

Reference ID: 11226439006\_Gutsche

---

**Reference ID:** 11226439006\_Gutsche

**Submission Date and Time:** 12/16/2019 12:58:11 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

*Consent Statement:* "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

### **Author Names & Affiliations**

Submitting author: Oliver Gutsche - Fermilab

**Additional authors:** For the U.S. CMS S&C Operations program with special acknowledgement to; James Letts, UCSD; Frank Wuerthwein, UCSD; Brian Bockelman, Morgridge Institute; Mike Hildreth, Notre Dame University; Ken Bloom, University of Nebraska–Lincoln

**Contact Email Address** (for NSF use only): (hidden)

### **Research domain(s), discipline(s)/sub-discipline(s)**

High Energy Physics;LHC;HL-LHC

### **Title of Response**

Accessing HL-LHC data from HPC centers, campus infrastructures and commercial clouds for central processing

### **Abstract**

In the HL-LHC era, we estimate more than an Exabyte of new data per year will be recorded and produced. We expect that HPC centers, campus infrastructures and commercial clouds that were not explicitly designed to meet the needs of the LHC and usually don't have co-located disk storage, will provide the bulk of the CPU needed for central production activities of CMS. A transformative change of the central processing infrastructure is needed to move significant volumes of data from archival facilities, where the data is kept mainly on high-latency low cost storage media, to these processing facilities, and archive the output of the processing back to these archival facilities.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

The HL-LHC is expected to record an unprecedented number of proton-proton collisions during its planned run from 2027-2037. Compared to LHC (starting its 3rd running period in 2021 and ending in 2024), the number of events recorded by the CMS detector will increase by almost an order of magnitude, from 7 billion collisions to around 50 billion collisions per data taking year, with similar increases for simulated collisions needed for physics analysis from 15 billion to around 100 billion events. The LHC uses a distributed infrastructure of computing resources interconnected through the R&E networks in Europe, Asia and the U.S. Currently, the CMS infrastructure is comprised of over 250k x86-compatible compute cores, 150 PB of disk, and 300 PB of tape to support central activities to record and simulate proton-proton collisions and perform their subsequent reconstruction to extract physics quantities from the measurements, as well as subsequent analysis activities driven by the physicists of the 2,500 large CMS collaboration. LHC analysis is based on individuals or groups of physicists processing centrally provided reconstructed events. We distinguish "central production" from "community data analysis" in that a small production team produces and curates data for the global CMS community of thousands of scientists. This response to the RFI covers only the central production aspects. For LHC today, central production primarily uses processing facilities that are designed to meet the needs of the LHC experiments. Many other domains, especially in EU and Asia, are using these same facilities, proven the versatility of their design across domains. CPU resources at these facilities are also co-located with disk storage to allow production activities to access input data like the detector data and intermediate data produced during central processing. We are concerned that this may change significantly for HL-LHC. The overall CPU needs will increase exponentially, because the complexity of the proton-proton collisions increases. Extrapolating from today, millions of x86-compatible CPU cores will be needed. In the HL-LHC era, we expect that HPC centers, campus infrastructures and commercial clouds that were not explicitly designed to meet the needs of the LHC and usually don't have co-located disk storage, will provide the bulk of the CPU needed for central production activities of CMS. Efficient and performant data access requires a transformational change in how these facilities are integrated into the central LHC processing workflows.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).**

Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

CMS uses a hierarchy of data formats from comprehensive raw detector data (RAW) and derived analysis object data (AOD) to refined analysis data in reduced size (MINI and NANO). During HL-LHC the data size per collision is expected to range from ~7MB for RAW to ~4kB for NANO. Processing time is dominated by the reconstruction of RAW to AOD, and corresponding simulations. MINI is produced from AOD, and NANO from MINI. The more refined data formats are expected to be remade more often to adjust to improvements in physics object definitions, and to significantly reduce the need of the general CMS scientific community to access AOD and perform all analyses using MINI and NANO. As a result, RAW data is accessed only as an organized production activity by a small production team while MINI and NANO are accessed by the entire global CMS community of a few thousand scientists. The RAW data volume of CMS is estimated to be around 500 PB per data taking year in the HL-LHC era. A full reconstruction pass is planned at the end of each data taking year, applying the best calibrations of the over-time evolving understanding of detector and data taking conditions for that year. This end-of-year pass is desired to take no more than two months, which means processing about 8 PB of data per day and producing 2-3PB of output per day. Facilities used for this reconstruction pass would have to be able to ingest the RAW data and return the reconstruction outputs for archiving and distribution for analysis. These resources would also have to support simulation activities during the year, which don't need access to RAW data, but still have high input data requirements due to the need to simulate parasitic collisions of the HL-LHC running period (called PileUp). The same output requirements as for data reconstruction apply. In totality, we estimate more than an Exabyte of new data per year during the HL-LHC era. An infrastructure supporting the central production needs of CMS would have to be able to access a diverse set of shared CPU resources like HPC centers, campus infrastructures and commercial clouds. The infrastructure would have to be able to move significant volumes of data from archival facilities, where the data is kept mainly on high-latency low cost storage media, to these processing facilities. The infrastructure has to orchestrate the execution of the CMS reconstruction (and simulation) workflows taking into account specialized hardware architectures provided by the facilities. At the end of the workflows, the infrastructure will have to archive the output of the processing and provide access to it for analysis. This problem is not unique to the LHC, as many other experiments and disciplines share the same characteristics of moderate to low data to CPU ratios. There is a lot of potential to share solutions and infrastructures.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

-- End Submission --