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

Carlos Pérez Dengra (PIC-CIEMAT)[1], **José Flix Molina** (PIC-CIEMAT)[1], **Anna Sikora** (Universitat Autònoma de Barcelona)[2] **on behalf of the CMS Collaboration.**
[1] Carrer de l'Albareda Edifici D · Campus UAB · 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain. [3] Carrer de les Sitges Edifici Q · Campus UAB · 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain. 2

[1]-cperez@pic.es

Abstract. The Large Hadron Collider (LHC) will enter a new era for data acquisition by 2029 within the High-Luminosity Large Hadron Collider (HL-LHC) program, where the LHC will raise the proton and ion collisions up to unprecedented levels. This increase will imply a factor 10 in terms of luminosity as compared to the current values, having an impact in the way the experimental data is stored and analyzed. This work focuses on the research and development of new data cache solutions that are adopted towards the new era, studying its effects at the Spanish WLCG sites that support CMS activities, namely the PIC Tier-1 and CIEMAT Tier-2. We propose a model to access and cache the most popular CMS datasets in Spain by deploying XCache cache systems in both PIC and CIEMAT sites. This work is complemented with dedicated studies to better understand and optimize these new solutions in the region.

1. Introduction

Since the start of LHC in 2009, the involved experiments have stored more than 1 Exabyte of simulated and collision data as a result of the underlying proton and ion collisions. These data has being stored on disk and tape, and processed in a worldwide distributed computing infrastructure comprising 170 centers in 35 countries, known as WLCG (World-wide LHC Computing Grid), with sites categorized by their size and functions, namely Tier-0 (at CERN), 13 Tier-1s and 150 Tier-2s **[1]**. The high estimates of data produced by LHC experiments in the HL-LHC period and the computation required to process this data will be a challenge for the expected budget. This situation led scientists to search for new mechanisms to alleviate these increases. Facing these challenges, the LHC experiments have launched an extensive Research and Development program to reduce the overall cost of storage, in terms of hardware and operations **[2]**. It also contemplates enabling efficient delivery of data at scale to large, remote and heterogeneous computing resources that are expected to be integrated into a network-centric and global infrastructure (Data Lake model). This program is developed in conjunction with other non-LHC data-intensive sciences that have similar computational challenges, since most of them use the same computational clusters available worldwide.

The WLCG Data Lake model presents a way to optimize the cost operations and reduce the hardware deployed by consolidating the storage resources in fewer sites. The sites would be also chosen for concentrating their efforts and investments specializing on running large computing farms, storage systems or both. Among the benefits of the model is that storage systems can be deployed as a distributed service accessed by remote facilities as a single-entry endpoint. Deploying resources in this manner allows a certain number of sites to share and offer resources as data federations, i.e. a collection of disparate storage resources that are transparently accessible across a wide area via a common namespace, taking into account geographic distance and latency.

The four LHC experiments, namely ATLAS, CMS, LHCb and ALICE work together in these R&D activities, under the WLCG DOMA working group **[3]**. Our contribution focuses on new mechanisms to cache popular datasets required by the CMS experiment **[4]**, based on the commissioning of XCache in the Spanish region. In the last years, the CMS community has put effort into the adoption of this technology **[5][6]**. The main goal is to reduce the data access latencies that affect the execution time of jobs and reduce or dispense with storage deployed at CMS Tier-2s (CIEMAT CMS Tier-2s in our case). These latencies are produced by the load time on Storage Element (SE) while accessing data stored in disk drives. The other cause of latency in job time executions are the limitations of bandwidth. Caches in worker nodes reduce the latency of jobs execution by providing data locally in those workflows where data is remotely accessed. Also, cached data is not subjected to custodial policies or cataloged, permitting the sites to deploy these caches in old or obsolete hardware. These advantages propagate directly on costs of hardware deployment and bandwidth at the sites. This work also presents dedicated studies on popular datasets and the benefits that cache techniques could bring to the experiment.

2. CMS data popularity in the spanish region

Popularity studies of Monte Carlo and collision data CMS files in the PIC Tier-1 and CIEMAT Tier-2 sites have been carried out to understand the data access patterns by the jobs executed at these sites. These studies have demonstrated that the improvement of the cache performance in some sites can improve as much as 20% **[7]**. Figure 1 and Figure 2 show the overall number of accesses to Monte Carlo and collision data files at both sites for the first ten months of 2021 (test datasets and test accesses have been excluded from the analysis).

File access information is provided for each job at run time by the CMS Software (CMSSW) and results are stored at the CERN HDFS (Hadoop Distributed File System) **[8]**. Many CMS job input files are accessed remotely via the CMS XRootD federation, known as AAA **[9]** (Any data, Anytime, Anywhere). The cmssw-popularity data source uses an internal service within CMSSW itself to send a small UDP packet to a server at CERN with details on the file access via the cmsRun application. Since this popularity study is based on the access of jobs to files from the PIC and the CIEMAT, the files not accessed from the local perspective of storage deployed on the site are not considered.

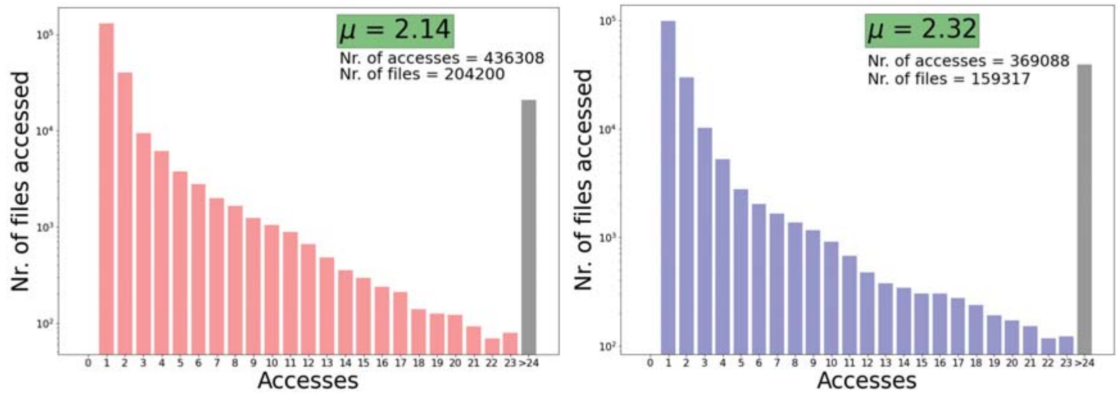

Figure 1: Data popularity distributions in 2021 (number of accesses by file) for the PIC Tier-1 (first and second captions) and CIEMAT Tier-2 (third and fourth captions), for both Monte Carlo (red) and collision data (blue) files accessed by executed jobs at the site.

We exclude from the analysis files such as tests, explicitly HammerCloud, SAM and the rest. Also, we exclude unmerged data and those files that we identified as badly reported by cmssw-popularity. Hence, keeping popular files in cache storage systems might help improving the efficiency of compute tasks, as well as reducing the deployed storage in the region if popular files could only be stored in these new storage cache elements. Since this popularity study is based on the access of jobs to files from the PIC and the CIEMAT, the files not accessed from the local perspective of storage deployed on the site are not considered.

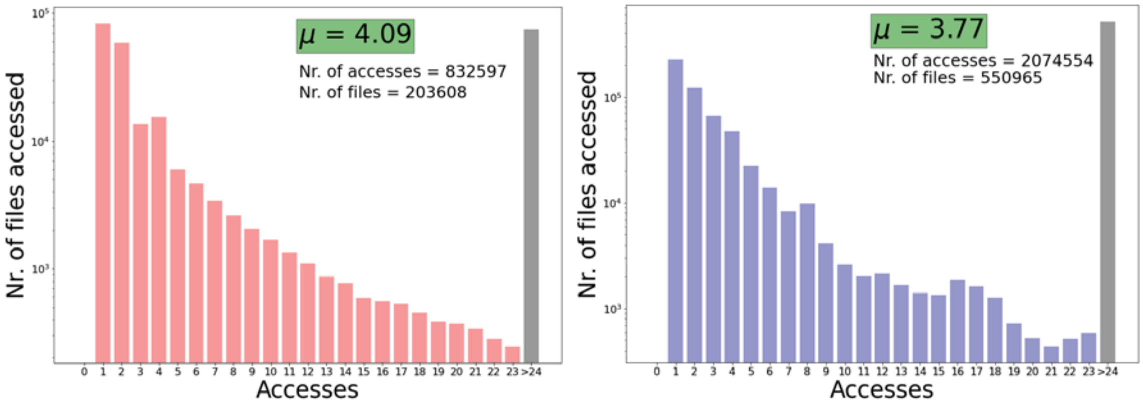

Figure 2: Data popularity distributions in 2021 (number of accesses by file) for the CIEMAT Tier-2, for both Monte Carlo (red) and collision data (blue) files accessed by executed jobs at the site.

3. Data volumes and mean access

Each CMS data tier accessed (experimental data types) has its own average number of accesses. An event in CMS consists in the signal of the involved particles in a single or several interectactions joined. These events are collected in the data formats, either collision or simulated data (Monte-Carlo). Finally these are categorized as data tiers regarding its level of processing and reconstruction. RAW consists of the first primary data from detectors, turning into RECO after the events reconstruction. When data is reconstructed including all the vertices, tracks, jets, electrons, etc; it is refined into RECO subsets that include more localized information called AOD. This data can still be re-processed into lighter data-tiers MINIAOD and NANOAOD. In this contribution we refer to all AOD formats as *AOD*.

This, in turn, can be compared with the total volume of unique data accessed from said data tier (the total TB that data would occupy if it had only been accessed once). This identifies the CMS Data Tiers in which the data is re-read often, and with a suitable size for being cached by a system deployed at each of the sites or in the region (CIEMAT and PIC are placed at 600 Km distance, ~10 ms RTT - studies showed no significant degradation in CPU efficiency if a federated storage model is deployed in the region **[10]**). The mean number of accesses per file have been computed over the total accessed files in the first ten months of 2021. The results are presented in Figure 3. They show that files of type *AOD* are good candidates to be placed in caches systems, since MINIAOD and MINIAODSIM have the higher maximum values of mean access and are *AOD* derived formats. A fraction of popular CMS files accessed at runtime by the CMS applications are not read completely, so many of these files could be cached partially (by blocks). This is being considered by the developers of cache services, such as XCache.

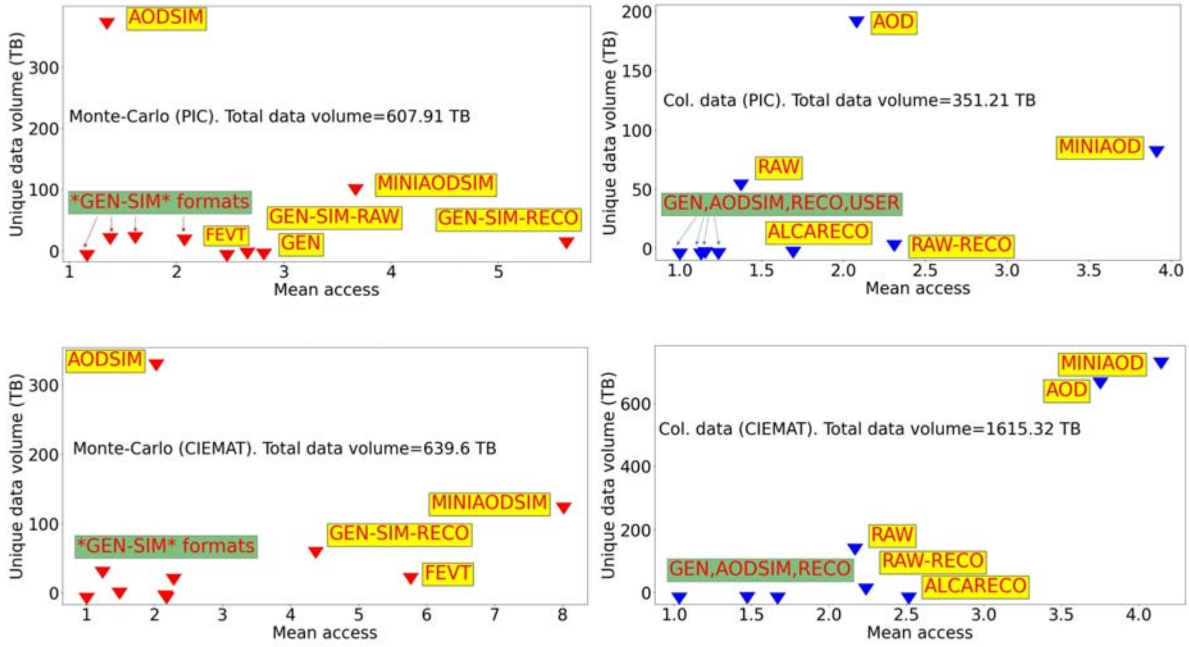

Figure 3: Data popularity mean accesses for the PIC Tier-1 and CIEMAT Tier-2 for both MC and DATA, separated by data tiers and displaying the unique data volume accessed, in 2021. The ideal candidates shown in the figure are the *AOD* formats.

4. Cache simulation at PIC Tier-1 for *AOD* MC and DATA

Based on the real executed jobs file accesses in 2021, we can simulate the expected behavior of a cache within a CMS site. The main objective of performing a cache simulation is to be able to evaluate the best features and how to increase their performance without having to deploy them in production in the real cache. The simulation is performed by adding the size of unique files accessed by jobs in real chronological order, in the same way that a physical cache would do and can be performed considering two approaches. The first approach consists in caching all files required by the executed jobs or caching just the files that are not already at the site storage system. Once the cache is full, a LRU algorithm ('Least Recently Used') is used to identify the files that are suitable for deletion. The algorithm organizes the files by order of use, and date of inclusion, and sets for deletion those which have not been accessed in the largest period of time. This deletion process is based on watermarks, thresholds that indicate a certain range of cache occupancy, as seen in Figure 4. If the cache occupancy overcomes "a high watermark HW", the least accessed files are removed until reaching "a low watermark LW". In these simulations, the watermark levels, as well as the size of the cache, can be adjusted in order to optimize the cache use by checking the hits - the number of times a file is present in the cache when is required by executed jobs at the sites. The simulation carried out Figure 5 with LW=0.5, HW=0.95 and 150TB of cache size gave an average hit-rate of 0.52 and 1100 hits/day filtering *AOD* files. Using the values of LW=0.5, HW=0.95 and 150TB (the current configuration for the cache deployed at PIC), the average hit-rate increased up to 0.63 with 1144 hits/day. Figure 5 shows monthly mean number of hits by file for the *AOD* kept in the simulated cache (*AOD* files stored at local disks present a mean lifetime of 41 days for collision data and 52 days for Monte-Carlo **[11]**). This metric has been chosen because as a cache fills up, the data that survives deletions is the most popular. Consequently, the average number of hits per month and per file increases over time and stabilizes. When it stabilizes, matches the average popularity we computed in Section 2, in our case, after 9 months, simulation shows that the monthly mean hits by file for *AOD* files are 2.6. Caching all *AOD* files would save the current custodial space dedicated to keep these data tiers at the sites.

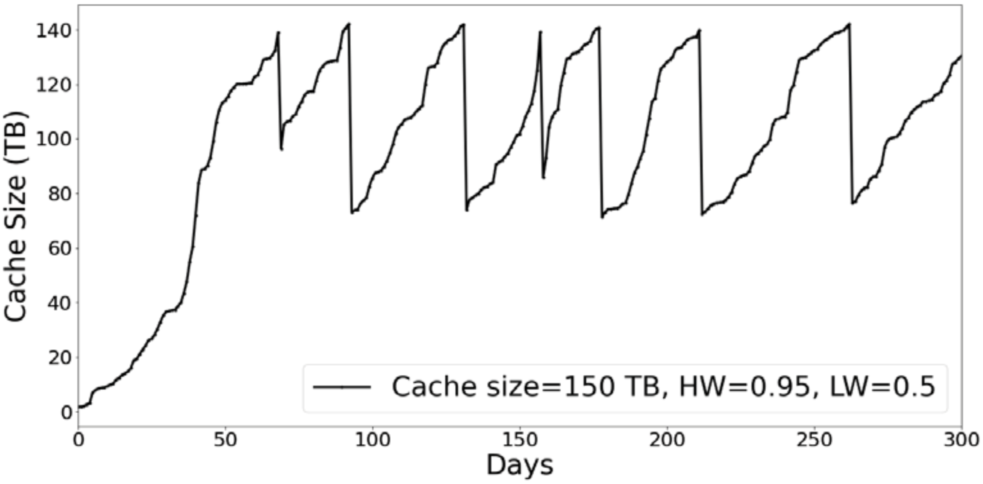

Figure 4: Evolution of a simulated 150TB cache size using LRU removal algorithm with high-watermark (HW) of 0.95 and low-watermark (LW) of 0. 5 at PIC Tier-1, using the data file accesses in 2021.

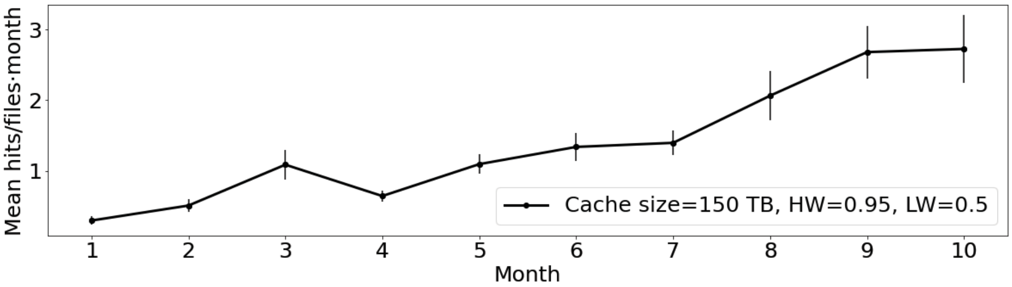

Figure 5: Monthly mean number of hits by file for *AOD*files kept in the simulated cache, using the historical file access information for PIC jobs in 2021. The trend increases during the year because the cache accumulates the most popular *AOD* files, whose number of hits is greater than one, and deletes those that have not been accessed again due to the LRU algorithm.

5. Deploying CMS XCache in the Spanish region

Real XCache test nodes have been introduced in both PIC and CIEMAT sites including a data-filter based on results of Section 4 for *AOD* files. For example, the PIC XCache instance (150 TB) was initially deployed in May 2020. XCache service provides data caching through Xrootd protocol via the AAA CMS infrastructure. This node was initially commissioned, caching all the requested files by jobs run at a single compute node. Later in 2021 this PIC service has been configured to cache only *AOD* files requested by CMS jobs running at the site, if those files are not presented in the local storage (i.e. the service caches all of those remote reads that are popular, for these data types). The service is currently running Xrootd 4.8.4, with the watermarks for LRU deletion set to 95% and 90%, for the HW and LW, respectively. The total disk drive storage deployed at PIC is 10 PB (3.4 PB dedicated to keep CMS custodial data). Part of the available storage for CMS is expected to be saved by introducing XCache instances at PIC and CIEMAT nodes by caching the total *AOD* data volume stored at the sites. During the first 300 days of 2021, 0.69PB of unique *AOD* files have been accessed by jobs through PIC compute nodes (corresponding to the 73% of the total unique volume accessed at the site). Local and remote *AOD* data accessed will be managed by the caching system and saved from the custodial space. The PIC CMS cache instance is part of the Spanish network of strategic data nodes.

6. Conclusions and outlook

Studies on the popularity of CMS data read by jobs that are executed in the Spanish region have been presented, with a quantitative identification of the most suitable files for being stored in cache systems, based on the repeatability of accesses. Further tasks can include Machine Learning techniques to analyze with more precision the most popular datasets or predict their popularity beforehand **[12]**. Based on the historical information of accesses, the cache service has been first simulated, in order to find the optimal configuration to keep popular files at both PIC and CIEMAT sites. Simulations based on caching-all *AOD* data formats have been carried out, and they are currently being refined, since much of this data used was already stored at the sites. The expected storage space saved on our sites corresponds to custodial space dedicated to *AOD* files. Since the jobs have accessed a unique data volume of 0.69 PB at PIC for the filtered data tier, the quantification of local and remote accessed bytes and the current volume kept of *AOD* at the sites will lead to a realistic estimation of the potential storage savings. Preliminary results of the simulation show an average hit-rate of 0.52 for the presented simulation and 0.63 with the cache configuration deployed at PIC. Despite not being optimal values, considerations such as optimizing the size of the cache, the introduction of data in fallback (not found on the site) and improving the identification of the most popular data will improve the predicted results. These results cannot yet be compared with the actual cache performance because the local service monitoring tasks are not yet finished. However, these simulations could help us to understand if these popular datasets used for analysis could only be served from local XCache systems at both PIC and CIEMAT, or from a unique cache service serving data to the whole region. The XCache service brings popular data close to compute nodes, hence the overall CPU efficiency of underlying jobs improves. This could also be quantified with dedicated and controlled analysis jobs that read from remote sites or local XCaches, a task that is currently being investigated by the authors.

References

[1] WLCG project, Consulted on 6th Jun of 2016: http://wlcg.web.cern.ch/

[2] J. Albrecht, et al, "A Roadmap for HEP Software and Computing RD for the 2020s", Computing and Software for Big Science volume 3, Article number: 7 (2019) https://doi.org/10.1007/s41781-018-0018-8

[3] X. Espinal, et al, ”The Quest to solve the HL-LHC data access puzzle. The first year of the DOMA ACCESS Working Group”, International Conference on Computing in High Energy and Nuclear Physics (CHEP), Adelaide, Australia, 4-8 november 2019, viewed 5 November 2019

[4] S. Chatrchyan et al. “The CMS Experiment at the CERN LHC”. In: JINST 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004

[5] D. Ciangottini et.al, (2019). “Integration of the Italian cache federation within the CMS computing model.” 014. 10.22323/1.351.0014.

[6] Edgar Fajardo, Matevz Tadel, Justas Balcas, Alja Tadel, Frank Würthwein, Diego Davila, Jonathan Guiang, Igor Sfiligoi “Moving the California distributed CMS XCache from bare metal into containers using Kubernetes”. EPJ Web Conf. 245 04042 (2020). DOI: 10.1051/epjconf/202024504042.

[7] 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.

[8] 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. 10.1088/1742-6596/1085/3/032055.

[9] K. Bloom et al. “Any Data, Any Time, Anywhere: Global Data Access for Science”, arXiv:1508.01443 [physics.comp-ph]

[10] C. Acosta-Silva, A. Delgado Peris, J. Flix, J. M. Guerrero, J. M. Hernández, A. Pérez-Calero Yzquierdo, F. J. Rodriguez Calonge, J. Gómez del Pulgar Ruano A 2019 “Lightweight site federation for CMS support” , International Conference on Computing in High Energy and Nuclear Physics (CHEP), Adelaide, Australia, 4-8 november 2019, viewed 3rd June of 2021

[11] Delgado Peris, A., Flix Molina, J., Hernández J., Pérez-Calero Yzquierdo, A., Pérez Dengra, C., Planas, E., Rodríguez Calonge, J., Sikora, A 2019 "CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2", EPJ Web Conf., 245 (2020) 04028

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

Acknowledgements

The authors of this work thanks the PIC and CIEMAT teams for their support in these studies and by deploying novel cache services for the CMS experiment in the Spanish region.

This project is partially funded by the Spanish *Ministerio de Ciencia e Innovación* MINECO, within the project grants with references FPA2016-80994-C2-1-R, PID2019-110942RB-C22, DATA-2020-1-0039, and BES-2017-082665. It has been also supported by the MICIN-Spain under contract PID2020-113614RB-C21 and by the Generalitat de Catalunya GenCat-DIUiE (GRR) project 2017-SGR-313.