focus should be put on the data accesses that the CMS jobs perform. The cmssw-popularity information is then the relevant source of information used for the deployment strategy of the Spanish XCache server.

The examination of data access and usage patterns within the PIC and CIEMAT storage systems, coupled with an assessment of job usage both locally and externally, reveals that the most popular CMS data tiers are understandably those associated with Analysis jobs, specifically the AOD and its derived data tiers. This aligns with the initial expectations, since the end-users are the ones who read the same datasets often when performing analysis. The structured and closely monitored workflow of CMS production and processing jobs primarily focuses on RAW and RECO data tiers, integral to the processing and reconstruction task chain, aiming to efficiently produce final reduced data for subsequent analysis. In contrast, AOD formats are more used in Physics analysis conducted by CMS users, presenting a more unpredictable usage pattern and variability in accessibility. Several factors, including proximity to current scientific conferences involving this data and active involvement of Physics groups utilizing this data, contribute to the dynamic nature of AOD usage and its utility time range.

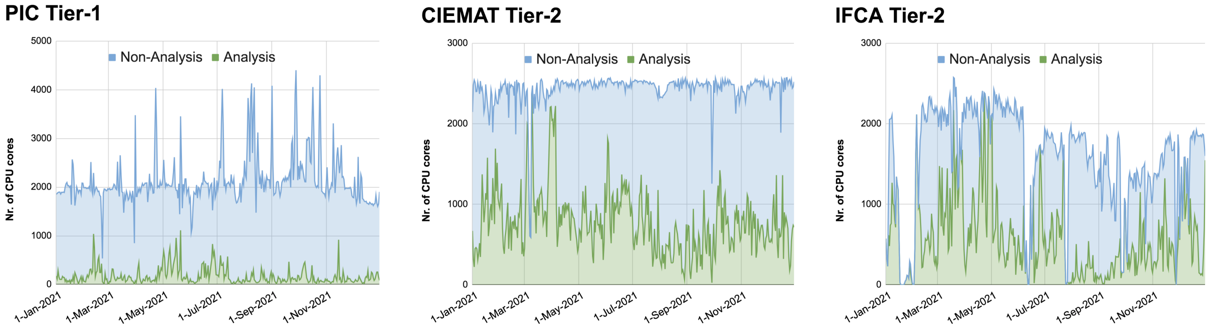


Figure 7.12: CPU cores used for Analysis and Non-Analysis activities at the Spanish Tier sites that support CMS computational activities (2021 year period).

Figure 7.12 shows the CPU cores used for Analysis and Non-Analysis activities at the Spanish CMS sites, namely the PIC Tier-1, the CIEMAT Tier-2 and IFCA Tier-2 computing sites. As expected, the CPU usage by end-users is higher at the Tier-2 sites. During 2021, around 8% of the CPU consumed in PIC was devoted to Analysis activities, as compared to the 32% and 37% of CPU consumed at CIEMAT and IFCA, respectively. Not all of the analysis activities read data from remote centers, but from the cmssw-popularity plugin it was estimated that around 25% of data was read from remote centers for Analysis jobs that were executed at PIC and CIEMAT sites. On the contrary, around 75% of data was read from remote sites for Analysis jobs executed in IFCA Tier-2.

It is worth mentioning that end-users can overwrite data locality when submitting CRAB jobs, and it seems that most end-users prefer to use the CPU at IFCA site while reading data from remote centers. This is important, since a deployment of a single XCache server in the region would be populated by these types of Analysis activities that run on all of these sites, and the Tier-2 centers would play an important role when populating and using the data cache service deployed.

7.4 Storage JSON rules

As explained in the beginning of the chapter, each CMS Site counts with a set of rules written in a code that enables the mapping of LFNs to PFNs. This set of rules can be currently found through a JSON file called *storage.json*. This JSON file defines a set of rules for translating logical file names to physical file names in a storage system. It uses the XRootD cache protocol and provides read-only access for the site. The rules are defined as an array of objects, each with two properties: LFN and PFN. The LFN property is a regular expression that matches the logical file name, and the PFN property is the physical file name to which the logical file name is mapped. Since the conclusions of the studies carried out in this Chapter suggested the data used by Analysis jobs were the most suitable to leverage from XCache, these rules were set in order to cache data from these jobs and avoid the rest of data used by other jobs. Here is an example of the storage JSON rules set at PIC to tell XCache when the data requested has to be downloaded:

```
{  "protocol": "xrootdcache",
    "access": "site-ro",
    "rules": [
       {  "lfn": "/+store/test/xrootd/T1_ES_PIC/store/(.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/$1"
       },
       {  "lfn": "/+store/test/xrootd/(.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/test/xrootd/$1"
       },
       {  "lfn": "/+store/(.*/.*/.*/.*PREMIX.*/.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/$1"
       },
       {  "lfn": "/+store/mc/(.*PrePremix.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/mc/$1"
       },
       {  "lfn": "/+store/unmerged/(.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/unmerged/$1"
       },
       {  "lfn": "/+store/(.*)",
          "pfn": "root://xcachecms.pic.es:1094//store/$1"
       },
       {  "lfn": "/+store/(.*)",
          "pfn": "root://xrootd-es.pic.es:1096//store/$1"
       }
```

The elements of the code can be read above. "protocol": specifies the protocol to be used, which is "xrootdcache" in this case. The "access": "site-ro": indicates the access level, which is "read-only" for the site. On the other hand, "rules": are the array of rules for translating the LFNs to PFNs. Each rule consists of two properties: (a) "lfn", being a regular expression that matches the LFN, and (b) "Pfn", which is the PFN to which the LFN is mapped. The JSON file also provides a structured way to access and manage data within the storage system. For example, if you want to access the file */+store/test/xrootd/T1_ES_PIC/store/data.txt*, one can use the following PFN: *root://xrootd-es.pic.es:1096//store/data.txt*. This logic applies for the rest of the rules. Again, rule 1 and rule 2 map LFNs in the */+store/test/xrootd* directory to different PFNs, depending on whether the LFN contains the substring T1_ES_PIC. The modification of the file is allowed to filter by indicating in the LFN which types of data have to be cached. In this case *PREMIX*, *PrePremix* and *unmerged* data are set apart from being cached due being used by non-Analysis jobs.

Chapter 8

Controlled environment test for XCache's CDN-based evaluation: benchmark jobs

The previous studies were based on an exhaustive analysis of real data access at PIC and CIEMAT centers. These studies had two main objectives: to understand data access patterns and to identify the most popular data to optimize the performance of the XCache. Overall, this task aimed to identify the ideal properties and configurations of a cache that covers remote XRootD accesses for CMS jobs that are executed in Spain. Beyond that, several aspects of the service must be evaluated to ensure the optimal configuration of an XCache-based CDN in Spain.

Despite its successful implementation in LHC experiments for many years, XCache still holds unexplored aspects that require further investigation, particularly in the context of its physical deployment. These aspects include the potential latency improvements during data access by jobs,the optimization of data transfer rates and network utilization. This Chapter presents the studies conducted in a controlled environment by submitting real CMS jobs, which have been named as "XCache benchmark jobs", with the objective to evaluate the benefits in latency and cost savings in walltime for the execution of CMS analysis jobs.

8.1. XCache benchmark job characteristics

Among all types of data generated by the CMS experiment, the final analysis files, the AOD and derived types, are the most frequently accessed data by the users of the CMS computing infrastructure. To optimize the performance and efficiency of the XCache system, it is crucial to carefully consider the types of files that are most suitable for caching, and study the potential benefits when using the service. Based on the research and usage patterns studies in this work, MINIAOD files accessed by user’s analysis jobs have been identified as the most appropriate files to be stored in a data cache. The CMS AOD files have a reduced set of reconstructed Physics objects for higher-level analysis, and MINIAOD files contain only the relevant information for faster processing and quick analysis, with the latter being the most accessed when performing final analyses for publications. To assess the potential advantages of caching MINIAOD samples, a comprehensive set of controlled jobs was executed in a production-like environment within the region.

The focal point involved employing a benchmark job [216] that required MINIAOD as input data, both with and without the XCache infrastructure deployed at PIC. This specific job, part of the muon Physics Object Group (POG), executed a tag and probe analysis targeting events stored in a MINIAOD file. The MuonAnalyzer package generates flat ROOT tag-and-probe trees and converts ROOT TTree into a *.root* file. This functionality signifies its involvement in muon data analysis, particularly in the field of tag-and-probe techniques. The package uses C++, the program that operates within the ROOT framework. The tag/probe method encompasses a loop traversing all events within the file, selecting a "tag" particle that satisfies the designated selection criteria, which could range from particle type identification to specific energy requirements. During the execution, certain tasks extend beyond the main event loop, such as initialization and output file writing. While a smaller event count reduces stage-out times and outputs, the execution task overhead remains relatively constant, irrespective of the number of processed events. This overhead was well-defined, occurring both before and after the event loop.

An initial research was conducted to determine the optimal number of events to analyze from the chosen MINIAOD file, given that the main event loop heavily influences the task's CPU efficiency. The benchmark job was run in single-core mode in a PIC compute node reserved for the tests (with 2 E5-2640 v3 CPUs, 64 GB RAM, and 10 Gbps NIC). This job required approximately 28 seconds for initialization before entering the event processing loop. Processing the entire 110,323 events in the template MiniAOD file (with a size of 2.9 GB) took approximately 6.1 ± 0.5 HS06·hours, with a CPU efficiency of 98.3 ± 0.2%. The application's peak memory usage stood at 1.47 ± 0.05 GB, and the average input file read throughput during job execution hovered around 2.46 ± 0.07 MB/s. Figure 8.1 showcases the CPU efficiency and memory utilization during the early phases of benchmark job execution. The initial execution time of the benchmark job is depicted, highlighting the initialization phase and the event loop phase.

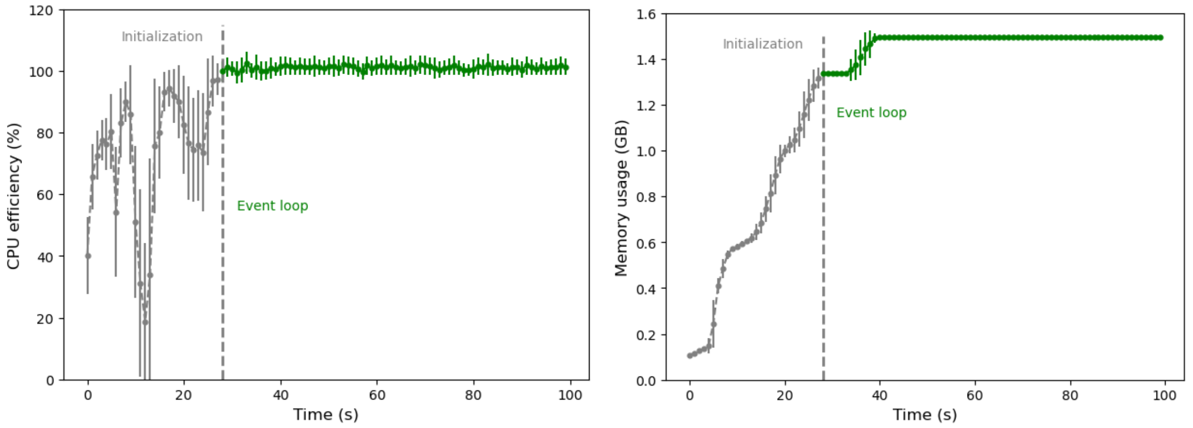


Figure 8.1: Initial execution time of the benchmark job where we can clearly see the initialization phase and the event loop phase. The CPU efficiency (left) and the memory usage (right) are shown.

On the other hand, Figure 8.2 presents how CPU efficiency reaches a plateau as the number of events processed increases. The efficiency stabilizes at 1% levels after processing 15k events and at 1‰ level at 65k events, relative to the final CPU efficiency obtained when processing the complete MINIAOD file. Consequently, fewer events could be leveraged for extensive testing of remote reads from various sites, showcasing no detrimental impact on test outcomes. The average file read rate approximates ∼2 MB/s, and the task's CPU efficiency remains remarkably high, affirming that the selected analysis is not I/O intensive in nature.

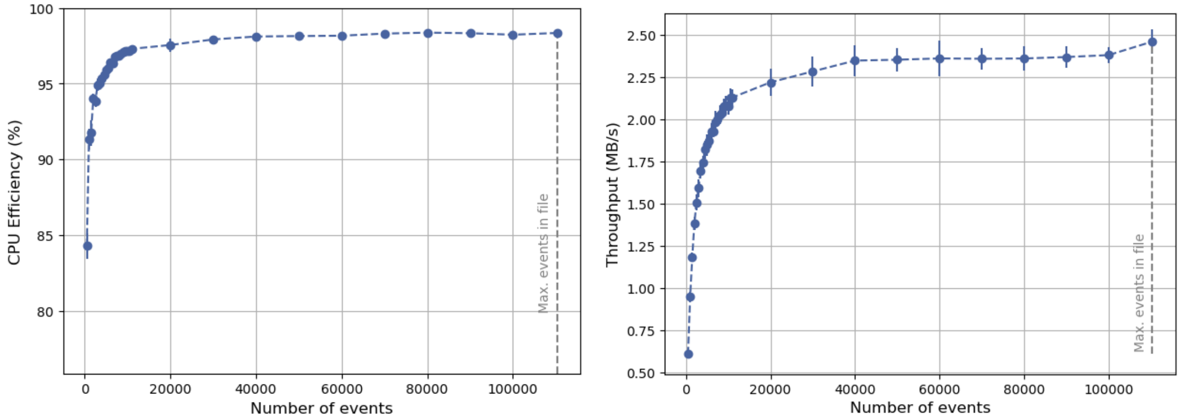


Figure 8.2: CPU efficiency (%) and events throughput (MB/s) as a function of the total number of events processed, from a task executed in PIC compute node reading from local XCache.

The slight variation in the CPU efficiency values during the final phase of processing the file is likely attributable to the overhead associated with initiating and terminating the processing of the file. Processing only a portion of the file necessitates multiple starts and stops, potentially leading to a marginal decrease in overall CPU efficiency.

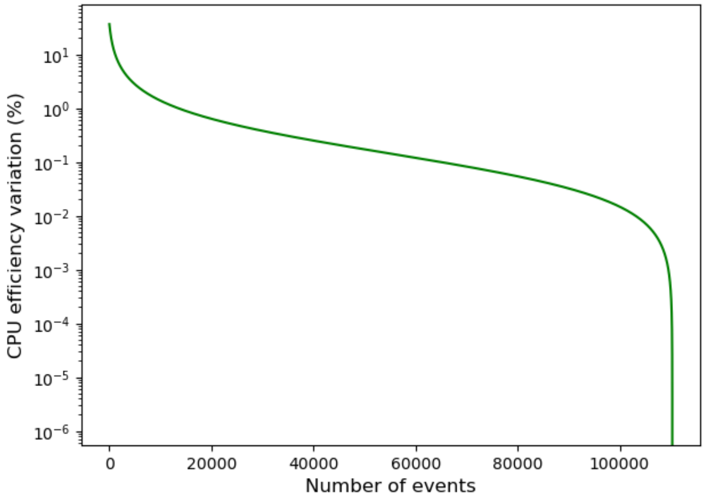


Figure 8.3: CPU efficiency variation (in %), with respect to the CPU efficiency obtained when processing the complete file. The CPU efficiency stabilizes at 1% levels after processing 15k events, and at 1‰ level at 65k events.

8.2. Submission of benchmark jobs to CMS Sites

Once the benchmark job was selected and its characteristics sufficiently known after preliminary results [217], a controlled environment at PIC was designed to submit the jobs that accessed MINIAOD files from local and many remote sites, with the aim to understand how the CPU efficiency of the benchmark job degrades when reading from distant sites. The PIC compute node used was emptied and separated from the HTCondor pool, to avoid conflicts or interferences with any other running jobs in the compute node. In the preliminary results, the CPU efficiency for a series of analysis jobs (serially) executed at PIC were studied, reading the input MINIAOD data files from both the PIC and CIEMAT caches. The serialization of the job executions ensures that only one job was accessing the input data at a time. First analyses using the environment highlighted valuable insights to achieve some of the objectives of this Thesis. After validating the controlled environment, around 25 sites were carefully selected to extend the results. This benchmark job utilized a single CPU core and the LHCOPN and LHCONE networks to access the input MINIAOD data. Approximately 100 jobs were executed, each reading events from similar MINIAOD files at every site, from the first to the 25th site. This process was repeated 100 times to maintain consistency across all selected sites and ensure a comprehensive set of tests over time. Then, the average CPU efficiency of the benchmark jobs was assessed when reading data remotely from these sites.

The entire test spanned 25 days, accounting for a total of 6.5k HS06·hours in the PIC compute node used for the test. In Figure 8.4, the average CPU efficiency of the benchmark jobs conducted at PIC is depicted, when reading data either from local storage (labeled as

T1_ES_PIC, with data stored in PIC XCache) or remote SEs, for sites placed in Europe (EU) and outside Europe (non-EU), as a function of latency (round-trip time, in ms). With the exception of a few sites displaying significantly poor or outstanding performance despite their latency, a noticeable trend of CPU efficiency degradation is evident for tasks reading from remote sites-.

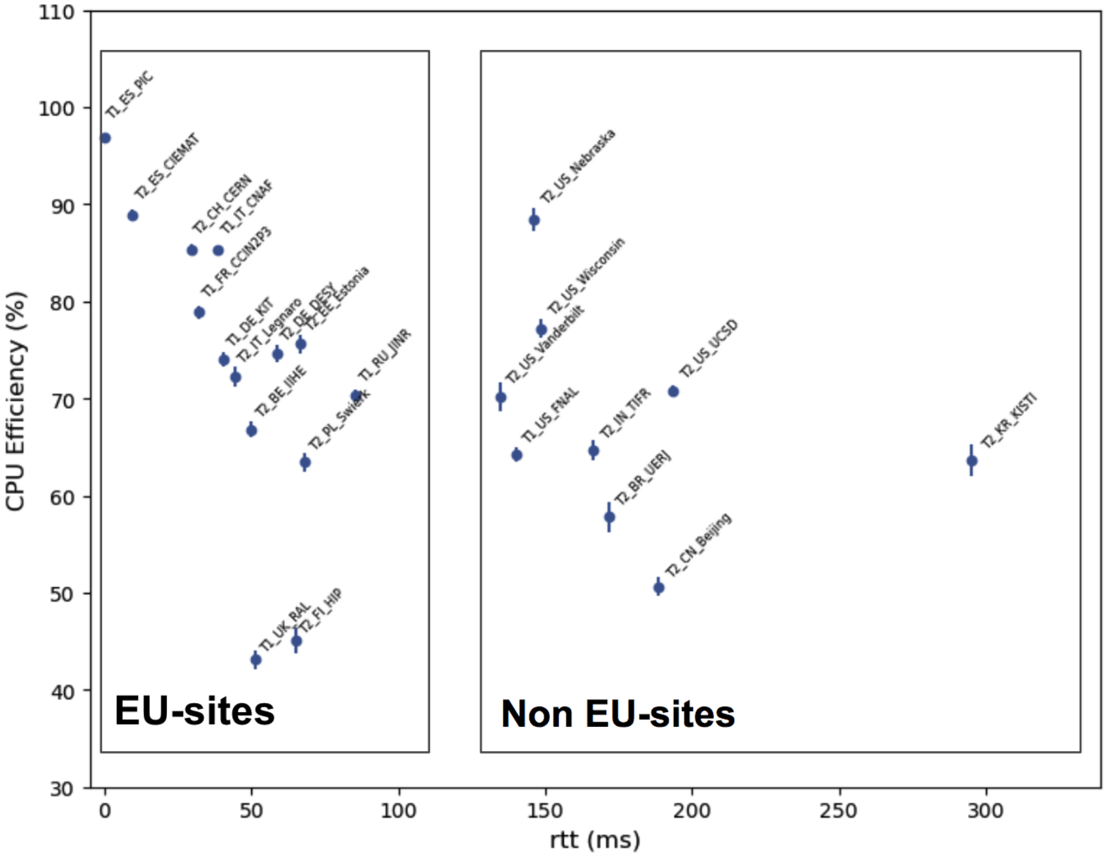


Figure 8.4: Average CPU efficiency of the benchmark analysis jobs executed at PIC when reading data from local, EU and non-EU sites, as a function of the site's latencies (round-trip-time [rtt], in ms).

Accessing data from sites in France, Italy, or CERN while the job runs at PIC leads to a considerable decline in CPU efficiency, dropping from 98% to 80-85% levels. These sites are located at distances of 650 km, 1,100 km, and 1,000 km, respectively, and exhibit similar round-trip time (rtt) values whose distributions can be observed in Figure 8.5. The FNAL Tier-1 site in Chicago, USA, situated around 7,000 km away from PIC center (with a latency of 150 ms), experiences a significant reduction in the mean CPU efficiency of these jobs, decreasing to approximately 65%. The furthest site tested in this study was in South Korea (KIST [218]), at a distance of approximately 10,000 km from PIC.

Although remote data access over transatlantic or transpacific networks is not the typical CMS practice, this study was conducted to evaluate the benefits of bringing data from exceedingly distant locations closer to compute nodes. Given that the CPU time for the executed task remains constant, the loss in walltime incurred by the job while accessing remote data can be determined by contrasting it with the maximum achievable CPU efficiency, which is attained when reading the data locally. The total computation of walltime lost by the CRAB jobs, which access remote data, will provide an evaluation of the benefit in terms of power consumption by the computer nodes at PIC.

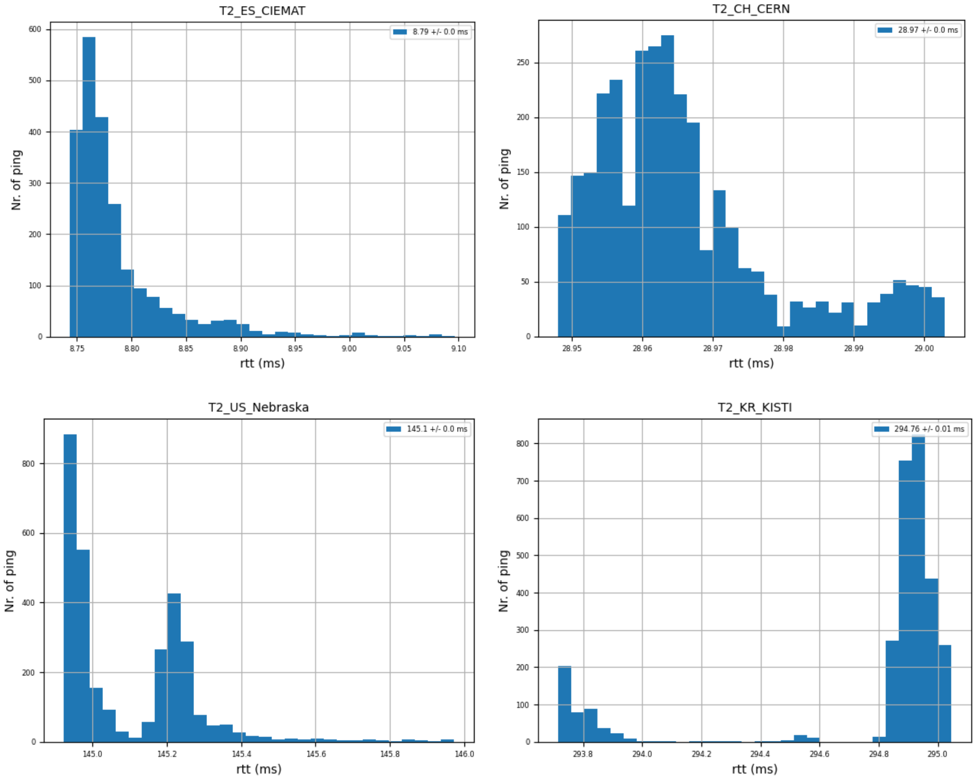


Figure 8.5: Examples of average ping RTT from PIC to distant sites (CIEMAT, CERN, Nebraska [219], and KISTI) within the period March - May of 2023 (round-trip-time [rtt], in ms).

Figure 8.6 shows the results of fractional walltime losses for jobs accessing external sites remotely, compared to if they had accessed them locally. The results of fractional walltime losses for jobs accessing external sites remotely are computed afterwards, compared to if they had accessed them locally. It has been estimated a potential saving of 1.8k HS06·hours, equivalent to 28% of the total walltime, of the the total 6.5k HS06·hours spent during the test could have been achieved. This outcome underscores the significant impact of latency and data placement on enhancing the efficiency of CMS analysis tasks. Therefore, it demonstrates that XCache improves the efficiency and walltime of these tasks by fetching data to deliver it locally, hence reducing the latency.

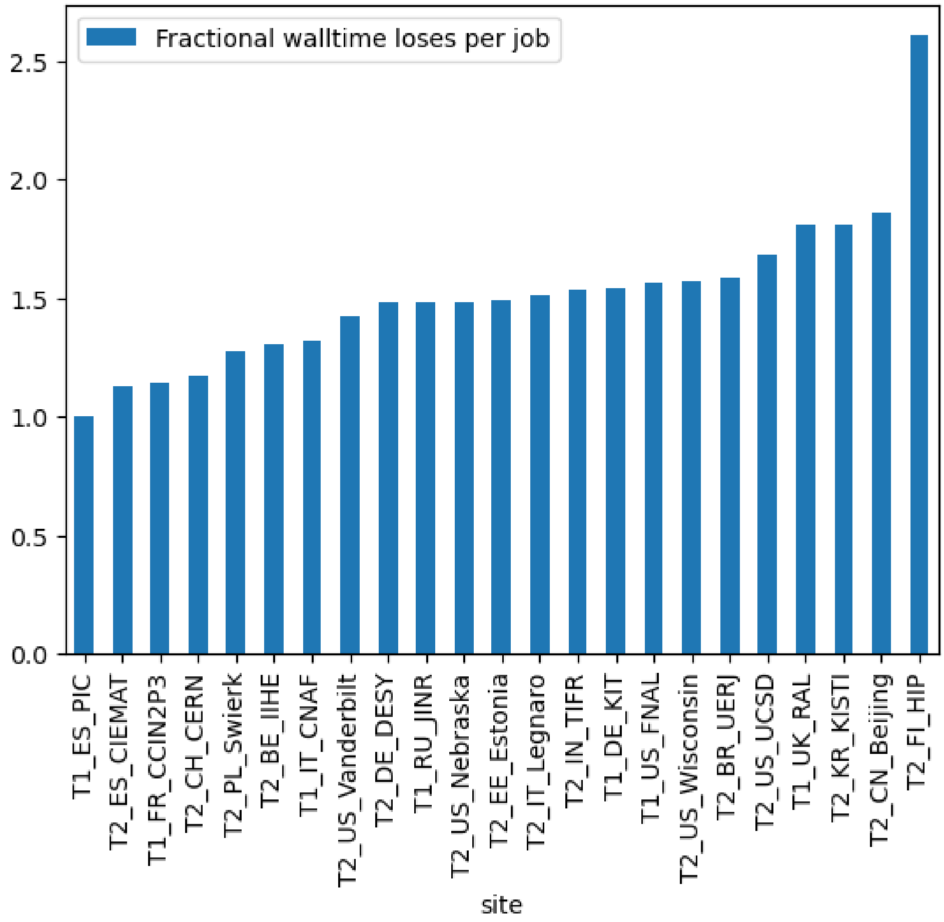


Figure 8.6: Fractional walltime losses per job of the benchmark analysis jobs executed at PIC when reading data from local, EU and non-EU sites, as a function of the site's latencies (round-trip-time [rtt], in ms).

Chapter 9

Results of XCache in production

After the benchmark jobs confirm the enhancement of CMS Analysis jobs with XCache and local data, the next step is to assess whether the observed positive impact extends beyond the controlled environment. This involves deploying the service in production for real CMS users executing jobs on the compute nodes in the Spanish region. To demonstrate if the CMS users' jobs have been properly executed and if they have accessed and/or downloaded data through XCache, while showcasing the improvement in CPU efficiency and walltime metrics, one must consult the vast individual job information provided by the monitoring services at CERN. These services provide information about these metrics and can also indicate, for example, whether the data has been read from local storage, or remotely from another site. For that purpose, comparisons between the jobs performance can be made with those jobs that have leveraged the deployed XCache in the region. CMS experiment has moved all of their monitoring infrastructure to widely used scalable, and non-SQL tools, such as Hadoop, InfluxDB, and ElasticSearch. These services are available through MONIT, a central monitoring infrastructure provided by the CERN IT department, and allow for the easy deployment of monitoring and accounting applications using visualization tools such as Kibana and Grafana. Relying on these CMS monitoring services, any user of the CMS experiment can access valuable information about data transfer movements between sites, job processing metrics, and the status of both computing sites and vital services. These measurements allow for a comprehensive understanding of how the CMS data is accessed and utilized, providing a wide understanding of CMS data utilization for optimizing CPU efficiency and job walltime metrics.

The analyzed raw data of monitoring is stored in the distributed HDFS storage system, while being accessed and analyzed through dedicated Big Data platforms deployed at CERN utilizing Spark, Hadoop, and Yarn [220]. These platforms can be easily exploited to perform data analysis through the dedicated web-based analysis portal called Service for Web-based Analysis (SWAN), offering to CMS researchers a powerful tool to conduct data analysis directly from their web browsers, without the need for specific software installations on their local machines, or connecting to specific remote interfaces. Through its web interface, Jupyter Notebooks can be executed, which by default utilize standard CPU and storage resources pre-assigned to individual users in EOS at CERN. However, SWAN also allows easy access to some other CERN resources, such as the Spark cluster which is used for Big Data analysis. This offers the gateway to access the whole CMS monitoring data which is stored in CERN HDFS, with an approximate total size of 12.5 PB. A schematic view of the access to Spark within SWAN in order to query data from the HDFS is displayed in Figure 9.1. Analyzing such a massive amount of data would be impossible without parallelization and Big Data tools. In the analysis performed in this chapter, the reduced information has been as well processed in PIC resources, using Dask and the PIC Jupyter platform.

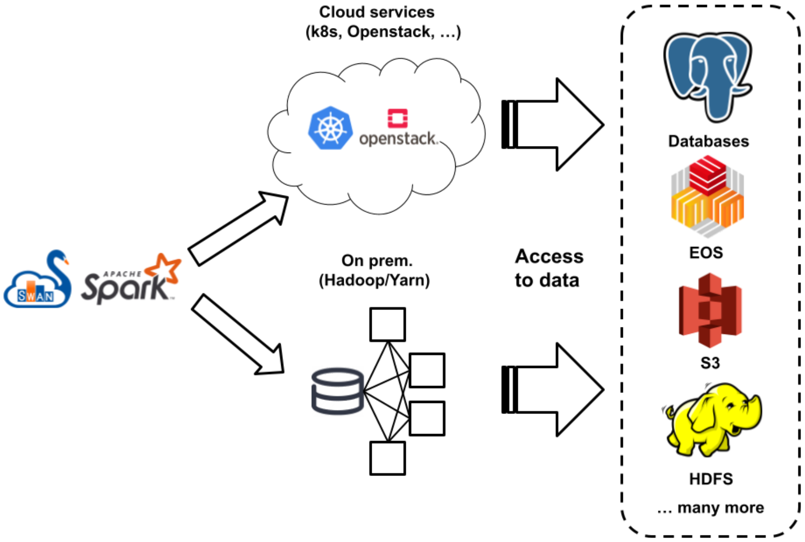


Figure 9.1: A schematic view of the access to different storage resources at CERN by allowing SWAN instances to use Spark cluster.

9.1. Locality determination of data accessed by CRAB jobs

Determining if a CMS job has remotely accessed data is a challenging task, since it has been learnt during the process of this Thesis that this information is not fully complete in the available standard monitoring views. Typically, the majority of the jobs read data from local SEs, but sometimes the central workflows or CRAB analysis jobs may ignore data locality, and read complete input files from remote computing sites. The end-users exploit this capability, if they want to run their analysis in CPU-stable sites, while keeping the input data on their host institution deployed Tier-2, or at CERN SE.

The currently deployed CMS monitoring infrastructure is not capable of fully monitoring all of the input data file accesses, since it is currently based on the cmssw-popularity (or PopDB) plugin. This plugin relies on UDP packets, and sometimes either the packets are lost or they are lacking the full job measures. Another possible source of information could come from the HTCondor job description (*jdl*) or the CMS job configuration file, but the full information of input files being accessed is not totally complete. Focusing on the CRAB jobs, all of the individual logs for each one of the executed jobs in any CMS site are sent and stored for three months in the CMSWEB infrastructure at CERN. These logs contain full information of the job execution tasks,in particular, any access to any file during job execution and the host that serves the data is clearly seen out of these logs. This provides users a way to check their particular CRAB log to identify failures, or debug their applications. Ongoing, the end-users can access these logs by HTTPs using unique URLs assigned to each of the jobs. Other specialized CRAB monitoring utilities such as CRABMon and the CMS Dashboard are available, providing detailed insights into submitted jobs, task status, configurations, parameters, user code, and the URLs to access to specific log files. For the purpose of better understanding the data caching implications, the information contained in the CRAB logs can be used to properly determine the input files characteristics, and be combined to other metrics that are stored in HDFS to access more relevant information, such as the CPU consumption of the jobs that are executed in the Spanish region. This approach provides a wide overview of each CRAB job's performance, detailing the read input files and their sources. This is crucial to understand which effect the XCache would have in the executed end-user jobs in the Spanish region.

9.1.1. *Processing the logs*

By means of the information stored in HDFS, one can easily get the complete list of CRAB log URLs for jobs completed at a particular site, in a given period of time. Taking into account that only the most recent CRAB logs (last 3 months) are kept at CERN CMSWEB service, one needs to build a machinery that constantly monitors the URLs pinpointing the log contents and parses the information before they disappear from the central services. CMS runs approximately 15 million CRAB jobs per month in its computing infrastructure. Considering the variability in the log sizes for each CRAB job, an average log size of 0.5 MB can be assumed. Consequently, the monthly aggregation of CRAB logs across all CMS sites would occupy approximately 7.5 TB/month. Since Spain represents the 4% of the total deployed resources, the access to the Spanish information would be of the order of 0.3 TB/month. Accordingly, the scale of data at hand necessitates the Big Data paradigm and sequential task parallelization. The performed study focuses on successfully completed CMS CRAB jobs executed on Spanish sites, which amount is depicted in Figure 9.2. The complete list of URLs that point to the CRAB logs is obtained from CERN HDFS using the CERN Spark Big Data Platform, accessed through the SWAN interface. As a reference, 100k logs per month are obtained for PIC Tier-1, which correspond to 100k completed CRAB jobs in the Tier-1 per month, on average, since this log-parsing tool has been available. The average number of CRAB logs accessed per month at CIEMAT and IFCA Tier-2 sites are 180k and 60k, respectively. Therefore, the list of URLs that point to CRAB logs is stored in *parquet* format [221], and the process of parsing the log is performed by other tasks.

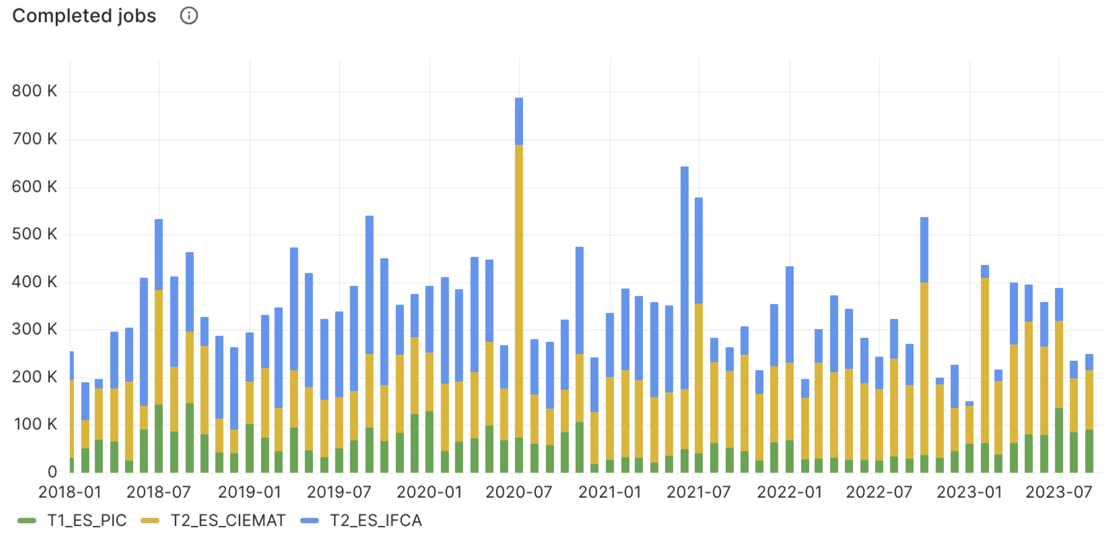


Figure 9.2: Completed CMS CRAB jobs per month at PIC, CIEMAT and IFCA (since 2018).

The procedure for parsing each individual log is outlined in Figure 9.3. The code developed for this purpose is designed to access the URL of each log, employing a personal and trusted x509 certificate within the pipeline for authentication against the CMSWEB server. It then proceeds to open the file, load its content into memory, and subsequently applies the parsing algorithm. This algorithm extends its functionality to include processing HTTP server requests by applying regular expressions, parsing, and iteratively navigating through all relevant fields to find matches. Ultimately, a data frame is obtained with the mentioned job's fields of interest, focusing primarily on which files the job has opened and from which site, determining if they were truly read from local SE or not. To expedite and enhance the efficiency of the process, the parsing code is parallelized across multiple threads. For this purpose, a Dask cluster at PIC was used, being one of the first formal analyses carried out through this platform at PIC. Dask is a Python library designed for parallel computing on extensive datasets, that enables the scalability of computations across multiple machines or cores without code modifications (or minimal adaptations). Dask offers a range of data structures and functions that simplify workload distribution and parallelization. For example, Dask enables the distribution of sizable *NumPy* functions across multiple machines and the utilization of its functions (in this case, the one used for accessing the URL and parsing the log) for parallel computations on the array. Furthermore, Dask extends its support to widely used Python libraries like Pandas and Scikit-Learn, making it a versatile tool for data processing and analysis.

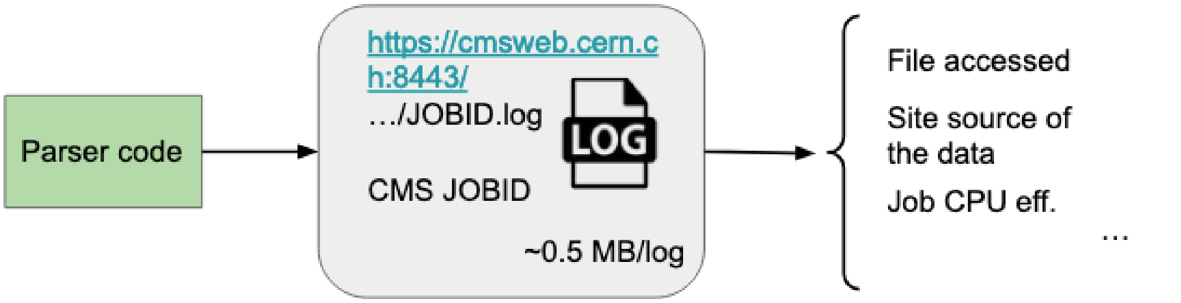


Figure 9.3: Individual parsing procedure for a single CRAB URL's log.

As shown in Figure 9.4, the original code is fine-tuned to parallelize the function that performs the URL request, *regex* matching, and parsing loop in a maximum of 10 threads simultaneously (in order to not be banned by the CMSWEB alarm system), one for each log, finally producing a single data frame including all the relevant information for the posterior analysis. Figure 9.5 illustrates an example on the interface monitoring the task stream, bytes used by compute nodes, and completion percentage of the parallelized parsing function for the CRAB log analysis. The application made a total of 37,996 function calls, including 36,793 primitive (built-in) calls while parsing 10k logs. The application's runtime was approximately 4.217 seconds, with the algorithm providing 23 logs parsed per second using 10 Dask executor nodes. This results in a speedup of the whole application in a factor 10 with respect to running the code in a single thread.

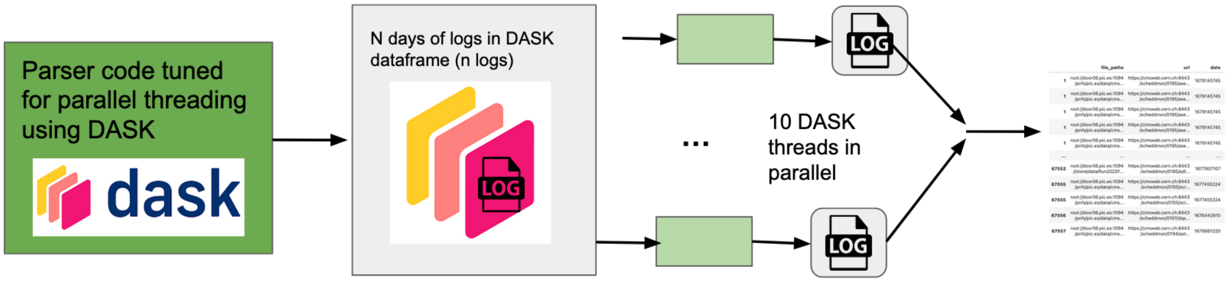
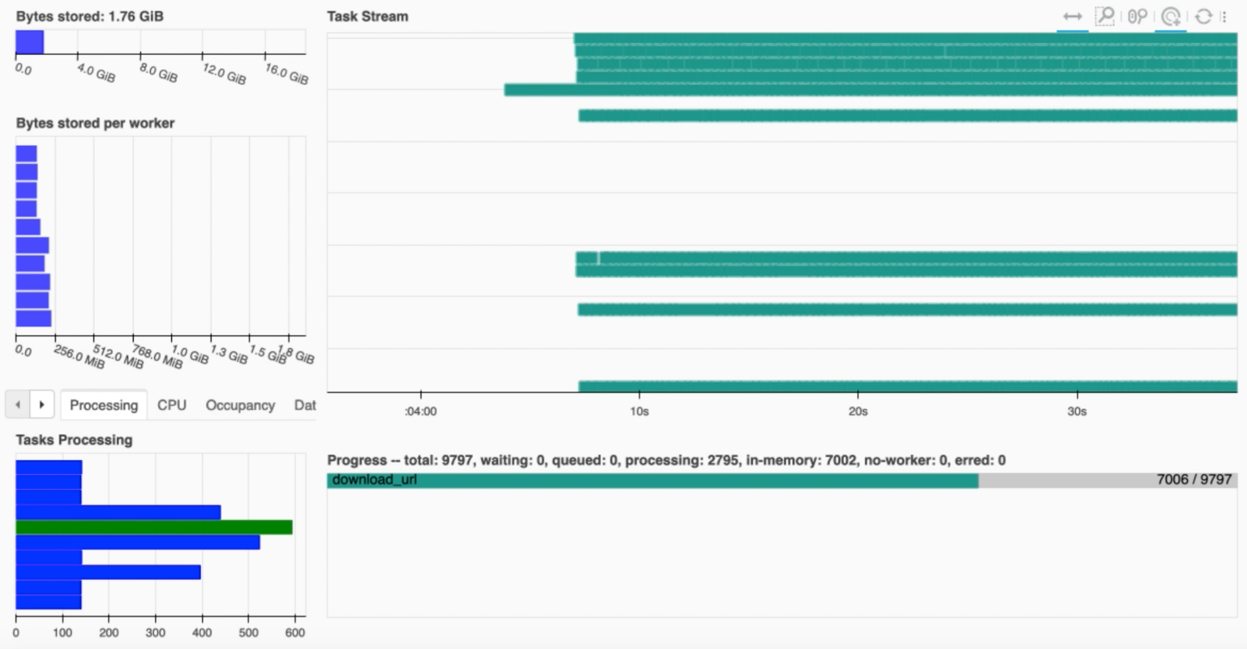

Parser code tuned for parallel threading using DASK
dask
N days of logs in DASK dataframe (n logs)
LOG
…
10 DASK threads in parallel
LOG
LOG
file_paths
url
date

Figure 9.4: Parallel parsing procedure using DASK for several CRAB URL's logs.



Progress -- total: 9797, waiting: 0, queued: 0, processing: 2795, in-memory: 7002, no-worker: 0, erred: 0

Figure 9.5: Interface view of DASK parallel executed function in PIC's dedicated cluster.

9.2. Enhancement of CRAB jobs efficiency by using XCache

Since the access to all of the individual CRAB job logs was enabled, half of the CIEMAT compute nodes were configured to use the PIC XCache as a fallback mechanism, while the remaining half directly utilized the CMS XRootD redirectors system. In both scenarios, if the input data files were locally available, they read directly from the local storage at CIEMAT. If not, two possibilities could happen for jobs reading remote data: half of the jobs would access the XCache server at PIC and benefit from the caching techniques, while the other half would still use many remote (and distant) sites to get the input data. This setup allowed us to investigate the impact of caching techniques as compared to the standard operational approach, since we can compare the performance of all of these CRAB job types.

Generally, users analyze popular datasets, leading us to anticipate that the XCache system performs effectively for analysis jobs, hence anticipating that we should observe higher CPU efficiencies for those CRAB jobs that use the PIC XCache service. As the logs were explored, many interesting insights about the diversity and behavior of the jobs were revealed. For example, many analysis jobs open more than one file, so the log parsing algorithm developed in this Thesis had to enable associating the unique Job ID with all the different files it has opened and from which SE or site. This additional feature is not provided by any of the existing CMS monitoring tools. Another phenomenon discovered in this study was finding that unique jobs can remotely open different files from different locations (mixed remote reads), although this phenomenon has been observed to occur only in about 1% of CRAB jobs analyzed. In particular, this phenomenon is of vital importance, because there are more than one distinct measures of data access latency affecting job efficiency. To ease the posterior analysis, all of these jobs with mixed remote reads were excluded from the study, and focusing only on jobs that have opened one or more files from the same SE. This decision is made because it would be essential to evaluate the degradation of CPU efficiency considering various site combinations and the distribution of I/O among them, a characterization that, as mentioned earlier, is beyond the scope of this thesis.

Over a period of 100 days (from January to April 2023), we measured and compared the average CPU efficiency for CRAB jobs executed at CIEMAT compute nodes with XCache enabled or disabled. Consequently, the computed results showed that the average CPU efficiency for jobs with XCache enabled was 77.2± 0.9%, whereas it was 70.4±1.0 % for jobs with XCache disabled. The normalized CPU efficiency distribution of CRAB jobs run at both parts of the CIEMAT compute farm is shown in Figure 9.6. Notably, the XCache's ability to serve popular files with reduced latency significantly improved the overall efficiency of analysis tasks on CIEMAT compute nodes when XCache was enabled. During this test, approximately ∼20% of the files were re-accessed from the PIC XCache server.

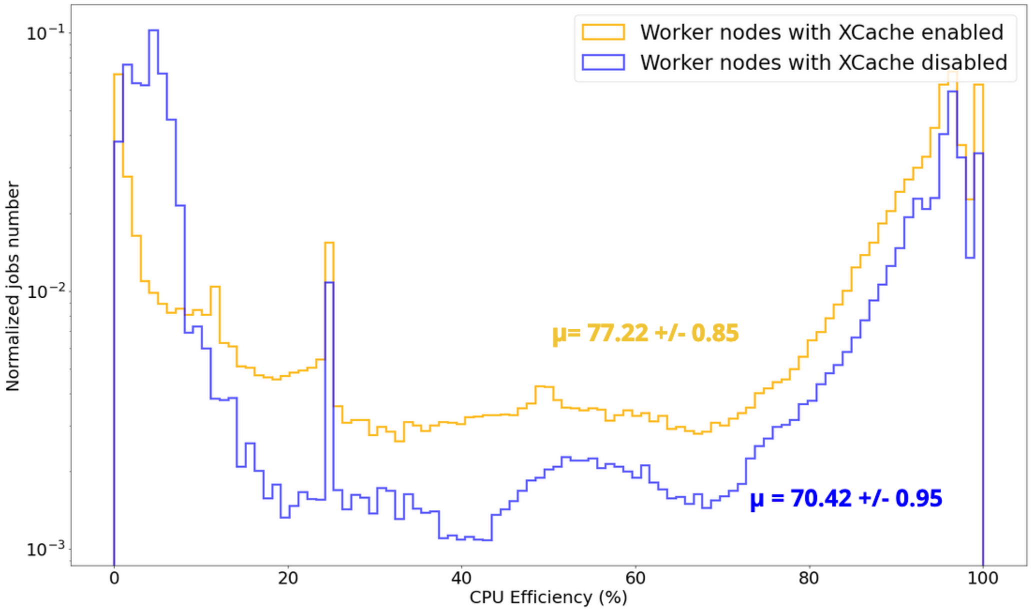


Figure 9.6: CPU efficiency distribution for CRAB jobs executed at CIEMAT compute nodes when XCache is enabled or disabled, in the period covering 100 days (from January to April 2023).

Figure 9.7 illustrates the breakdown of the average CPU efficiency for CRAB jobs executed at CIEMAT, considering data access from local SE, XCache at PIC, or remote CMS sites. When reading data locally (T2_ES_CIEMAT), the average CPU efficiency is remarkably high, at approximately ∼95%. Notably, when data is readily available in the PIC XCache and served to CIEMAT, the observed CPU efficiency closely mirrors that of local storage access. Conversely, the figure highlights the degradation in CPU efficiency when reading from remote and distant sites. In addition to network latency, several other factors can impact CPU efficiency, including the load on the remote storage system and WAN configuration and/or network load. It is remarkable that the degradation of the CPU efficiency for CRAB jobs reading from PIC XCache is not very significant. This result suggests that serving data from PIC to CIEMAT is an efficient process, since they are not very distant in terms of *rtt*.

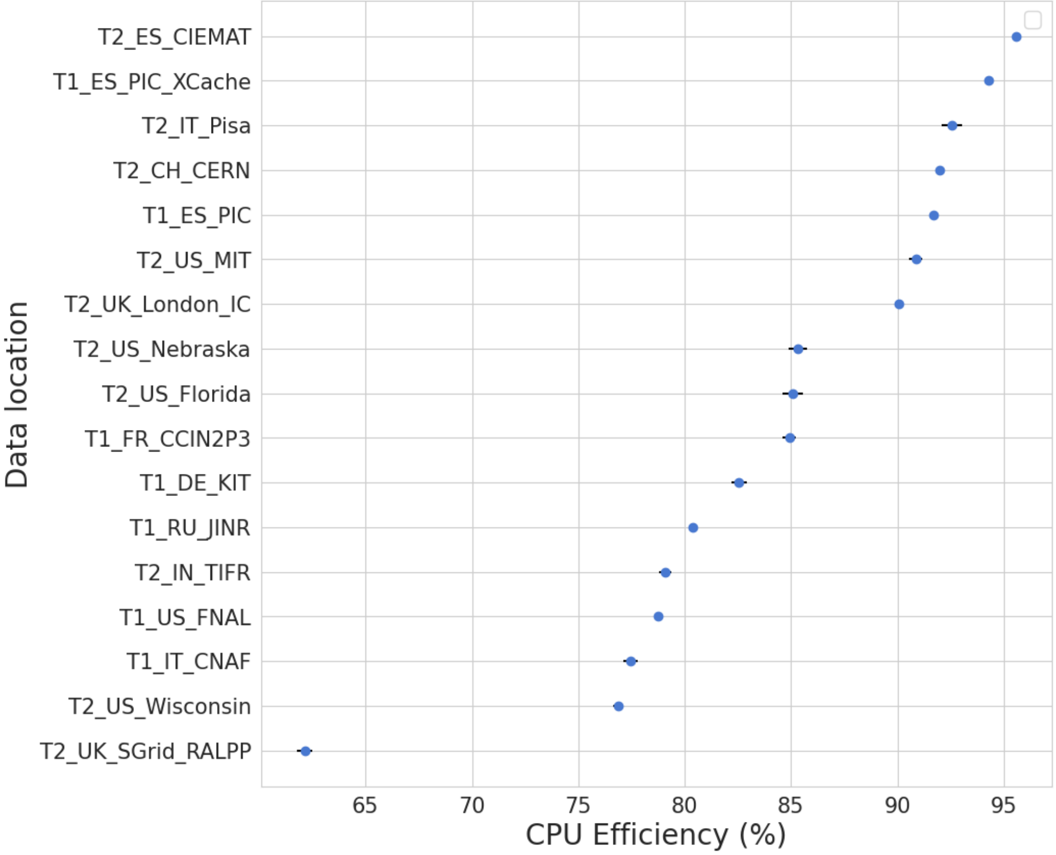


Figure 9.7: Average CPU efficiency of CRAB jobs executed at CIEMAT when reading data from local,XCache or remote sites, in the period covering 100 days (from January to April 2023).

Considering the consumed HS06·hours and CPU efficiency of these jobs, it is estimated that enabling access to PIC XCache for all CIEMAT compute nodes during this period of time could have resulted in approximately 13% savings in the total HS06·hours spent by these jobs at CIEMAT as can be observed in Figure 9.8. This result highlights the potential for performance improvement and resource savings in the region through the utilization of the XCache service. Either CIEMAT can perform +11% more computational work, or deploy

11% less CPU resources to perform the same work as previously done without using the XCache service.

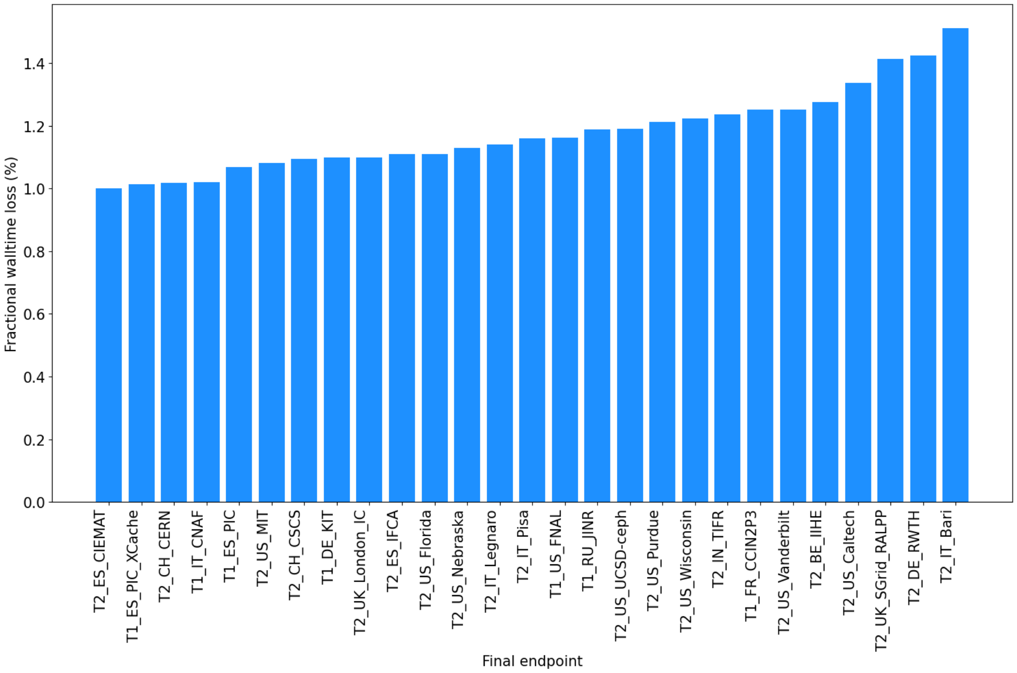


Figure 9.8.: Fractional walltime losses per CRAB jobs executed at CIEMAT when reading data from local, PIC XCache or remote sites, in the period covering 100 days (from January to April 2023).

Ultimately, the results of this study provide valuable insights into the potential for improving CPU efficiency for CMS tasks using an XCache service. This demonstrates that CMS CRAB jobs executed at CIEMAT compute nodes with enabled remote reads from PIC XCache show better performance than similar jobs executed at CIEMAT using the CMS global XRootD re-director infrastructure. The results also show that the CPU efficiency for tasks executed at CIEMAT reading from their local storage, or reading from PIC's XCache (if data was already in the cache), is very similar. This indicates that a single cache placed in PIC Tier-1 could effectively serve data to all Spanish CMS Tier-2 sites without a significant impact on the application performance. All of these results, among the final studies on benchmark jobs, were presented in an Oral contribution at CHEP 2023 in Norfolk, Virginia, USA [222].

9.3. Estimation of remotely accessed data rates by CMS jobs

A question that often arises within the CMS experiment community concerns the volume of data accessed remotely through XRootD by the executed CRAB tasks. However, accurately addressing this query has proven to be quite intricate, but by means of the developed analysis shown in this Chapter, we can estimate the importance of remote reads, beyond the Spanish region. The main contributors to remote XRootD reads are typically the centrally managed MC workflows, since they access and read events from the pile-up libraries that are located at CERN and FNAL (the so-called PREMIX files). These libraries have a size of ~1 PB, and a priori one could think they are ideal candidates to the cache mechanism, but they are not, indeed. Each MC job accesses several PREMIX files from these libraries to get a few random events that are then added to the simulated event. In general, the level of file repetition is very low, and each of the PREMIX files is opened to get a small number of random events.

To make this addition process to be statistically consistent, the PREMIX files contain many events, and the files are very big (approximately, a factor 10 higher than data files). Even if this access behavior is not suitable for caching techniques, we tested caching PREMIX files, which is an open issue for MC generations that are run in HPC centers (i.e. placing the pile-up samples in a cache close to HPC resources, in the context of integrating CPU resources from the Barcelona Supercomputing Center). This resulted in memory overloads on the disk server hosting the XCache, leading to very low data reusability. Consequently, this caused substantial thrashing in the cache, amplifying the network usage. If the XCache needs to properly perform to access such big files, it would require further developments. An estimation can be made to evaluate the amount of PREMIX data which is read via XRootD by the MC simulation campaigns.

CMS has been running an average of 6k MC simulation jobs in the WLCG infrastructure. This particular workflow consumes pile-up data at 2 MB/s, so the aggregated estimated global throughput for reading pile-up though XRootD is on average ~12 GB/s worldwide. As a comparison, the data which is currently moved by the FTS service across CMS sites is as well at ~10 GB/s levels. With the ability to parse the CRAB logs, the amount of data movements generated by remote CRAB reads via XRootD can be estimated as well. With the accumulated data of 2023 for the PIC Tier-1, CIEMAT Tier-2 and IFCA Tier-2 sites, the measured data rates from remote reads via XRootD accounts for an average value of ~0.4 GB/s in the region. Since these sites contribute to CMS for about 4% of resources, if the rest of regions behave as Spain, the accumulated data rates generated by remote CRAB reads via XRootD would be at the order of ~10 GB/s.

Using the same techniques applied for Spain, understanding all of the file accesses would have been a very difficult task, given the huge amount of data to analyze. However, leveraging the Big Data and Parallelization framework outlined in section 9.1.1, extensive data analysis of the related logs to the overall jobs executed in the Grid by CMS was feasible, but for just one month, parsing about ~15M CRAB job logs (~7.5 TB of logs to query!). Figure 9.9 shows the percentage of CRAB jobs that read files remotely or locally, for the selected month of April 2023, and for all of the CMS sites worldwide. Note that not all of the sites behave similar, and that different regions might differ as compared to Spain. A more refined analysis would be required at this point, but during April 2023, approximately 43% of the CRAB jobs executed in the WLCG by CMS accessed remote data via XRootD, which represented around ∼30% of the total CPU usage by CRAB jobs. In Spain, the amount of CRAB jobs that access remote data is estimated to be around 20%, taking all of the cumulative data from 2023. Hence, the XRootD traffic that the CRAB jobs impose into the global CMS network might be higher than 10 GB/s. In other words: the traffic that CRAB jobs generate might be at the order of the sum of the one generated by FTS transfers and the one generated by the read of PREMIX libraries. Unfortunately, this XRootD traffic generated by CRAB jobs that read remote files is not yet accounted for in any of the existing CMS monitoring tools. Nevertheless, this outcome highlights the role of XRootD in handling traffic loads for CMS analysis tasks, a facet that is often overseen.

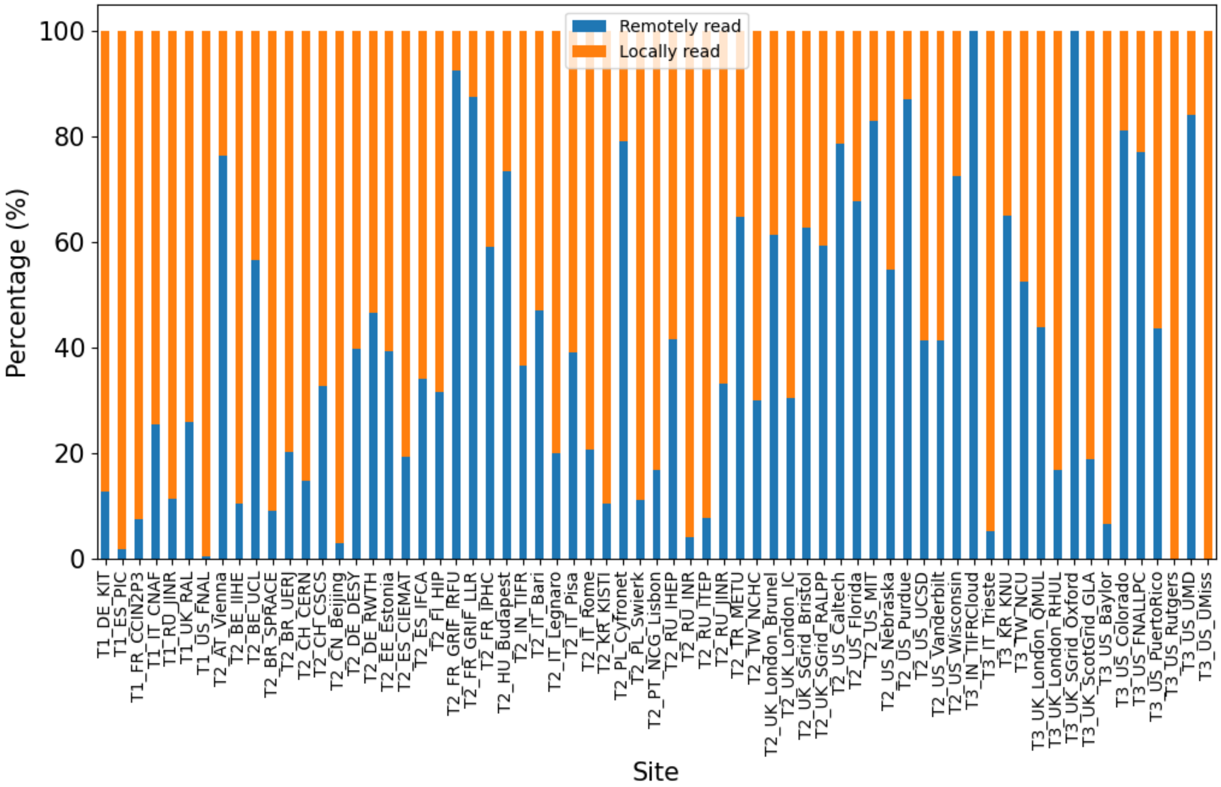
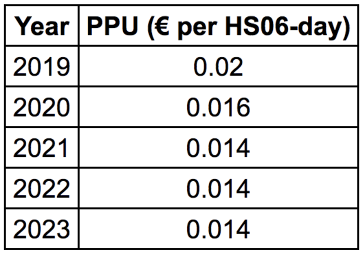


Figure 9.9: Percentage of CRAB jobs that accessed remote/local data through XRootD at each of the CMS sites (in April of 2023).

9.4. Estimation of costs-savings by deploying data caches

Including data caches into the system and improving job performances have several benefits. Either more 'work' can be done with the same amount of CPU resources deployed, since, as seen in the previous sections, an improvement of ∼10% can be obtained on the CPU efficiency of end-user submitted tasks. Since data latency accesses are reduced by XCache, the improvement goes directly into the walltime of the submitted tasks, hence the users get their results 10% faster. This means that, with the same CPU pledges, an additional 10% of ‘work’ can be performed with the use of data caches. An alternative is to provide 10% less CPU resources to continue offering the same amount of delivered work, although the pledges in WLCG are measured in CPU power, rather than delivered outcome (such as events/s produced or processed). Another benefit resides in the amount of storage to be deployed in the region. Typically, the storage resources are managed by the experiment, but a fraction of the local storages at Tier-2s are devoted to the local Physics Groups. Maybe, this space could be lowered with the use of data caches. This is particularly interesting, since the storage resources are the most expensive to operate in the Grid. However, the complete pledges declared to WLCG include both space managed by the experiments, and the one devoted to local users. The latest is typically extended beyond the pledge by many countries to satisfy the local requirements, and there the data caches might be of great interest to be explored. Additionally, managing a computing center with no storage and reading data from regional caches is growing in WLCG. It not only saves costs associated with the storage services, but also personnel costs to manage and operate such services.

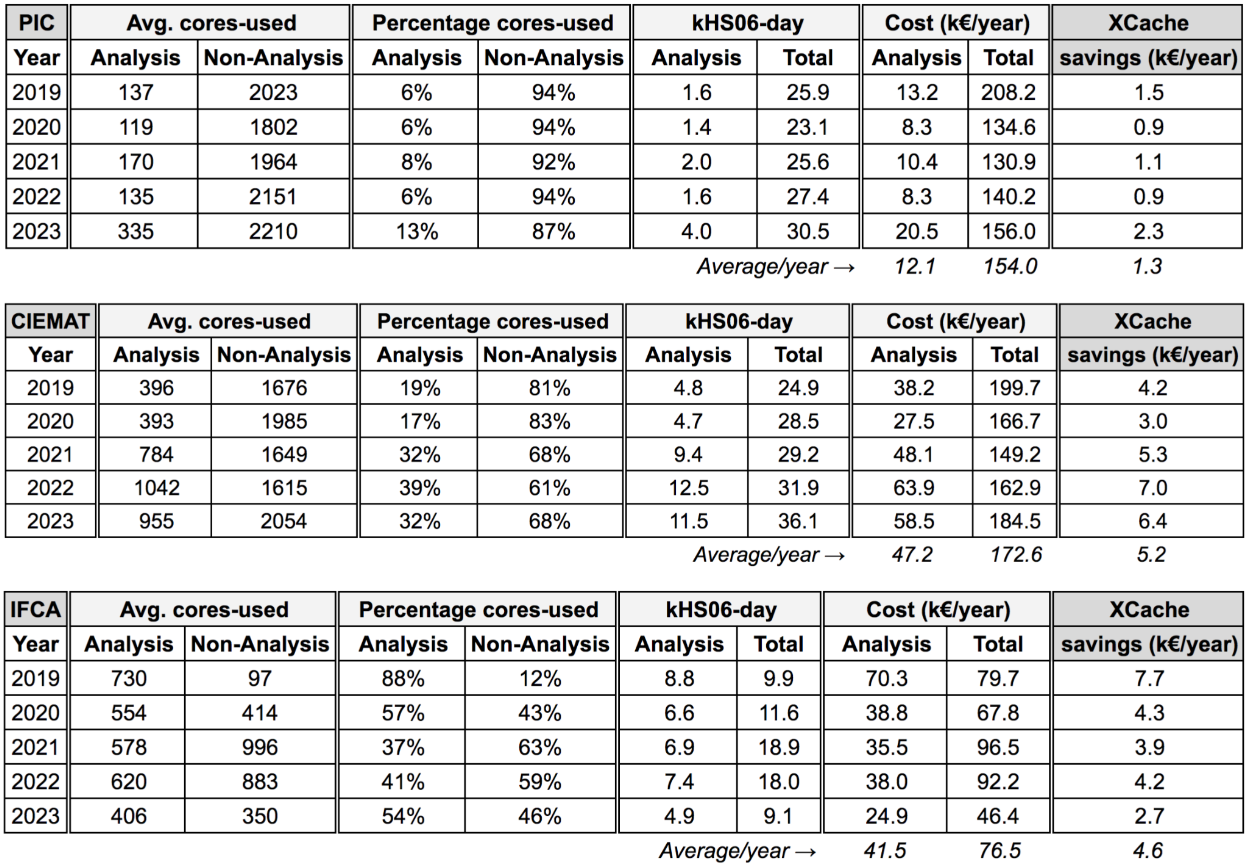
In order to assess the costs-benefits of deploying a XCache service in a region, the last 5 years historical information of CRAB jobs run at the Spanish sites is taken. The average number of CPU cores used by CRAB jobs can be translated into costs by applying the official PIC Pay-Per-Use (PPU) CPU metrics, that are shown in Table 9.1. Beyond that, Table 9.2 shows the CRAB jobs executed in PIC Tier-1, CIEMAT Tier-2 and IFCA Tier-2, and the associated costs for running these end-users activities, which have been estimated using the PPU metrics from PIC Tier-1. These values that can be considered very representative of the running costs for CPU resources in a site (including all of the overhead and personnel costs to operate the services).



Year	PPU (€ per HS06-day)
2019	0.02
2020	0.016
2021	0.014
2022	0.014
2023	0.014

Table 9.1: Pay-Per-Use (PPU) for CPU usage at PIC Tier-1. These values take into account the cost of consuming HS06-day at PIC, including hardware and operation costs .

The average cost per year for running these analysis activities is estimated to be 12k€, 47k€ and 42k€, for PIC, CIEMAT and IFCA, respectively, as quoted in Table 9.2. An average of 12 HS06/core of power has been applied to estimate these costs. Since the XCache usage would yield up to a ∼11% cost reduction for CPU deployment, this would imply a reduction of about 11k€ per year for the whole Spanish CMS sites when running Analysis tasks. As compared to the total cost spent in CPU per site, the XCache cost saving is at 1% level for PIC Tier-1, 3% for CIEMAT Tier-2 and 6% for IFCA Tier-2. The average amount of budget spent in CPU resources deployed in the Spanish Grid infrastructure for CMS is of the order of ∼350k€/year, so the amount of CPU cost savings by deploying an XCache service in the region would be around ∼3%. However, it should be noted that half of the CPU pledges from the Spanish CMS sites will be taken from the Barcelona SuperComputing Center (BSC) CPU resources from 2024 on. These resources are used for MC production, hence the Analysis activities will run only on the Grid sites. Hence, from the Grid resources deployment perspective, the use of the XCache in the region would have a cost benefit at the order of 6%.



PIC	Avg. cores-used		Percentage cores-used		kHS06-day		Cost (k€/year)		XCache
Year	Analysis	Non-Analysis	Analysis	Non-Analysis	Analysis	Total	Analysis	Total	savings (k€/year)
2019	137	2023	6%	94%	1.6	25.9	13.2	208.2	1.5
2020	119	1802	6%	94%	1.4	23.1	8.3	134.6	0.9
2021	170	1964	8%	92%	2.0	25.6	10.4	130.9	1.1
2022	135	2151	6%	94%	1.6	27.4	8.3	140.2	0.9
2023	335	2210	13%	87%	4.0	30.5	20.5	156.0	2.3
						Average/year →	*12.1*	*154.0*	*1.3*

CIEMAT	Avg. cores-used		Percentage cores-used		kHS06-day		Cost (k€/year)		XCache
Year	Analysis	Non-Analysis	Analysis	Non-Analysis	Analysis	Total	Analysis	Total	savings (k€/year)
2019	396	1676	19%	81%	4.8	24.9	38.2	199.7	4.2
2020	393	1985	17%	83%	4.7	28.5	27.5	166.7	3.0
2021	784	1649	32%	68%	9.4	29.2	48.1	149.2	5.3
2022	1042	1615	39%	61%	12.5	31.9	63.9	162.9	7.0
2023	955	2054	32%	68%	11.5	36.1	58.5	184.5	6.4
						Average/year →	*47.2*	*172.6*	*5.2*

IFCA	Avg. cores-used		Percentage cores-used		kHS06-day		Cost (k€/year)		XCache
Year	Analysis	Non-Analysis	Analysis	Non-Analysis	Analysis	Total	Analysis	Total	savings (k€/year)
2019	730	97	88%	12%	8.8	9.9	70.3	79.7	7.7
2020	554	414	57%	43%	6.6	11.6	38.8	67.8	4.3
2021	578	996	37%	63%	6.9	18.9	35.5	96.5	3.9
2022	620	883	41%	59%	7.4	18.0	38.0	92.2	4.2
2023	406	350	54%	46%	4.9	9.1	24.9	46.4	2.7
						Average/year →	*41.5*	*76.5*	*4.6*

Table 9.2: Summary of CRAB Job Execution and Associated Costs at Tier-1 and Tier-2 CMS Sites in Spain according to the PPU estimated metrics.

From the end-user experience, the fact that jobs end faster and that additional ∼10% of end-user jobs can be executed in the infrastructure (if the pledge is kept as it is today) would be much appreciated, considering that typically these jobs compete with production workloads. Recalling that the Spanish case seems to be a factor 2 lower than the global value seen when considering the fraction of CRAB jobs that read remote files. It might be that for some regions these costs-savings might be higher at about ~10% levels if using XCache data services. With respect to the storage resources, no in-depth study has been done in terms of cost savings, but the fact that only 20% of CRAB jobs executed in Spain read remote datasets (as compared to the 40% observed globally), might be an indication that the Physics Groups in Spain tend to copy to their local SE the datasets they are interested to process, hence there might be some room for improvement. The space devoted to local groups is typically 15% of the pledge at Tier-2s. In Spain, the average cost of deploying the local SE at the Tier-2s is around 130k€/year, so the expected savings if fully using a data cache to access datasets of interest from local groups would be at the order of ~20k€/year, a value that could slightly increase if storage is as well saved in the Tier1 site. In any case, it is worth mentioning that, since the XCache service is an uncatalogued and non-critical service, it can be deployed in old hardware which is typically retired from production after 5 years of operation time. If the XCache service fails at any time, the CRAB jobs reading remote data would use all the CMS XRootD infrastructure, so the system is deployed in a way that it is not disruptive, in case of service failure. From this perspective, any old server could be then configured and deployed to start caching new requested data at any time.

Chapter 10

Simulating a regional CDN for the CMS experiment in Spain

Upon discovering a method to access information about CRAB job access and data usage, as well as efficiently processing the substantial volume of data they accumulate, this information can be employed to realistically model the behavior of the data cache. The goal is to predict and evaluate the impact of different policies and configurations on cache performance metrics without the need to implement them in a production environment. For this purpose, several attempts of cache modeling have been performed trying to shed light on which are the best configurations and which are the relevant metrics for each deployment in CMS [223]. In addition, many studies regarding the optimization of cache systems have been published in the last decades, and some of the classic cache algorithms and policies include write everything, Least Recently Used (LRU), Least Frequently Used (LFU), etc. For example, a similar study conducted by INFN employing an AI approach to evaluate the best cache algorithms for a regional cache for CMS, showcased that QCACHE AI algorithm outperformed traditional caching algorithms. Specifically, this outperformance surpassed classic algorithms between 100TiB and 200TiB cache sizes [224]. Since these analyses started to gain popularity over the last years, they opened the door to discuss which are the proper cache configurations to deliver experimental data in regional CDNs. However, data sources employed by these studies in CMS come from cmssw-popularity, a database which lacks certain critical information in order to use it in simulations and estimations, such as the data location, which jobs use them and which amount of events they process. Even a study that was performed during this Thesis used the cmssw-popularity data [225], and it was incomplete.

Harnessing the Big Data capabilities provided by SWAN and the parallelization framework of Dask, we process logs from CRAB users who have run experiments in Spain. This enables us to generate the essential information needed to model and simulate various cache configurations, marking the first utilization of this data source for such purposes. Moreover, this realistic approach to caches will provide valuable insights about the suitability of classic caching algorithms for CMS regional XCaches and the proper cache sizes. The work presented in this Chapter has been submitted to the upcoming International Symposium of Grids and Clouds (ISGC 2024) conference [226].

10.1. Analysis of CRAB jobs executed in Spain

In order to simulate the effect of a cache serving data to the whole Spanish region, the CRAB jobs executed in the Spanish sites for 4 months have been analyzed. On average, about 9.5k jobs have been executed per day in PIC, CIEMAT and IFCA sites, with daily peaks up to ∼30k jobs. The majority of the jobs (50%) have been executed in CIEMAT, while 34% and 16% of the jobs have been completed in PIC and IFCA, respectively. Figure 10.1 shows the number of completed CRAB jobs per day in these sites, from June to October 2023.

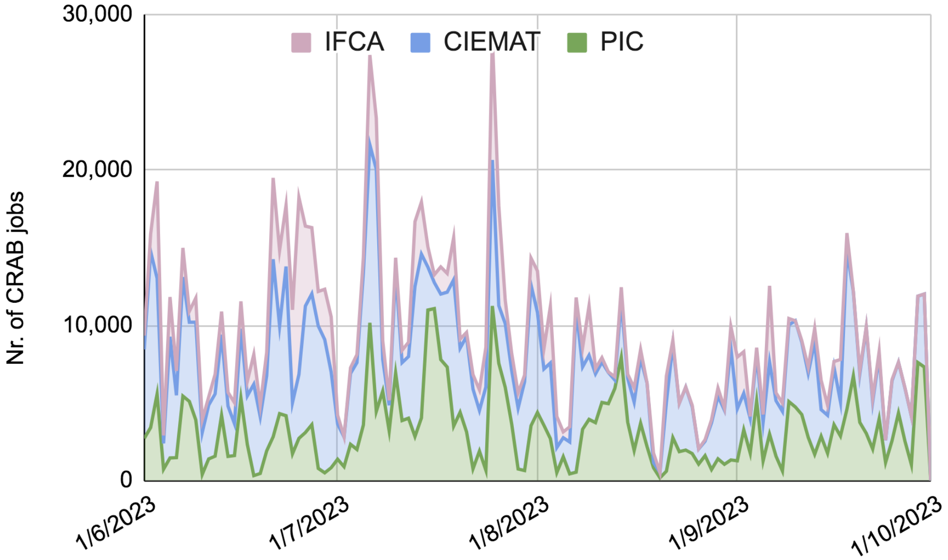


Figure 10.1: Number of CRAB jobs completed in PIC, CIEMAT and IFCA for the period from June to October 2023.

Each CRAB job is able to access more than one input file. By using the developed techniques explained in Chapter 9, the number of input files and their location have been obtained from each CRAB job log file. Figure 10.2 shows the number of accessed files per day for all of the executed CRAB jobs in the Spanish sites. On average, ∼25k files have been opened per day, which means that on average ∼2.7 input files have been opened for each of the CRAB jobs executed in the region in this period. Furthermore, Figure 10.3 shows the daily percentage of files opened from local SEs or from remote SEs for all of the Spanish sites in the selected period. For PIC, CIEMAT and IFCA, ∼22%, ∼33% and ∼77%, of the input files were obtained from remote SEs, respectively. A total of ∼3.1M files have been accessed in this period of time, with 1.1M files being accessed from remote locations.

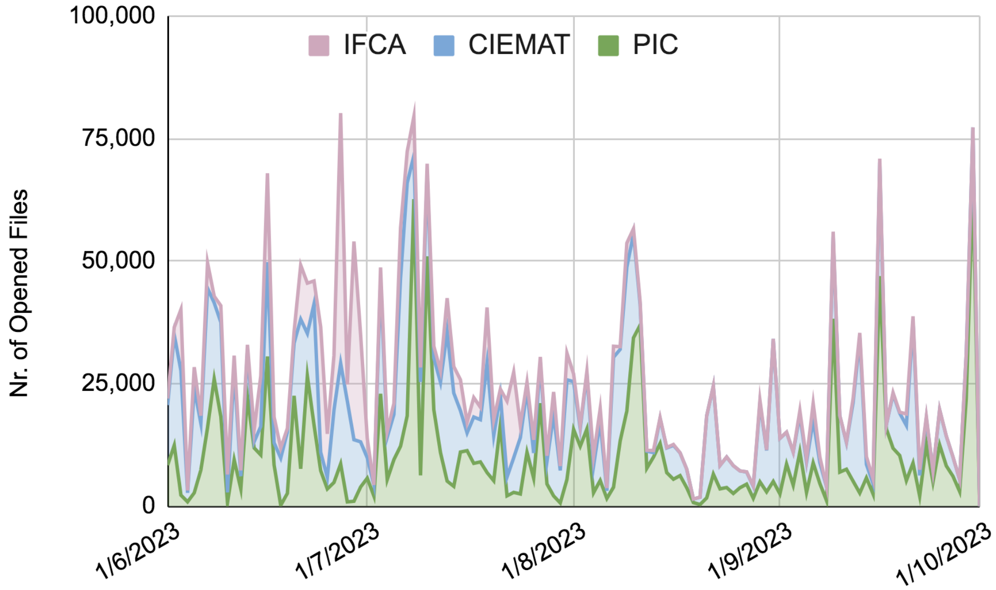


Figure 10.2: Number of opened files for the CRAB jobs completed in PIC, CIEMAT and IFCA in the period from June to October 2023.

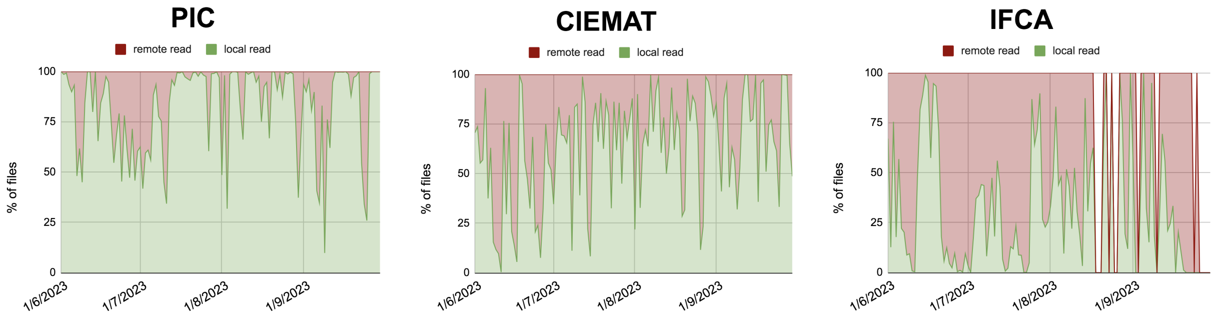


Figure 10.3: Percentage of input files that have been read from local SE (green) or from remote sites (red) in PIC, CIEMAT and IFCA for the period from June to October 2023.

CRAB jobs executed in the region also access a variety of input files. Aside from *users* or *Physics Groups* generated data, the main accessed files in this period were official CMS DATA and MC files, with a total of 1.9M and 950k files, respectively. Figure 10.4 shows the breakdown of input files by data tier, for both DATA and MC, and for all of the considered sites. Overall, the most significant DATA files accessed were of type MINIAOD (∼30%), RAW (∼20%) and AOD (∼16%). The most significant MC accesses were of type

GEN-SIM* (16%) and MINIAODSIM (~11%). It is worth noting that the access to DATA files from the Tier-2 sites show a similar trend.

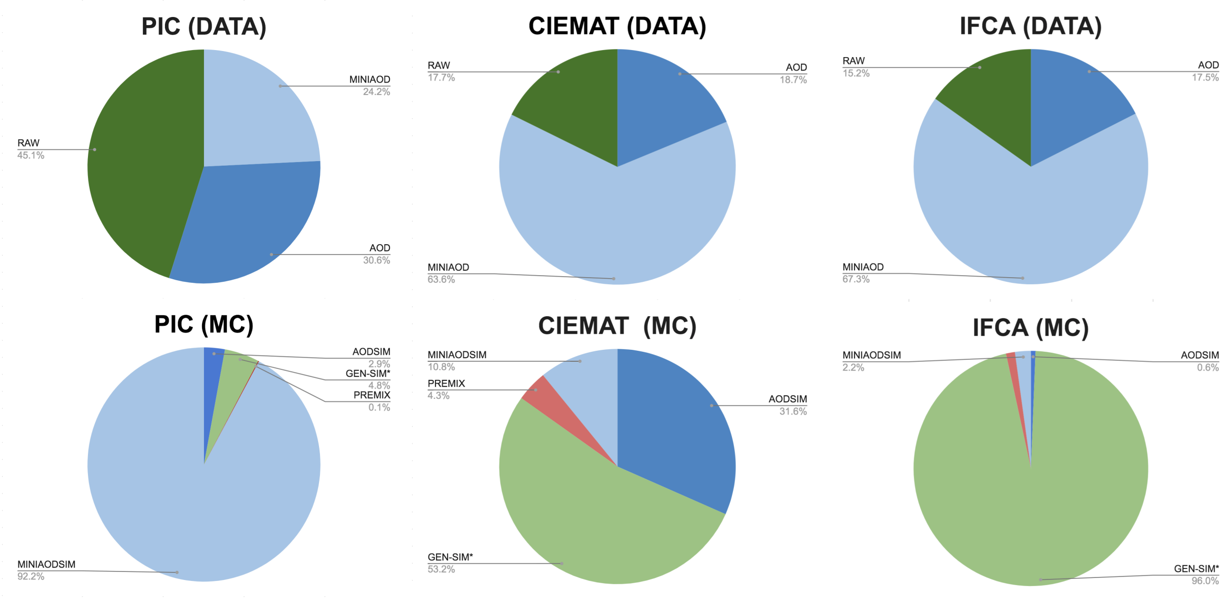


Figure 10.4: Breakdown of total (local and remote) input files by data tier, for both DATA and MC, and for all of the considered sites in the period from June to October 2023.

10.2. Completeness of input file reads

The information shown in Section 10.1 would be used to simulate a data cache behavior in the region for the analysis jobs that are executed in the regional compute nodes. The XCache deployed in PIC has enabled a pre-fetching mechanism, so files are not completely downloaded to the cache if the applications do not need to read the files completely, using read-ahead techniques. Chunks of 10 blocks of 50 Kb are fetched when there are read requests, and typically not all of the files in the cache are complete. This is important to know, if a realistic simulation of a data cache is being performed. One has to know if the input files are completely cached, or not.

Snapshots of the deployed XCache at PIC Tier-1 are regularly inspected, and it has been learnt that typically most of the input files are downloaded complete, with the exception of DATA RAW and MC GEN-SIM files. Many CRAB jobs, if not all, access to AOD* files to perform analysis, but they do also access DATA RAW and MC GEN-SIM files to get additional information, in the same job execution. It has been observed that these files that are opened at execution time do not always need to be read completely. Sometimes, these files are opened and closed, even without reading any byte.

Figure 10.5 shows the fraction of the DATA RAW and MC GEN-SIM files downloaded to the XCache at PIC. The observation is that DATA RAW files have a spread on the percentage of file read, regardless of the number of accesses the file has. The MC GEN-SIM shows a similar trend, but only on first access. The subsequent data file accesses typically read the complete file. Since these two types of files are frequently accessed (~ 35% of the accesses go to DATA RAW and MC GEN-SIM), it is important to include this read bytes behavior when simulating an XCache service for the region.

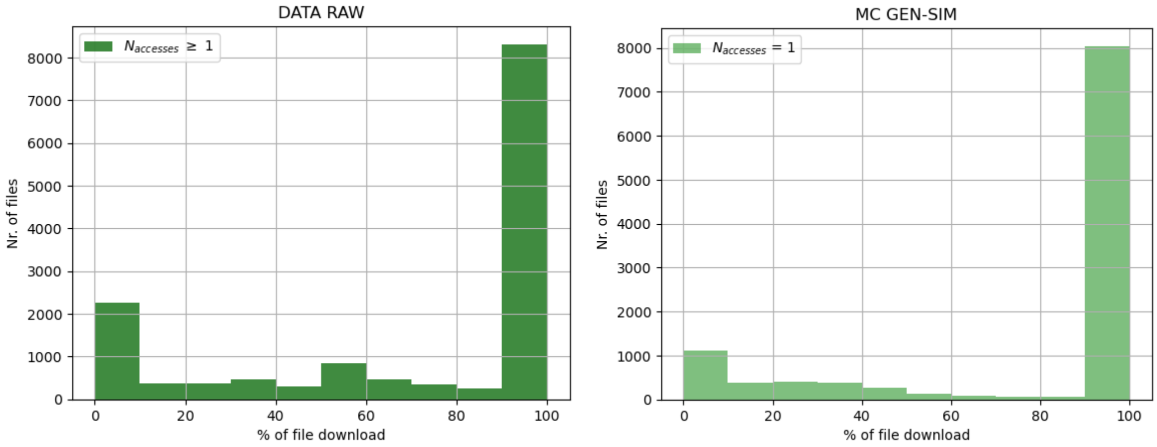


Figure 10.5: Fraction of the DATA RAW and MC GEN-SIM files downloaded to the XCache at PIC, from a snapshot of the XCache content taken on 1st October 2023.

10.3. Caching algorithms in CDNs

Caching algorithms are introduced to improve network and content delivery efficiency of a cache system [227]. A caching algorithm consists in a set of rules to decide which data to store in a cache and which items to delete from the cache when it needs to make room for new data. The rules to store popular data are typically performed on similar analysis as performed in this Thesis, from users' access patterns to data. With respect to data deletions from the cache, several mechanisms have been explored, the most popular being the Least Recently Used (LRU) algorithm. The LRU works by evicting the least recently used data item from the cache when a new data item needs to be added. LRU is a simple and effective algorithm, but it can be less effective for caching data items that have different access frequencies.

Another common deletion mechanism is the Least Frequently Used (LFU) algorithm. LFU works by evicting the least frequently used data item from the cache when a new data item needs to be added. LFU is more effective than LRU for caching data items that have different access frequencies, but it can be less effective for caching data items that have different sizes. In addition to classic caching algorithms like LRU and LFU, CDNs employ a variety of other data retention strategies to optimize performance. Other notable algorithms include Least Frequently Recently Used (LFRU) and Least Recently Used Working Set (LRU-WS). LFRU is a hybrid approach that divides the cache into privileged and unprivileged partitions to accommodate data with different access frequencies. LRU-WS considers an application's working set to evict the least recently used data outside that set, making it ideal for applications with well-defined working sets.

The XCache service employs the LRU algorithm as the default policy for cached data replacement. Deletion is triggered by watermarks representing specific occupancy thresholds. When the occupancy exceeds the High-Watermark (HW) of 95%, the algorithm initiates file deletion until reaching the lowest occupancy range, the Low-Watermark (LW) of 90%. Given this characteristic, the simulations conducted in this Chapter will adhere to the LRU algorithm, with the same watermarks as used in production, aligning with the behavior of the physical XCache that the computations seek to replicate.

10.4. Simulating an XCache for the Spanish CMS Tiers

All of the ingredients to simulate an XCache behavior for the Spanish CMS Tiers have been identified and addressed: all of the remote input files that are accessed from CRAB jobs executed in the region, the level of file reads completion based on the data type, and the data retention algorithm that mimics the XCache (LRU with 95%-90% watermarks). Figure 10.6 shows the XCache data cache population for all of the remote requests from CRAB jobs executed in PIC, CIEMAT and IFCA for the period from June to October 2023.

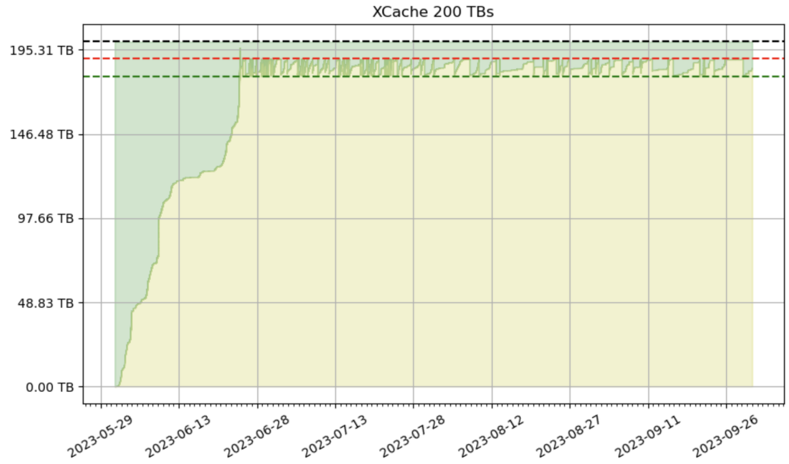


Figure 10.6: Simulation of a 200 TB XCache that caches all of the remote reads from CRAB jobs executed in PIC, CIEMAT and IFCA, in the period from June to October 2023.

This example XCache is simulated with a size of 200 TBs (black dashed line), and the cached data deletions are handled by an LRU with 95%-90% watermarks, shown in the figure with red and green lines, respectively. It takes less than 1 month to saturate the XCache disk, as Figure 10.7 shows the number of files created and deleted per day in the simulated XCache.

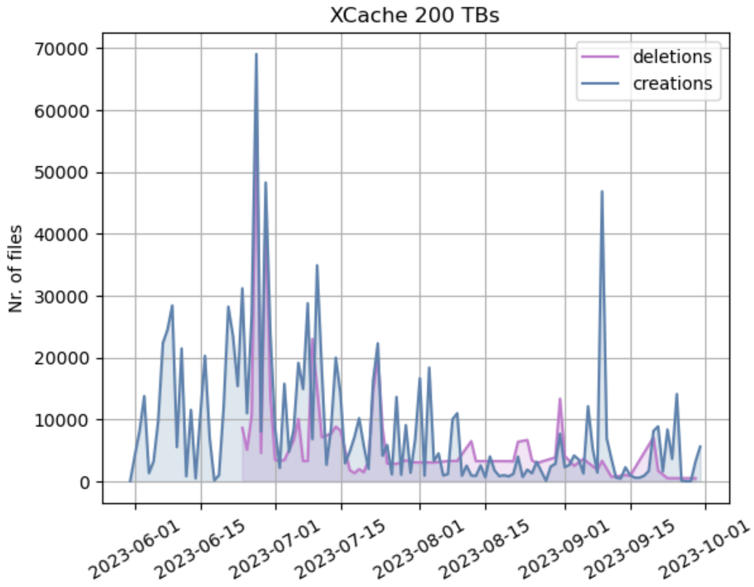


Figure 10.7: Number of files created and deleted for the simulated 200 TB XCache example.

An important aspect of a data cache, aside from its size, is the data import and export, since it conditions the network connectivity the cache server should have available. Figure 10.8 shows the data import and exports for the simulated XCache of 200 TB to illustrate it.

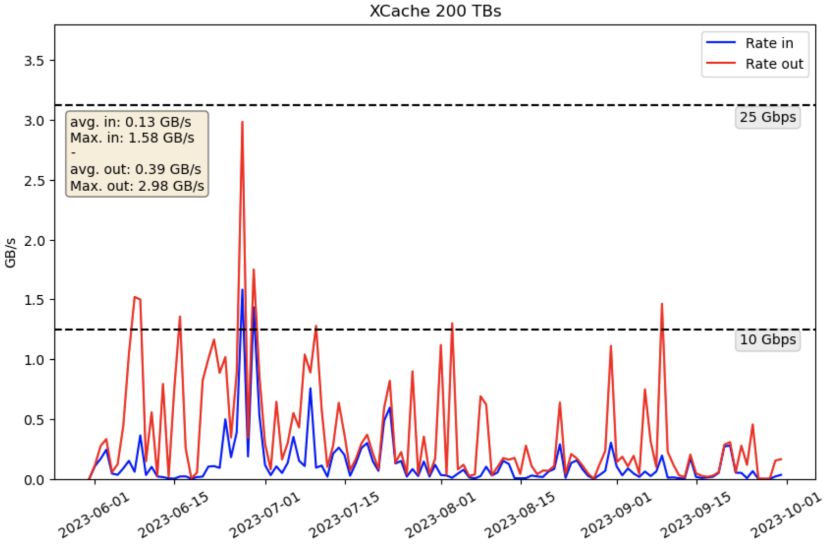


Figure 10.8: Data export and import for the simulated 200 TB XCache example.

On the other hand, a well sized cache should contain the most frequently accessed data files, while keeping the minimum number of non-accessed data files. Even if the LRU algorithm is in charge of deleting data, a minimal sized deployed cache would be inefficient, since it would not keep the popular files enough. Consequently, the LRU would trigger data cache repopulation with popular files that have been previously stored in the cache, and that would need to be re-cached again. A very large data cache would suffer from holding old unaccessed data which is not totally flushed from the LRU data deletion cycles. Characterizing the XCache is important in order to set the most optimal size to be deployed. The *Hit Rate* measure can be introduced as the number of cache hits (i.e. the number of accesses to files that were present in the cache) over the total number of accesses (i.e. number of cache hits and cache misses). A cache miss is a file that is not present in the cache and needs to be cached. The *Hit Rate* is then calculated as:

$$Hit\ Rate = \frac{hits}{hits + misses} = \frac{hits}{N_{accesses}} \tag{6}$$

It can be expressed in percentage, and it can be calculated in a cumulative way, since the data cache starts being populated. Figure 10.9 shows the percentage of cache hits, cache misses and cumulative *Hit Rate* for the simulated 200 TB XCache.

10.5. Optimal XCache size for the Spanish CMS Tiers

Based on the cumulative *Hit Rate* and network aspects, one can simulate several sized data caches to find the most optimal and performant data cache that should be deployed to serve the region, both on total size and network connectivity. The procedure is the same as described in the previous section, but varying the XCache size to evaluate the relevant metrics and then identify the most optimal working point. It is worth noting again that these simulations are based on remote input file accesses from real jobs executed in the Spanish region for a 5 month period.

Figure 10.10 shows the cumulative *Hit Rate* (%) and the percentage of *Hit Rate* gain or benefit when transitioning to a bigger data cache (a line is drawn at 1% level, for reference). The cache saturates at around 61.2% in cumulative *Hit Rate* for a data cache size of 400 TBs. The relative gains in the cumulative *Hit Rate* go below 1% levels for data caches above ~100 TBs. Deploying a data cache at about 200 TBs seems to be sufficient to serve the region with optimal performance from the cumulative hit rate point of view.

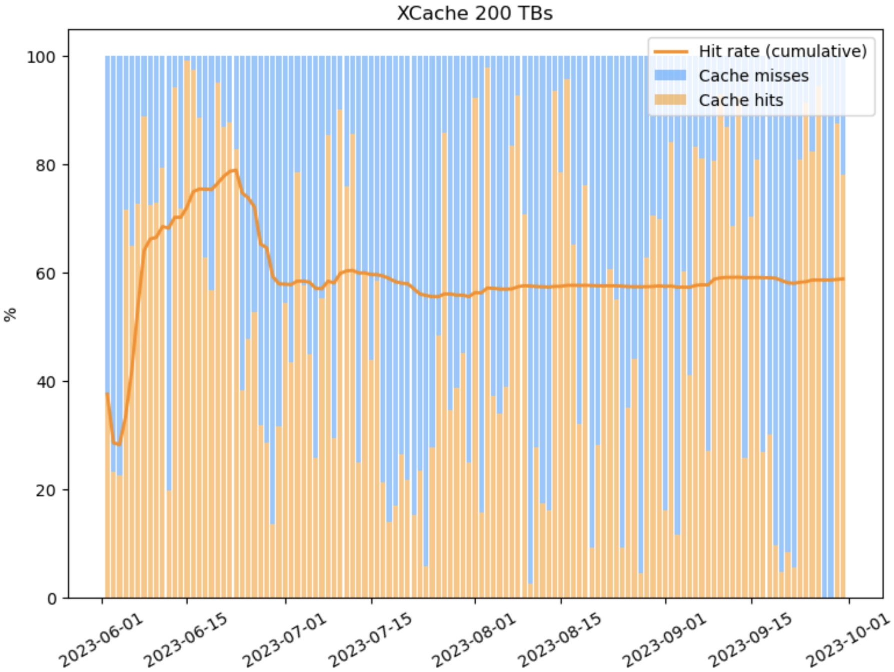


Figure 10.9.: Percentage of cache hits, cache misses and the cumulative Hit Rate for the simulated 200 TB XCache example.

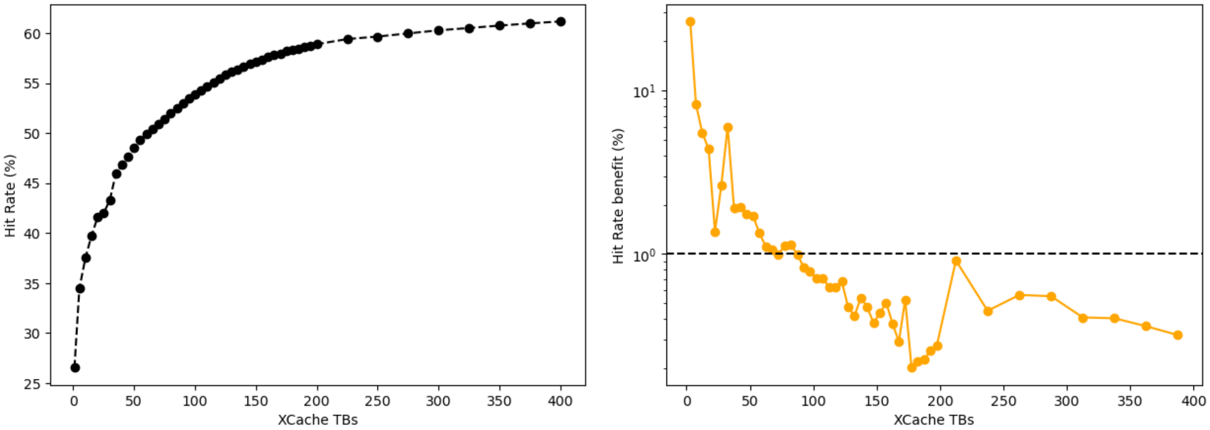


Figure 10.10.: Cumulative hit rate (%) [left] and the percentage of hit rate gain or benefit when transitioning to a bigger data cache (a line is drawn at 1% level, for reference) [right].

Figure 10.11 shows the network performance metrics for the simulated caches. For all of the simulated data caches, a disk server equipped with 25 Gbps NIC is necessary to satisfy the daily peaks observed both in data imports and data exports (indeed, a 100 Gbps NIC would provide a bit more headroom, since the values shown are daily averages and the peaks during the day might be higher than these estimated values). For a data cache bigger than 200 TBs, the cache would serve in average three times more data to the regional compute nodes than it gets from remote CMS sites, as seen in Figure 10.12.

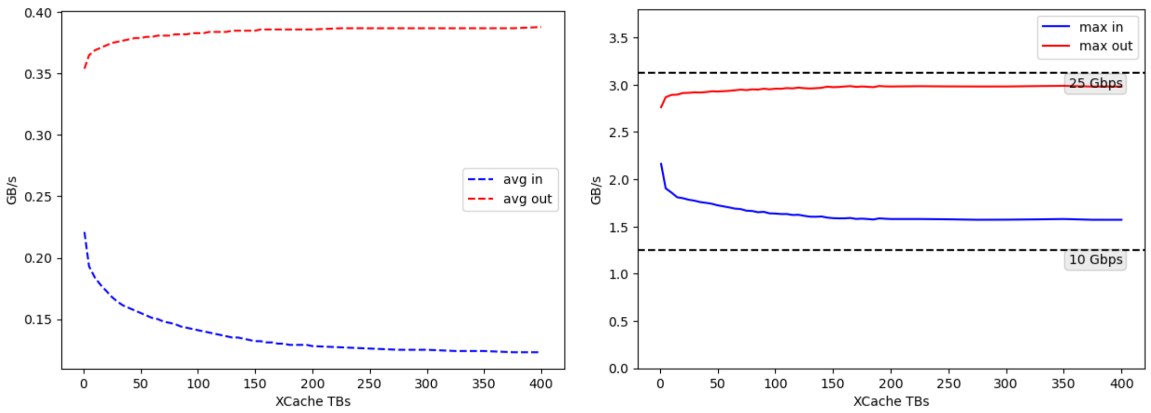


Figure 10.11.: Network performance metrics for the simulated caches. Average in and out data rates [left] and maximum in and out data rates [right].

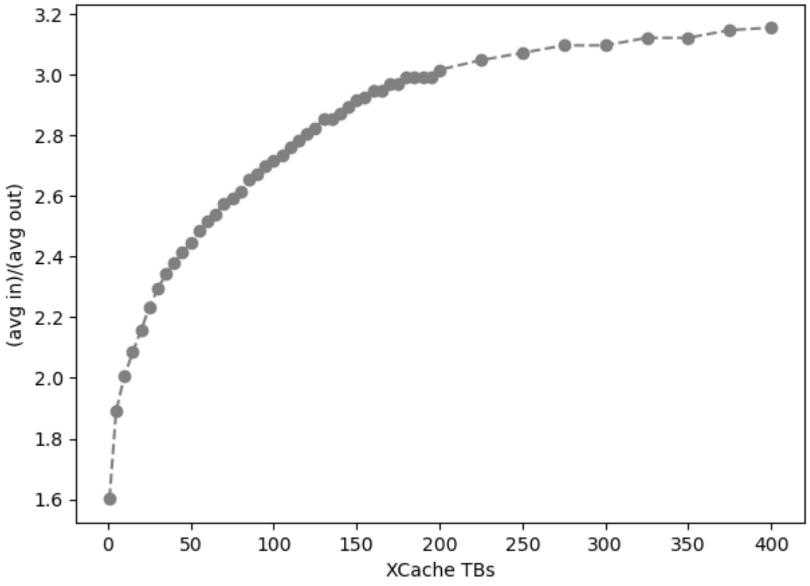


Figure 10.12.: The ratio of average in to average out data rates, as a function of simulated XCache size.

Chapter 11

Conclusions

The research conducted on this Thesis primarily focused on overcoming the challenges associated with managing the vast volumes of data generated by the CMS experiment at the LHC. One of the most prominent issues was the escalating demands for storage with constrained budgets within the Physics community. Different strategies were set in order to optimize data access and utilize available resources efficiently, especially in preparation for the future HL-LHC period. This Thesis aimed to demonstrate that deploying a CDN to deliver data for the CMS experiment through a centralized XCache system in Barcelona at the PIC Tier-1, serving data to all three CMS sites in Spain, enhances end-users job efficiencies and might contribute to cost savings in the region.

During this Thesis, the potential of cache systems has been assessed, particularly those leveraging XCache, when strategically positioned near CMS WLCG sites in Barcelona, at PIC Tier-1, serving CIEMAT Tier-2 in Madrid and, potentially IFCA Tier-2 in Cantabria. This approach involves creating an intermediary caching layer that brings frequently accessed data closer to compute nodes. The main objective has been to establish an enhanced data delivery system that, eventually, optimizes data management and reduces associated costs. Additionally, the practical implementation of the XCache in the region has allowed PIC and CIEMAT sites to efficiently explore and understand how the service handles the data and operates in interaction with the real end-users' CMS jobs.

Initially, an in-depth analysis of data usage patterns was conducted at these centers, offering valuable insights into which experimental CMS data tiers would benefit the most from innovative caching strategies in the region. The investigation of data access and patterns within PIC and CIEMAT storage systems, as well as from the jobs run on the PIC and CIEMAT compute nodes, revealed that CMS analysis jobs would gain the greatest advantages from these caching strategies. This was primarily due to the frequent re-accessibility of data widely utilized by analysis jobs, particularly for the AOD-derived data formats.

Consequently, after having identified the jobs with the most potential for benefiting from XCache, additional studies were carried out to evaluate the advantages of XCache in terms of latency and cost savings for executing CMS analysis jobs. The studies were conducted in a controlled environment using actual CMS jobs, referred to as "XCache benchmark jobs". These tests submitted at PIC while accessing data remotely from several CMS sites worldwide demonstrated that, during the trial period, accessing data locally estimated that there can be a potential saving of 1.8k HS06·hours, which is equivalent to 28% of the total walltime spent in the test. This result highlights the importance of latency hiding and data placement in improving the efficiency for the CMS analysis tasks. It shows that accessing data locally can significantly enhance the overall performance of CMS Analysis jobs.

Following the benchmark studies for XCache, CMS CRAB Analysis logs were analyzed to demonstrate that efficiency enhancement is also experienced by end-users in production. For that purpose, Big Data and parallelization techniques were employed to analyze several TBs of data collected from these logs for all of the jobs executed in PIC and CIEMAT while accessing remote data or doing it locally. By enabling access to PIC XCache for all CIEMAT compute nodes, there could have been approximately 13% savings in the total HS06·hours spent by CRAB jobs at CIEMAT. This indicated a potential for performance improvement and resource savings in the region through the utilization of the XCache service. In terms of delivered work, it means that either CIEMAT can perform 11% more computational work or deploy 11% less CPU resources to perform the same work as previously done without using the XCache service.

Another aspect delved into using this Big Data and parallelization techniques is the estimation of the XRootD traffic generated by CRAB jobs that read remote files. Currently, this traffic is not yet accounted for in any of the existing CMS monitoring tools. The techniques employed in user logs are also permitted to apply the same methodology by doing the same with the whole CMS sites within a month. The estimations showed that the actual amount of traffic imposed by CRAB jobs on the global CMS network might be higher than the estimated value of 10 GB/s. In other words, the traffic generated by CRAB jobs could be the sum of the traffic generated by FTS transfers and the traffic generated by the reading of PREMIX libraries, about an additional 30% with respect to the expected values. The realistic estimations of this traffic, along with the precise measurement of the total remote data accessed by XRootD during a month by the CMS sites, showcase that more extensive deployments similar to the regional in Spain could be beneficial for the global performance of CRAB jobs.

Another relevant outcome from these studies is that the average cost per year for running analysis activities is estimated to be 12k€, 47k€, and 42k€ for PIC, CIEMAT, and IFCA, respectively. The usage of XCache would result in a cost reduction of approximately 11% for CPU deployment, leading to a reduction of about 11k€ per year for the Spanish CMS sites when running analysis tasks. On the other hand, the cost savings with XCache usage of these studies concluded in the order of 1% for PIC Tier-1, 3% for CIEMAT Tier-2, and 6% for IFCA Tier-2. These estimations translate in approximately 3% of the CPU cost savings by deploying an XCache service in the region. From the perspective of Grid resources deployment, the use of XCache in the region would have a cost benefit at the order of 6%. On the other hand, enhancing the execution time of analysis jobs, as demonstrated by the 10% improvement measured in this Thesis, would also lead to a more favorable end-user experience.

The Spanish case shows that the cost savings from using XCache data services elsewhere in the WLCG sites used by CMS might be higher at around 10%. In terms of storage resources, only 20% of CRAB jobs in Spain read remote datasets, indicating that there is potential for improvement. The average cost of deploying the local SE at Tier-2s in Spain is around 130k€/year, concluding that the expected savings from fully utilizing a data cache would be around 20k€/year. The XCache service can be deployed on old hardware and if it failed, the CRAB jobs would use the CMS XRootD infrastructure without disruption.

The results of LRU cache simulations using real CMS data are presented with the purpose of evaluating and fine-tuning aspects such as cache sizes and configurations of XCache, without the need for production environment testing, and including the missing site that was not tested with the XCache production service (namely IFCA Tier-2). These simulated computations provided crucial insights into the performance and efficiency of XCache in a controlled environment, using real CMS data. All the necessary components for simulating the behavior of XCache in the Spanish CMS Tiers have been identified and addressed. These include considering remote input files accessed from CRAB jobs in the region, file read completion levels based on data type, and the data retention algorithm, being the LRU with 95%-90% watermarks in correspondence with XCache default configuration. The results of the simulations have helped to identify potential bottlenecks and determine the necessary network configurations. In this case, it is recommended to have a 25 Gbps NIC installed in a single XCache service serving the entire Spanish region to avoid the excess of traffic in a regular XCache service for a single node. Finally, based on the cumulative hit rate and network aspects, the most optimal and performant data cache size to serve the Spanish CMS Tiers has been determined. Varying the XCache size, the relevant metrics can be evaluated to identify the optimal working point, in particular the hit-rate, which measures the probability of having the desired data in the cache. In this case, the results showed how the cache saturates at around 61.2% in cumulative hit-rate for a data cache size of 400 TBs, showing in the simulation that deploying a data cache of about 200 TBs is sufficient for optimal performance from the cumulative hit rate perspective. In terms of network, for all simulated data caches, a disk server equipped with a 25 Gbps NIC is necessary to handle the daily peaks observed in data imports and exports. In conclusion, cache larger than 200 TBs would serve, on average, three times more data to the regional compute nodes than it receives from remote CMS sites, showcasing to be an ideal size for the use-case in Spain for the CMS experiment.

11.1. Future work

Overall, the changes implemented in this Thesis have a positive impact and are currently being deployed in production in Spain. These changes have also been evaluated at scale and will be further improved during the LHC Run3. Additionally, the Thesis demonstrates the potential benefits of deploying a CDN and centralized cache system in Barcelona at the PIC for the CMS experiment in the Spanish region by enhancing job efficiency, contributing to cost savings and reducing workload completion times. Hence, the knowledge gained from this research will also have broader implications for other regions and data-intensive scientific endeavors. In particular, the service deployment has been selected as a strategic project from the RES perspective, and funds to deploy this service and perform R&D activities has been granted.

Despite the satisfactory study and deployment of XCache to serve CMS data for a Tier-1 and Tier-2 facility in Spain, several tasks remain to be carried out and aspects to explore in order to improve and extend the service, not only within Spain, but also for the entire CMS community.

The positive impact of placing an XCache for both PIC and CIEMAT on analysis jobs has been demonstrated, but the current XCache service does not exclusively serve data to these types of jobs. Therefore, future work should focus on evaluating whether there is an improvement in other job types when accessing data through XCache and how to fine tune its production performance.

On the other hand, benchmark job studies were conducted from the compute nodes at PIC, accessing data from various CMS sites worldwide. These tests objectively measured the benefits in terms of walltime, efficiency, and latency gained by adding the XCache service in PIC. This study could be extended to various CMS sites in the same way they were conducted at PIC to objectively assess the savings achieved by adding XCache for each of the regions.

Furthermore, the deployment of the cache in Spain has shown similar benefits in production as indicated by the benchmark jobs over CRAB jobs. While the real-world scenario outside a controlled environment may have higher error accumulation and may not yield as significant improvement values, the benefits are still observable. Therefore, if XCache deployment is expanded to more sites worldwide, the real impact on production and the enhancement of CRAB jobs could be determined. It is essential to highlight that accessing and processing this data involves an intensive process, requiring various Big Data techniques and parallelization methods that may not be easily transferable to other users and administrators at different sites. The necessity of incorporating this information through monitoring services of CMS is crucial in order to assess the benefits of the models explored during this Thesis and, therefore, the results obtained could help for future inclusion of this information into production.

Moreover, cache simulations have the potential to go beyond the models presented in this work, specifically: determining the best replacement algorithm, whether through Machine Learning techniques or the identification of new metrics that more characterize the cache performance more accurately for the use case. Metrics such as hit rate and those coming from data popularity have proven to be suitable indicators for cache performance, with the LRU algorithm demonstrating favorable results in both production and simulation. Nevertheless, the integration of Machine Learning techniques, such as Random Forests and Neural Networks, seem promising in the selection of relevant features from among the various potential metrics to potentially better fit the specific use-case. This approach can aid in identifying data replacement algorithms that align better with the access patterns of CMS data in the Spanish region, potentially extending this methodology to areas with CMS sites facing similar conditions as those explored in this study for Spain.

Regarding configuration, future work should explore the possibility of studying the production inclusion of multi-node XCache deployments to distribute data load across multiple servers, thus avoiding bottlenecks. Additionally, future studies should consider the impact of caches on the upcoming CMS Analysis Facility being deployed at CIEMAT, which aims to provide data and computing resources to CIEMAT's analysis users. An analysis facility is a concept rapidly evolving to include dedicated hardware infrastructures and computing services optimized for the effective analysis of CMS data samples. This concept aims to replace the existing final data analysis model, which relies on data reduction within the Grid infrastructure and subsequent interactive analysis of manageable-sized samples on individual computers of physicists. Recent results exploring the diverse features of an Analysis facility at CIEMAT, showed that reading from XCache or SEs yielded similar results in most of the Analysis facilities without bottlenecks in the networks; even faster data access through XCache deployed in SSDs [228]. The study to further enhance XCache performance in this data delivery context and the evaluation of the feasibility of running computational resources without relying on persistent storage are tasks that CMS scientists in Spain plan to address in the future.

Keeping faith that a modest contribution to the understanding of using data caches for the CMS experiment in WLCG has been made, and hoping that at least this effort will motivate further and more complete studies, this Thesis ends here.

References

[1] Glashow, S. (1959). The renormalizability of vector meson interactions. Nucl. Phys. 10, 107.

[2] Salam, A., Ward, J.C. Weak and electromagnetic interactions. Nuovo Cim 11, 568–577 (1959). https://doi.org/10.1007/BF02726525

[3] Weinberg, S. (1967). A model of leptons. Physical Review Letters, 19(21), 1264-1266. https://doi.org/10.1103/PhysRevLett.19.1264

[4] T. Kajita et al., "Evidence for Oscillation of Atmospheric Neutrinos," Physical Review Letters, 1998. (DOI: 10.1103/PhysRevLett.81.1562)

[5] A. B. McDonald et al., "Measurement of the Rate of νe + d → p + p + e− Interactions Produced by 8B Solar Neutrinos at the Sudbury Neutrino Observatory," Physical Review Letters, 2002. (DOI: 10.1103/PhysRevLett.89.011301)

[6] Higgs, P. W. (1964). Broken symmetries, massless particles, and gauge fields. Physical Review Letters, 13(16), 508-509. doi:10.1103/PhysRevLett.13.508

[7] "The Large Hadron Collider," CERN. https://home.cern/science/accelerators/large-hadron-collider (accessed Jun. 30, 2023).

[8] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," Physics Letters B, vol. 716, no. 1, pp. 1–29, Jun. 2012, doi: 10.1016/j.physletb.2012.08.020.

[9] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," Physics Letters B, vol. 716, no. 1, pp. 30–61, doi: 10.1016/j.physletb.2012.08.021.

[10] A. G. Ruggiero and A. Zichichi, "Hadron Colliders at the Highest Energy and Luminosity," in Hadron Colliders at The Highest Energy and Luminosity, Apr. 1998. Accessed: Jul. 04, 2023. [Online]. Available: http://dx.doi.org/10.1142/9789814528931

[11] "Performance of the ALICE experiment at the CERN LHC," International Journal of Modern Physics A, vol. 29, no. 24, p. 1430044, Sep. 2014, doi: 10.1142/s0217751x14300440.

[12] ATLAS. WORLD SCIENTIFIC, 2018. Accessed: Jun. 30, 2023. [Online]. Available: http://dx.doi.org/10.1142/11030

[13] G. Petrucciani, "The CMS experiment at the CERN LHC," in The Search for the Higgs Boson at CMS, Pisa: Scuola Normale Superiore, 2013, pp. 15–58. Accessed: Jun. 30, 2023. [Online]. Available: http://dx.doi.org/10.1007/978-88-7642-482-3_2

[14] R. Quagliani, "The LHCb Detector at the LHC," in Springer Theses, Cham: Springer International Publishing, 2018, pp. 29–65. Accessed: Jun. 30, 2023. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-01839-9_2.

[15] CMS Collaboration. "WorkBook CMS Software Framework." [Online]. Available: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSSWFramework, accessed 1 November 2023.

[16] "Welcome to the Worldwide LHC Computing Grid," WLCG. https://wlcg.web.cern.ch/ (accessed Jul. 03, 2023).

[17] M. Adelholz et al., "Models of Network Analysis at Regional Centres for LHC experiments." [Online]. Available: https://monarc.web.cern.ch/MONARC/docs/phase2report/Phase2Report.pdf

[18] I. Foster, I. T. Foster, and C. Kesselman, "The Grid: Blueprint for a New Computing Infrastructure." Morgan Kaufmann, 1998.

[19] P. Charpentier, "LHC Computing: past, present and future," EPJ Web Conf., vol. 214, p. 09009, 2019. DOI: 10.1051/epjconf/201921409009.

[20] European Commission. "European Union." [Online]. Available: https://european-union.europa.eu/, accessed 1 November 2023.

[21] National Science Foundation (NSF). "National Science Foundation." [Online]. Available: https://www.nsf.gov/, accessed 1 November 2023.

[22] E. Martelli and S. Stancu, "LHCOPN and LHCONE: Status and Future Evolution," Journal of Physics: Conference Series, vol. 664, no. 5, p. 052025, Dec. 2015, doi: 10.1088/1742-6596/664/5/052025.

[23] Port d'Informació Científica (PIC). (2023, October 30). "PIC - Port d'Informació Científica." [Online]. Available at: https://www.pic.es/ (Accessed October 30, 2023).

[24] IFIC. "webific.ific.uv.es." [Online]. Available: https://webific.ific.uv.es/web/, accessed 1 November 2023.

[25] Universidad Autónoma de Madrid. "Tier2 Services." [Online]. Available: https://uam.es/FisicaTeorica/Tier2-services/1242686057197.htm?language=en&nodepath=Services, accessed 1 November 2023.

[26] Institut de Física d'Altes Energies (IFAE). (2023, October 30). "IFAE - Institut de Física d'Altes Energies." [Online]. Available at: https://www.ifae.es/ (Accessed October 30, 2023).

[27] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT). (2023, October 30). "CIEMAT - Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas." [Online]. Available at: https://www.ciemat.es/ (Accessed October 30, 2023).

[28] Instituto de Física de Cantabria (IFCA). (2023, October 30). "IFCA - Instituto de Física de Cantabria." [Online]. Available at: https://ifca.unican.es/ (Accessed October 30, 2023).

[29] Universidade de Santiago de Compostela. "Universidade de Santiago de Compostela." [Online]. Available: https://www.usc.gal/, accessed 1 November 2023.

[30] Universitat de Barcelona. "Universitat de Barcelona." [Online]. Available: https://web.ub.edu/es/, accessed 1 November 2023.

[31] CERN. "CERN Virtual Machine File System (CernVM-FS)." Accessed 2023-10-30. [Online]. Available: https://cernvm.cern.ch/fs/.

[32] A. Chervenak, R. Schuler, M. Ripeanu, M. Amer, S. Bharathi, I. Foster, A. Iamnitchi, and C. Kesselman, "The Globus Replica Location Service: Design and Experience," Parallel and Distributed Systems, IEEE Transactions on, 20, 1260 - 1272 (2009). 10.1109/TPDS.2008.151.

[33] A. A. Ayllon et al, "FTS3: New Data Movement Service For WLCG," J. Phys.: Conf. Ser. 513 032081 (2014).

[34] M. Ellert, A. Konstantinov, B. Kónya, O. Smirnova, A. Wäänänen, "The NorduGrid project: using Globus toolkit for building GRID infrastructure," Nuclear Instruments and Methods in Physics Research A, 502 (2–3), 407–410 (2003). doi:10.1016/S0168-9002(03)00453-4.

[35] D. Thain, T. Tannenbaum, and M. Livny, "Distributed Computing in Practice: The Condor Experience," Concurrency and Computation: Practice and Experience, Vol. 17, No. 2-4, pages 323-356, February-April, 2005, doi:10.1002/cpe.938.

[36] A. Tsaregorodtsev et al., "DIRAC: a community grid solution. Journal of Physics: Conference Series," DOI: 119. 062048 (2008). 10.1088/1742-6596/119/6/062048.

[37] Belle II Collaboration. "Belle II." [Online]. Available: https://www.belle2.org/, accessed 1 November 2023.

[38] International Linear Collider Collaboration. "International Linear Collider." [Online]. Available: https://linearcollider.org/, accessed 1 November 2023.

[39] [103] Consortium, The & Actis, M & Agnetta, G. & Aharonian, F. & Akhperjanian, A. & Aleksic, J. & Aliu, Ebenezer & Allani, D. & Allekotte, I. & Antico, Federico & Antonelli, Lucio Angelo & Antoranz, Pedro & Aravantinos, A. & Arlen, Timothy & Arnaldi, L.. (2011). "Design concepts for the Cherenkov Telescope Array CTA: an advanced facility for ground-based high-energy gamma-ray astronomy." Experimental Astronomy. 32. 193–316.

[40] M. Barisits et al., "Rucio: Scientific Data Management." Comput Softw Big Sci 3, 11 (2019). https://doi.org/10.1007/s41781-019-0026-3.

[41] ITU, "X.509: Information technology - Open Systems Interconnection - The Directory: Public-key and attribute certificate frameworks," ITU-T Recommendation X.509, 9.1 ed., Oct. 14, 2021. [Online]. Available: https://www.itu.int/rec/T-REC-X.509/. Accessed: Oct. 23, 2023.

[42] Withers, A., Bockelman, B., Weitzel, D., Brown, D., Gaynor, J., Basney, J., Tannenbaum, T., Miller, Z. (2018, July). SciTokens. In Proceedings of the Practice and Experience on Advanced Research Computing (pp. 83-92). ACM. doi: 10.1145/3219104.3219135.

[43] Xu, Liutong & Ai, Bo. (2003). FTPGrid: A new paradigm for distributed FTP systems. 3033. 895-898. 10.1007/978-3-540-24680-0_141.

[44] Google Developers. "Enable HTTPS." [Online]. Available: https://web.dev/articles/enable-https?hl=en&visit_id=638345331036731898-3627640507&rd=1, accessed 1 November 2023.

[45] A. Dorigo, P. Elmer, F. Furano, and A. Hanushevsky, "XROOTD - A highly scalable architecture for data access," WSEAS Transactions on Computers, vol. 4, no. 5, pp. 348-353, 2005.

[46] A. Aimar, A. Corman, P. Andrade, J. Fernandez, B. Bear, E. Karavakis, D. Kulikowski, and L. Magnoni, "MONIT: Monitoring the CERN Data Centres and the WLCG Infrastructure," EPJ Web of Conferences, vol. 214, p. 08031, 2019, doi: 10.1051/epjconf/201921408031.

[47] G. Apollinari, O. Bruening, T. Nakamoto, and L. Rossi, "High Luminosity Large Hadron Collider (HL-LHC)," CERN-2015-005.1, CERN, Geneva, 2017, doi: 10.5170/CERN-2015-005.1.

[48] CERN, "High Luminosity LHC Project," https://hilumilhc.web.cern.ch/content/hl-lhc-project, accessed 1 November 2023.

[49] The HEP Software Foundation., Albrecht, J., Alves, A.A. et al. A Roadmap for HEP Software and Computing R&D for the 2020s. Comput Softw Big Sci 3, 7 (2019). https://doi.org/10.1007/s41781-018-0018-8.

[50] CMS Collaboration, "CMS Offline Computing Results," CMS TWiki, https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults, accessed 1 November 2023.

[51] I. Bird, S. Campana, M. Girone, X. Espinal, G. McCance, and J. Schovancová, "Architecture and prototype of a WLCG data lake for HL-LHC," EPJ Web of Conferences, vol. 214, p. 04024, 2019, doi: 10.1051/epjconf/201921404024.

[52] “Pentaho, Hadoop, and Data Lakes,” James Dixon’s Blog, Oct. 14, 2010. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (accessed Jul. 03, 2023).

[53] M. Pedersen et al., “Nordugrid ARC Datastaging and Cache: Efficiency gains on HPC and cloud resources,” EPJ Web of Conferences, vol. 245, p. 03011, 2020, doi: 10.1051/epjconf/202024503011.

[54] “Data Organization, Management and Access (DOMA),” Institute for Research and Innovation in Software for High Energy Physics. https://iris-hep.org/doma.html (accessed Jul. 12, 2023).

[55] X. Espinal, "Conceptual sketch of the WLCG Data Lake," Fig. 1, EPJ Web Conf., vol. 245, p. 04027, 2020, doi: 10.1051/epjconf/202024504027.

[56] RedIRIS. "RedIRIS - Welcome to RedIRIS." (2023, October 30). [Online]. Available at: https://www.rediris.es/.

[57] RES, "The Spanish Supercomputing Network," https://www.res.es/en/about, accessed 1 November 2023.

[58] Bauerdick, L. & Bloom, K. & Bockelman, B. & Bradley, D. & Dasu, S. & Dost, Jeffrey & Sfiligoi, Igor & Tadel, A. & Tadel, Matevz & Würthwein, Frank & Yagil, A.. (2014). "XRootD, disk-based, caching proxy for optimization of data access, data placement and data replication." Journal of Physics: Conference Series. 513. DOI: 10.1088/1742-6596/513/4/042044.

[59] Pryce, D. (11 May 1989). "80486 32-bit CPU breaks new ground in chip density and operating performance. (Intel Corp.) (product announcement)." EDN.

[60] ARM, "A Brief History of ARM Part 1," ARM Community Blog, https://community.arm.com/arm-community-blogs/b/architectures-and-processors-blog/posts/a-brief-history-of-arm-part-1, accessed 1 November 2023.

[61] Bakoglu, H. B., Grohoski, G. F., & Montoye, R. K. (1990, January). The IBM RISC System/6000 processor: Hardware overview. *IBM Journal of Research and Development*, 34(1), 12-22. https://doi.org/10.1147/rd.341.0012

[62] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," PLOS ONE, vol. 12, no. 5, 2017, doi: 10.1371/journal.pone.0177459.

[63] B. Burns, "The History of Kubernetes & the Community Behind It," in Kubernetes Blog, Jul. 20, 2018. [Online]. Available: https://kubernetes.io/blog/2018/07/20/the-history-of-kubernetes-the-community-behind-it/. [Accessed: Oct. 14, 2023].

[64] F. Stagni, A. Valassi, and V. Romanovskiy, "Integrating LHCb workflows on HPC resources: status and strategies," in EPJ Web of Conferences, vol. 245, p. 09002, 2020. DOI: 10.1051/epjconf/202024509002.

[65] Douglas Benjamin, Taylor Childers, David Lesny, Danila Oleynik, Sergey Panitkin, Vakho Tsulaia, Wei Yang, and Xin Zhao, "Building and using containers at HPC centres for the ATLAS experiment," EPJ Web of Conferences, vol. 214, p. 07005, 2019. DOI: 10.1051/epjconf/201921407005.

[66] Barcelona Supercomputing Center (BSC-CNS). Available at: https://www.bsc.es/. Last accessed on 23rd Oct 2023.

[67] C. Acosta Silva, A. Peris, J. Molina, J. Frey, M. Hernández, M. Livny, A. Yzquierdo, and T. Tannenbaum, "Exploiting network restricted compute resources with HTCondor: a CMS experiment experience," EPJ Web of Conferences, vol. 245, p. 09007, 2020. DOI: 10.1051/epjconf/202024509007.

[68] Martin Barisits, Fernando Barreiro, Thomas Beermann, Karan Bhatia, Kaushik De, Arnaud Dubreuil, Johannes Elmsheuser, Alexei Klimentov, Mario Lassnig, Peter Love, Tadashi Maeno, Andrea Manzi, Ruslan Mashinistov, Andy Murphy, Paul Nilsson, Sergey Panitkin, and Tobias Wegner, "The Data Ocean Project - An ATLAS and Google R&D collaboration," EPJ Web of Conferences, vol. 214, p. 04020, 2019. DOI: 10.1051/epjconf/201921404020.

[69] WLCG Collaboration. "Signed Memoranda of Understanding." Accessed 2023-10-30. [Online]. Available: https://wlcg.web.cern.ch/mou/signed.

[70] CERN. "RRB - LHC Experiments Resources Review Boards." Accessed 2023-10-30. [Online]. Available: https://indico.cern.ch/.

[71] Valassi, A., Basset, R., Clemencic, M., Pucciani, G., Schmidt, S., & Wache, M. (2008). COOL, LCG conditions database for the LHC experiments: Development and deployment status. In 2008 IEEE Nuclear Science Symposium Conference Record (pp. 3021-3028). IEEE. doi: 10.1109/NSSMIC.2008.4774995.

[72] Blumenfeld, B & Dykstra, Dave & Lueking, L & Wicklund, E. (2008). "CMS conditions data access using FroNTier." Journal of Physics: Conference Series. 119. 072007. DOI: 10.1088/1742-6596/119/7/072007.

[73] Global Grid User Support (GGUS). Accessed 2023-11-03. [Online]. Available: https://ggus.eu/.

[74] Sfiligoi, Igor & Bradley, Daniel & Holzman, Burt & Mhashilkar, Parag & Padhi, Sanjay & Wuerthwein, Frank. (2009). "The Pilot Way to Grid Resources Using glideinWMS." In: 2009 WRI World Congress on Computer Science and Information Engineering, CSIE 2009. 2, 428-432. DOI: 10.1109/CSIE.2009.950.

[75] Maeno, T., De, K., Wenaus, T., Nilsson, P., Stewart, G., Walker, R., Stradling, A., Caballero, J., Potekhin, M., & Smith, D. (2011). Overview of ATLAS PanDA workload management. *Journal of Physics: Conference Series*, 331(7), 072024. https://doi.org/10.1088/1742-6596/331/7/072024.

[76] Saiz, P., Aphecetche, L., Buncic, P., Piskač, R., Revsbech, J.-E., & Šego, V. (2003). AliEn—ALICE environment on the GRID. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 502(2-3), 437-440. https://doi.org/10.1016/S0168-9002(03)00462-5.

[77] EGI. (October 30, 2023). Check-in Service. Retrieved from https://www.egi.eu/service/check-in/.

[78] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and S. Tuecke, "The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets," in Journal of Network and Computer Applications, vol. 23, no. 3, pp. 187-200, 2000. ISSN: 1084-8045. Available: https://doi.org/10.1006/jnca.2000.0110[1].

[79] Tim Berners-Lee. "The Original HTTP as Defined in 1991." World Wide Web Consortium. [Online]. Available: www.w3.org. Retrieved: 24 July 2010. Accessed: 31 August 2023.

[80] Stanford University. "GSI Protocol Specifications." [Online]. Available: https://xrootd.slac.stanford.edu/doc/gsidocs/XRootD-GSI-Protocol-Specifications.pdf. Accessed: 31 August 2023.

[81] ITU. "X.509: Information Technology - Open Systems Interconnection - The Directory: Public-Key and Attribute Certificate Frameworks." [Online]. Available: https://www.itu.int/rec/T-REC-X.509. Accessed: 6 November 2019.

[82] JSON Web Tokens (JWT). Accessed 2023-11-03. [Online]. Available: https://jwt.io/.

[83] Donno, Flavia, Abadie, Lanna, Badino, Paolo, Baud, J-P, Corso, Ezio, de Witt, Shaun, Fuhrmann, Patrick, Gu, Junmin, Koblitz, Birger, Lemaitre, Sophie, Litmaath, Maarten, Litvintsev, D., Lo Presti, Giuseppe, Magnoni, Luca, Mccance, Gavin, Mkrtchan, Tigran, Mollon, Rémi, Natarajan, Vijaya, Perelmutov, Timur, and Zappi, Riccardo. (2008). Storage Resource Manager version 2.2: design, implementation, and testing experience. Journal of Physics: Conference Series, 119. doi: 10.1088/1742-6596/119/6/062028.

[84] Williams, Brad, Tadlock, Justin, and Jacoby, John. (2020). REST API. 10.1002/9781119666981.ch12.

[85] dCache.org. "dCache: Distributed Storage for Scientific Data." [Online]. Available: https://www.dcache.org/. Accessed: 12 July 2023.

[86] DESY. "DESY - Deutsches Elektronen-Synchrotron." [Online]. Available: https://www.desy.de/. Accessed: 1 November 2023.

[87] Fermi National Accelerator Laboratory. "Fermilab - America's Premier National Lab for Particle Physics and Accelerator Research." [Online]. Available: https://www.fnal.gov/. Accessed: 1 November 2023.

[88] Norwegian Electronic Infrastructure Consortium. "NeIC - Nordic e-Infrastructure Collaboration." [Online]. Available: https://neic.no/. Accessed: 11 November 2023.

[89] Indiana University Knowledge Base. "Accessing Data Stored on NeIC." Archived from https://kb.iu.edu/d/agjv on 14 June 2018. Accessed 11 November 2023. Journal Format.

[90] Sandberg, R., Goldberg, D., Kleiman, S., Walsh, D., & Lyon, B. (1985). Design and Implementation of the Sun Network File System. Proceedings of the Summer 1985 USENIX Conference. Journal Format.

[91] Peters, AJ, Sindrilaru, EA, & Adde, G. (2015). EOS as the Present and Future Solution for Data Storage at CERN. Journal of Physics: Conference Series, 664, 042042. doi:10.1088/1742-6596/664/4/042042.

[92] Whitehead, E. James, & Goland, Yaron Y. (1999). "WebDAV." Ecscw' 99. Netherlands: Springer Science+Business Media. pp. 291–310. doi:10.1007/978-94-011-4441-4_16. ISBN 978-94-011-4441-4.

[93] Microsoft Learn. (n.d.). Overview of Microsoft SMB Protocol and CIFS Protocol. Retrieved from https://learn.microsoft.com/es-es/windows/win32/fileio/microsoft-smb-protocol-and-cifs-protocol-overview?redirectedfrom=MSDN (Accessed 11 November 2023).

[94] Libfuse/libfuse [Software]. GitHub. Available from: https://github.com/libfuse/libfuse [Accessed 11 November 2023].

[95] gRPC. (n.d.). gRPC Website. Retrieved from https://grpc.io/ (Accessed 11 November 2023).

[96] Davis, M., Bahyl, V., Cancio, G., Cano, E., Leduc, J., & Murray, S. (2019). CERN Tape Archive — from development to production deployment. EPJ Web of Conferences, 214, 04015. https://doi.org/10.1051/epjconf/201921404015.

[97] Italian Grid. "Italian Grid STORM Functional Description." Accessed 2023-11-03. [Online]. Available: https://italiangrid.github.io/storm/documentation/functional-description/1.11.2/.

[98] INFN-CNAF, "INFN-CNAF Website," https://www.cnaf.infn.it/, accessed 11 November 2023.

[99] EOSC Portal, "EOSC Projects," https://eosc-portal.eu/about/eosc-projects, accessed 11 November 2023.

[100] CERN. "CERN Disk Pool Manager (DPM)." Accessed 2023-11-03. [Online]. Available: https://lcgdm.web.cern.ch/dpm.

[101] F. Schmuck and R. Haskin, "GPFS: A Shared-Disk File System for Large Computing Clusters." USENIX, 2001. [Online]. Available: https://www.usenix.org/legacyurl/title-18.

[102] Lustre Home. Archived from the original on March 31, 2001. Retrieved September 23, 2013. [Online]. Available: https://lustre.org/lustre-home.

[103] Sage A. Weil, Scott A. Brandt, Ethan L. Miller, Carlos Maltzahn, and Charles H. Lee. "Ceph: Reliable, Scalable, and High-Performance Distributed Storage." In Proceedings of the 6th ACM Symposium on Cloud Computing, SoCC '10, pages 317-328, New York, NY, USA, 2010. ACM. doi:10.1145/1807128.1807164.

[104] Enstore-org. "Enstore-org/enstore." Accessed 2023-11-03. [Online]. Available: https://github.com/Enstore-org/enstore.

[105] "The HPSS Archive System." NERSC Documentation. Retrieved from https://docs.nersc.gov/filesystems/archive/ on November 7, 2023.

[106] IBM. "Tivoli® Storage Manager." Accessed 2023-11-03. [Online]. Available at: https://www.ibm.com/products/tivoli-storage-manager.

[107] Scientific Linux. "Scientific Linux." Accessed 2023-11-03. [Online]. Available: https://scientificlinux.org/.

[108] The CentOS Project. (2023, October 30). The CentOS Project. [Online]. Available: https://www.centos.org/.

[109] Red Hat. "Red Hat Enterprise Linux." Accessed 2023-11-03. [Online]. Available: https://www.redhat.com/en/technologies/linux-platforms/enterprise-linux.

[110] Sylabs. "Deploying Secure Container Solutions from Edge to Exascale". 2023. Sylabs. [Online]. Available: https://sylabs.io/.

[111] CERN IT Linux Team. (2023, October 30). CentOS 8 (CS8) - Linux @ CERN. [Online]. Available: https://linux.web.cern.ch/centos8/linux8/.

[112] Rocky Linux. 2023. Rocky Linux. [Online]. Available: https://rockylinux.org/.

[113] AlmaLinux OS Foundation. 2023. AlmaLinux OS - Forever-Free Enterprise-Grade Operating System. [Online]. Available: https://almalinux.org/.

[114] Oracle. "Oracle Tape Storage." (2023, October 30). [Online]. Available at: https://www.oracle.com/storage/tape-storage/.

[115] IBM. "IBM Tape Storage." (2023). [Online]. Available at: https://www.ibm.com/tape-storage.

[116] Morais, M. G. P. (2017). Optimization of the LHCb tape drive software for data acquisition. CERN-THESIS-2017-131. Retrieved from https://cds.cern.ch/record/2282014/files/CERN-THESIS-2017-131.pdf.

[117] Waczyńska, J., Martelli, E., Vallecorsa, S., Karavakis, E., & Cass, T. (2021). Convolutional LSTM models to estimate network traffic. EPJ Web of Conferences, 251, 02050. https://doi.org/10.1051/epjconf/202125102050.

[118] Busse-Grawitz, C., Martelli, E., Lassnig, M., Manzi, A., Keeble, O., & Cass, T. (2020). "The NOTED software tool-set improves efficient network utilization for Rucio data transfers via FTS." EPJ Web Conf., 245, 07022.

[119] CERN. "VOMS." CERN ServiceNow. https://cern.service-now.com/service-portal?id=functional_element&name=VOMS. Accessed 11 November 2023.

[120] OAuth 2.0 (2023, October 30). [Online]. Available: https://oauth.net/2/.

[121] OpenID Foundation (2023, October 30). [Online]. Available: https://openid.net/.

[122] Indigo IAM. iam: Identity and access management for Indigo. GitHub repository. Retrieved from https://github.com/indigo-iam/iam.

[123] EGI. (October 30, 2023). Check-in Service. Retrieved from https://www.egi.eu/service/check-in/.

[124] Bockelman, Brian & Ceccanti, Andrea & Collier, Ian & Cornwall, Linda & Dack, Thomas & Guenther, Jaroslav & Lassnig, Mario & Litmaath, Maarten & Millar, A. & Sallé, Mischa & Short, Hannah & Teheran, Jeny & Wartel, Romain. (2020). "WLCG Authorisation from X.509 to Tokens." EPJ Web of Conferences. 245. 03001. DOI: 10.1051/epjconf/202024503001.

[125] CMS CRIC Accounts. https://cms-cric.cern.ch/accounts/account/list/. Accessed August 8, 2023.

[126] Giordano, Domenico, Alef, Manfred, & Michelotto, Michele. (2019). Next Generation of HEP CPU Benchmarks. EPJ Web of Conferences, 214, 08011. https://doi.org/10.1051/epjconf/201921408011.

[127] University of Florida, Institute of High Energy Physics, "The CMS Experiment", https://ihepa.phys.ufl.edu/the-cms-experiment/, accessed 11 November 2023.

[128] CMS Collaboration. "CMS Detector Components." WorkBook CMS Experiment, CMS TWiki, https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperiment, accessed 11 November 2023.

[129] CMS Collaboration. "Schematic view of the CMS detector." WorkBook CMS Experiment, CMS TWiki, https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperiment, accessed 11 November 2023.

[130] Kuznetsov, Valentin & Evans, Dave & Metson, Simon. (2010). "The CMS data aggregation system." Procedia CS. 1. 1535-1543. DOI: 10.1016/j.procs.2010.04.172.

[131] P. Elmer, "CMS Trivial File Catalog and Site Local Configuration", https://indico.cern.ch/event/5490/contributions/1207463/attachments/979331/1391873/elmer_cms_trivial_file_catalog.pdf, accessed 11 November 2023.

[132] Megino, Barreiro, Cinquilli, M., Giordano, Domenico, Karavakis, Edward, Girone, Maira, Magini, N., Mancinelli, V., & Spiga, Daniele. (2012). Implementing data placement strategies for the CMS experiment based on a popularity model. Journal of Physics: Conference Series, 396. https://doi.org/10.1088/1742-6596/396/3/032047.

[133] Somnath Choudhury. (2015). "4-lepton invariant mass distribution in the CMS experiment and the Higgs boson mass peak over almost flat ZZ background in the mass range of interest." Measurements of the Higgs Boson at the LHC and Tevatron. EPJ Web of Conferences, 90, 05002. https://doi.org/10.1051/epjconf/20159005002.

[134] WMCore: Core workflow management components for CMS. GitHub repository. Retrieved from https://github.com/dmwm/WMCore.

[135] CERN. (2023, October 30). Preparation of new WMAgents. [Online]. Retrieved from https://twiki.cern.ch/twiki/bin/view/CMSPublic/PreparationNewWMAgent.

[136] A. Pérez-Calero Yzquierdo, M. A. Acosta Flechas, D. Davila, S. Haleem, K. P. Hurtado Anampa, T. T. Ivanov, F. A. Khan, E. Kizinevic, K. E. Larson, J. Letts, M. Mascheroni, and D. A. Mason, "Evolution of the CMS Global Submission Infrastructure for the HL-LHC Era," EPJ Web Conf. 245, 03016 (2020).

[137] A. Pérez-Calero Yzquierdo, J. Balcas, J. Hernandez, F. Khan, J. Letts, D. Mason, & V. Verguilov. (2017). "CMS readiness for multi-core workload scheduling." Journal of Physics: Conference Series, 898(5), 052030. https://doi.org/10.1088/1742-6596/898/5/052030.

[138] A. Pérez-Calero Yzquierdo. “The CMS Computing Infrastructure”. PIC Seminar presented on 20th of October of 2023.

[139] CMS Collaboration. CMSSW: The CMS Software Framework. GitHub repository. Retrieved from https://github.com/cms-sw/cmssw.

[140] CMS Collaboration, "Event displays of lead-lead ion collisions in the CMS detector during Run 3, 26 September 2023", CERN Document Server, https://cds.cern.ch/record/2872371/files/, accessed 11 November 2023.

[141] CERN. (2023, October 30). ROOT. [Online]. Retrieved from https://root.cern/.

[142] CMS Collaboration. "Events from a software point of view: The Event Data Model (EDM)." WorkBook CMS Software Framework. CMS TWiki. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBook-CMSSWFramework. Accessed 11 November 2023.

[143] CERN. (2023, October 30). SWGuideCrab: Software Guide on CRAB. [Online]. Retrieved from https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCrab.

[144] Apache ActiveMQ Project. ActiveMQ: Apache ActiveMQ Message Broker. [Online]. Retrieved from https://activemq.apache.org/.

[145] Elastic. (2023, October 30). Elasticsearch. [Online]. Available: https://www.elastic.co/.

[146] InfluxData. (Oct. 30, 2023). InfluxDB. Retrieved from https://www.influxdata.com/.

[147] K. Shvachko, H. Kuang, S. Radia , and R. Chansler, “The Hadoop Distributed File System,” *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, doi: 10.1109/MSST.2010.5496972.

[148] Elastic. (October 30, 2023). Kibana. Retrieved from https://www.elastic.co/es/kibana.

[149] Piparo, Danilo & Tejedor, Enric & Mato, Pere & Mascetti, Luca & Moscicki, Jakub & Lamanna, Massimo. (2016). SWAN: A service for interactive analysis in the cloud. Future Generation Computer Systems. 78. 10.1016/j.future.2016.11.035.

[150] Apache Software Foundation. "OpenSearch." https://opensearch.org/. Accessed 11 November 2023.

[151] Meoni, Marco & Kuznetsov, Valentin & Menichetti, Luca & Rumševičius, Justinas & Boccali, Tommaso & Bonacorsi, Daniele. (2017). "Exploiting Apache Spark platform for CMS computing analytics." Journal of Physics: Conference Series. 1085. DOI: 10.1088/1742-6596/1085/3/032055.

[152] Andrade, P. & Babik, M. & Bhatt, Kislay & Chand, Phool & Collados, David & Duggal, Vibhuti & Fuente, P. & Imamagic, Emir & Joshi, P. & Kalmady, R. & Karnani Gaur, Urvashi & Kumar, V. & Tarragon, Josep & Lapka, W. & Triantafyllidis, C.. (2012). "Service Availability Monitoring Framework Based On Commodity Software." Journal of Physics Conference Series. 396. 2008-. DOI: 10.1088/1742-6596/396/3/032008.

[153] Schovancová, Jaroslava & Girolamo, Alessandro & Fkiaras, Aristeidis & Mancinelli, Valentina. (2019). "Evolution of HammerCloud to commission CERN Compute resources." EPJ Web of Conferences. 214. 03033. DOI: 10.1051/epjconf/201921403033.

[154] DUNE Collaboration. (Oct. 30, 2023). "DUNE." Retrieved from https://www.dunescience.org/.

[155] South African Radio Astronomy Observatory. (2023, October 30). "The SARAO Project." [Online]. Retrieved from https://www.sarao.ac.za/about/the-project/.

[156] LSST Science Collaboration. (October 30, 2023). "LSST." Retrieved from https://www.lsst.org/.

[157] Campana Simone, «WLCG data challenges for HL-LHC - 2021 planning». Zenodo, sep. 27, 2021. doi: 10.5281/zenodo.5532452.

[158] Forti A., on behalf of the Data Challenge Team, "WLCG Network Data Challenges 2021: wrap-up and recommendations". Zenodo, Dic. 08, 2021. doi: 10.5281/zenodo.5767913.

[159] Forti A., on behalf of the Data Challenge Team, "Network Data Challenge and some Tape", GDB, 13 October 2021.

[160] Forti A., on behalf of the Data Challenge Team, "Network Data Challenge", GDB, 10 November 2021.

[161] McKee S., "DC24 Planning and Near Term Activities". Presented on WLCG DOMA General Meeting, Jun.23, 2023.

[162] Hannemann, A., Boote, J., Boyd, E., Durand, J., Kudarimoti, L., Lapacz, R., Swany, M., Trocha, S., Zurawski, J. (2005). PerfSONAR: A Service Oriented Architecture for Multi-domain Network Monitoring. In Proceedings of the 2005 Network Performance Measurement Workshop, 241-254. doi:10.1007/11596141_19.

[163] Netflix, Inc. (2023, October 30). "Netflix." Available at: https://www.netflix.com/ (Accessed October 30, 2023).

[164] Spotify AB (2023, October 30). "Spotify (International Spanish)." Available at: https://open.spotify.com/intl-es (Accessed October 30, 2023).

[165] Steinmetz, R.; Wehrle, K (2005). "What Is This 'Peer-to-Peer' About?". Springer Berlin Heidelberg. pp. 9-16.

[166] Facebook, Inc. (2023, October 30). "Facebook." Available at: https://www.facebook.com/ (Accessed October 30, 2023).

[167] Amazon.com, Inc. (2023, October 30). "Amazon.com." Available at: https://www.amazon.com/ (Accessed October 30, 2023).

[168] Alexa Internet, Inc. (2023, October 30). "Top 1000 Sites." Available at: https://www.alexa.com/topsites (Accessed October 30, 2023).

[169] Cisco Systems, Inc. (2023, October 30). "VNI Complete Forecast Highlights Global - Consumer - Business - Service Provider Internet Traffic." White Paper. [Online]. Available at: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html (Accessed October 30, 2023).

[170] Paul V. Mockapetris, Kevin J. Dunlap, "Development of the Domain Name System." Comput. Commun. Rev. 25(1): 112-122 (1995).

[171] Rooney, T. (2010). The Domain Name System (DNS) Protocol. In Computer Networks: A Systems Approach (5th ed., pp. 217-244). John Wiley & Sons. doi:10.1002/9780470880654.ch9.

[172] V. Jacobson and V. Aggarwal, "Akamai: A Global Content Delivery System," Proceedings of the ACM SIGCOMM '96 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, 1996, pp. 13-22.

[173] Davis, A.; Parikh, J.; Weihl, W. (2004). “Edge Computing: Extending Enterprise Applications to the Edge of the Internet”. 13th International World Wide Web Conference. S2CID 578337. doi: 10.1145/1013367.1013397.

[174] 23rd International Conference on Computing in High Energy and Nuclear Physics, CHEP 2018. 9 - 13 Jul 2018, Sofia, Bulgaria.

[175] Lamanna, G. (2023, May 8). "The ESCAPE Collaboration - long term perspective." Plenary Track 10 - Exascale Science Plenary Session, 26th International Conference on Computing in High Energy & Nuclear Physics, CHEP 2023, (Norfolk, VA, USA).

[176] Squid Development Group, "Squid Cache," https://www.squid-cache.org/, accessed 11 November 2023.

[177] L. Bauerdick, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, M. Ernst, R. Gardner, A. Hanushevsky, H. Ito, D. Lesny, P. McGuigan, S. McKee, O. Rind, H. Severini, I. Sfiligoi, M. Tadel, I. Vukotic, S. Williams, and W. Yang, "Using XRootD to Federate Regional Storage," Journal of Physics: Conference Series, vol. 396, no. 4, p. 042009, 2012, doi: 10.1088/1742-6596/396/4/042009.

[178] C. Fröhlich, J. Romero, H. Roth, C. Wehrli, B. Andersen, T. Appourchaux, V. Domingo, U. Telljohann, G. Berthomieu, P. Delache, J. Provost, T. Toutain, D.A.H. Crommelynck, A. Chevalier, A. Fichot, W. Däppen, D. Gough, T. Hoeksema, A. Jiménez, and R. Willson, "VIRGO: Experiment for helioseismology and solar irradiance monitoring," Solar Physics, vol. 162, no. 1, pp. 101-128, 1995, doi: 10.1007/BF00733428.

[179] StashCache Team, "StashCache," https://stashcache.github.io/, accessed 11 November 2023.

[180] Varnish Cache. (2023, October 30). "Varnish Cache - Web Application Accelerator." [Online]. Available at: https://varnish-cache.org/ (Accessed October 30, 2023).

[181] Apache Traffic Server. (2023, October 30). "Apache Traffic Server - A Web Performance Server." [Online]. Available at: https://trafficserver.apache.org/ (Accessed October 30, 2023).

[182] Hanushevsky, Andrew & Ito, Hironori & Lassnig, Mario & Popescu, Radu & Silva, Asoka & Simon, Michal & Gardner, Robert & Garonne, Vincent & Stefano, John & Vukotic, Ilija & Yang, Wei. (2019). "Xcache in the ATLAS Distributed Computing Environment." EPJ Web of Conferences. 214. 04008. DOI: 10.1051/epjconf/201921404008.

[183] NERSC, "National Energy Research Scientific Computing Center," https://www.nersc.gov/, accessed 11 November 2023.

[184] UC San Diego, "University of California San Diego," https://ucsd.edu/, accessed 11 November 2023.

[185] Caltech, "California Institute of Technology," https://www.caltech.edu/, accessed 11 November 2023.

[186] Fajardo, Edgar & Tadel, Matevz & Balcas, Justas & Tadel, A. & Würthwein, Frank & Davila, Diego & Guiang, Jonathan & Sfiligoi, Igor. (2020). "Moving the California distributed CMS XCache from bare metal into containers using Kubernetes."

[187] SLAC National Accelerator Laboratory, "SLAC National Accelerator Laboratory," https://www6.slac.stanford.edu/, accessed 11 November 2023.

[188] CERN, "RedirectorsSubscription," https://twiki.cern.ch/twiki/bin/view/Main/RedirectorsSubscription, accessed 11 November 2023.

[189] Ministerio de Ciencia e Innovación. "Infraestructuras Científicas y Técnicas Singulares (ICTS)." Retrieved September 26, 2023, from https://www.ciencia.gob.es/Organismos-y-Centros/ICTS.html.

[190] Universitat Autònoma de Barcelona. (2023, October 30). "UAB - Universitat Autònoma de Barcelona." [Online]. Available at: https://www.uab.cat/ (Accessed October 30, 2023).

[191] Ministerio de Asuntos Económicos y Transformación Digital. (2023, October 30). "Ministerio de Asuntos Económicos y Transformación Digital." [Online]. Available at: https://portal.mineco.gob.es (Accessed October 30, 2023).

[192] Generalitat de Catalunya. (2023, October 30). "Politics and Economy." [Online]. Available at: https://web.gencat.cat/en/temes/catalunya/coneixer/politica-economia/ (Accessed October 30, 2023).

[193] "MAGIC," The MAGIC Telescope Web Server on La Palma. http://www.magic.iac.es/ (accessed Jul. 11, 2023).

[194] "PAUCam – PAU." https://pausurvey.org/paucam/ (accessed Jul. 11, 2023).

[195] Dark Energy Survey Collaboration, "Dark Energy Survey," https://www.darkenergysurvey.org/, accessed 11 November 2023.

[196] "Euclid Consortium – A space mission to map the Dark Universe." https://www.euclid-ec.org/ (accessed Jul. 11, 2023).

[197] CELLS, "CELLS", https://www.cells.es/es/, accessed 11 November 2023.

[198] Proxmox VE, "Storage: RBD," https://pve.proxmox.com/wiki/Storage:_RBD, accessed 11 November 2023.

[199] Amazon Web Services, "Amazon S3," https://aws.amazon.com/es/s3/, accessed 11 November 2023.

[200] A. Thusoo et al., "Hive," Proceedings of the VLDB Endowment, vol. 2, no. 2, pp. 1626–1629, Aug. 2009, doi: 10.14778/1687553.1687609.

[201] Rocklin, Matthew. (2015). "Dask: Parallel Computation with Blocked algorithms and Task Scheduling." 126-132. DOI: 10.25080/Majora-7b98e3ed-013.

[202] JupyterHub Team, "JupyterHub," https://jupyter.org/hub, accessed 11 November 2023.

[203] Green Revolution Cooling. (2023, October 30). "GRC - Immersion Cooling Authority." [Online]. Available at: https://www.grcooling.com/ (Accessed October 30, 2023).

[204] Cabrillo, Ibán & Cabellos, Luis & Marco de Lucas, Jesus & Fernandez, Janfer & González Caballero, Isidro. (2014). "Direct exploitation of a top 500 Supercomputer for Analysis of CMS Data." Journal of Physics: Conference Series. 513. 032014. DOI: 10.1088/1742-6596/513/3/032014.

[205] Yoo, A. B., Plank, J. S., Thain, D., & Hensgen, D. A. (2003). "Slurm: Simple Linux Utility for Resource Management." In 10th IEEE International Symposium on High Performance Distributed Computing (HPDC '03) (pp. 114--122). IEEE. DOI: 10.1109/hpdc.2003.1219457.

[206] Acosta Silva, Carles, Peris, A., Flix, J., Guerrero, J., Hernández, J., Yzquierdo, A., Calonge, F., & Gómez-Pulgar, J. (2020). Lightweight site federation for CMS support. EPJ Web of Conferences, 245, 03013. doi:10.1051/epjconf/202024503013.

[207] Ciangottini, Diego & Bagliesi, Giuseppe & Biasotto, Massimo & Boccali, Tommaso & CESINI, Daniele & Donvito, Giacinto & Falabella, Antonio & Mazzoni, Enrico & Spiga, Daniele & Tracolli, Mirco. (2019). "Integration of the Italian cache federation within the CMS computing model." DOI: 10.22323/1.351.0014.

[208] Beermann, T., Maettig, P., Stewart, G., Lassnig, M., Garonne, V., Barisits, M., Vigne, R., Serfon, C., Goossens, L., Nairz, A., Molfetas, A., et al. (2014). Popularity Prediction Tool for ATLAS Distributed Data Management. Journal of Physics: Conference Series, 513(4), 042004.

[209] Meoni, M., Perego, R. & Tonellotto, N. Dataset Popularity Prediction for Caching of CMS Big Data. J Grid Computing 16, 211–228 (2018). https://doi.org/10.1007/s10723-018-9436-4

[210] dCache.org. (2023, October 30). "dCache Billing Database Configuration on a File System Hierarchy." [Online]. Available at: https://www.dcache.org/manuals/Book-2.8/config/cf-billing-db-fhs.shtml (Accessed October 30, 2023).

[211] PostgreSQL. (2023, October 30). "PostgreSQL - The world's leading open source relational database." [Online]. Available at: https://www.postgresql.org/ (Accessed October 30, 2023).

[212] Iiyama, Yutaro & Maier, Benedikt & Abercrombie, Daniel & Goncharov, Maxim & Paus, Christoph. (2021). "Dynamo: Handling Scientific Data Across Sites and Storage Media." Computing and Software for Big Science. 5. DOI: 10.1007/s41781-021-00054-2.

[213] Barrass, Newbold and Tuura 2005 'The CMS PhEDEx System: a Novel Approach to Robust Grid Data Distribution', UK e-science Programme All Hands Meeting, Nottingham, UK.

[214] A. Delgado Peris, J. Flix Molina, J. M. Hernández, A. P. C. Yzquierdo, C. P. Dengra, E. Planas, F. J. R. Calonge, and A. Sikora, "CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2," EPJ Web Conf. 245, 04028 (2020), doi: 10.1051/epjconf/202024504028.

[215] C. P. Dengra, J. F. Molina, and A. Sikora on behalf of the CMS Collaboration, "New storage and data access solution for CMS experiment in Spain towards HL-LHC era," J. Phys.: Conf. Ser. 2438, 012053 (2023), doi: 10.1088/1742-6596/2438/1/012053.

[216] G. Ramirez, "Muon Analysis: Muon Analyzer," GitLab repository, CERN, 2023, https://gitlab.cern.ch/garamire/muonanalysis-muonanalyzer/-/tree/master/.

[217] C. Pérez Dengra, J. Flix., and A. Sikora on behalf of the CMS Collaboration, "Deploying a cache content delivery network for CMS experiment in Spain," presented at the 21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 2023. (Submitted and accepted to be published in IOP Proceedings).

[218] KIST, "Korea Institute of Science and Technology," https://www.kisti.re.kr/eng/, accessed 11 November 2023.

[219] UNL, "University of Nebraska-Lincoln", https://www.unledu/, accessed 11 November 2023.

[220] Yarn Team, "Yarn", https://yarnpkg.com/, accessed 11 November 2023.

[221] Vohra, Deepak. (2016). "Apache Parquet." DOI: 10.1007/978-1-4842-2199-0_8.

[222] C. Pérez Dengra, J. Flix, A. Sikora, J. Casals, C. Acosta-Silva, C. Morcillo, A. Pérez-Calero Yzquierdo, A. Delgado Peris, and J. Hernández, "A study case of Content Delivery Network solutions for the CMS experiment" presented at the 21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Norfolk, VA, USA, May 8-12, 2023. (Submitted and accepted to be published in EPJ Conference Series).

[223] Daniele Spiga, Diego Ciangottini, Mirco Tracolli, Tommaso Tedeschi, Daniele Cesini, Tommaso Boccali, Valentina Poggioni, Marco Baioletti and Valentin Y. Kuznetsov,. "Smart Caching at CMS: applying AI to XCache edge services". EPJ Web Conf., 245 (2020) 04024. DOI: https://doi.org/10.1051/epjconf/202024504024.

[224] Tedeschi, Tommaso & Tracolli, Mirco & Ciangottini, Diego & Spiga, Daniele & Storchi, Loriano & Baioletti, Marco & Poggioni, Valentina. (2021). "Reinforcement Learning for Smart Caching at the CMS experiment." DOI: 10.22323/1.378.0009.

[225] C. Pérez Dengra, J. Flix, and A. Sikora on behalf of the CMS Collaboration., 2022 “Simulating a network delivery content solution for the CMS experiment in the Spanish” WLCG Tiers, International Symposium on Grids & Clouds (ISGC) 2022 Virtual Conference, 21-25 March 2022, last access on 25th of May of 2022: https://indico4.twgrid.org/event/20/contributions/1116/.

[226] P. Serrano, J. Flix, C. Pérez Dengra, A. Sikora, "Simulating a XCache for CMS Crab jobs executed in Spain" (Submitted to ISGC 2024). Link to the agenda: https://indico4.twgrid.org/event/33/.

[227] Tanenbaum, Andrew S. "Modern Operating Systems." 4th ed. Pearson Education, 2023.

[228] Hernández, J., Flix Molina, J., Rodriguez Calonge, F.J., Morcillo Perez, C., Leon Holgado, J., Cárdenas Montes, M., Pérez-Calero Yzquierdo, A. (2023). The Spanish CMS Analysis Facility at CIEMAT. In Proceedings of the 26th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2023) (pp. 1-8).