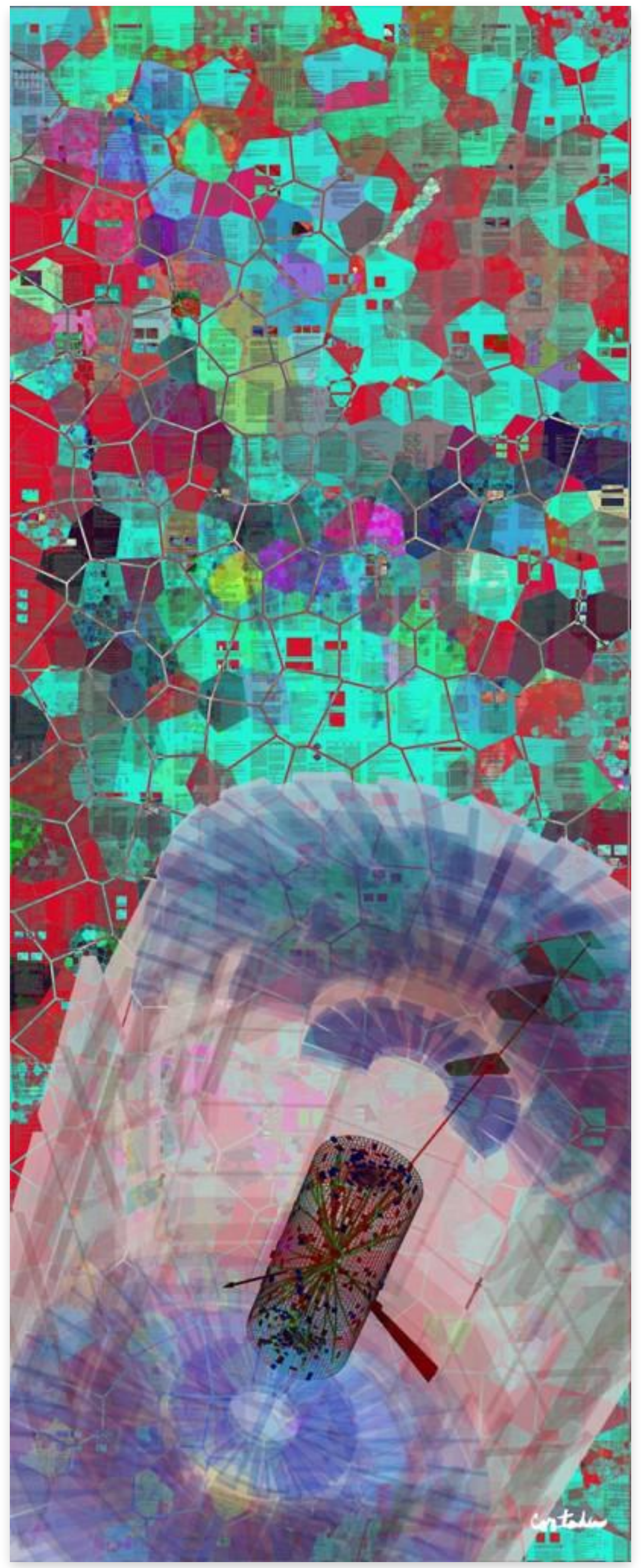
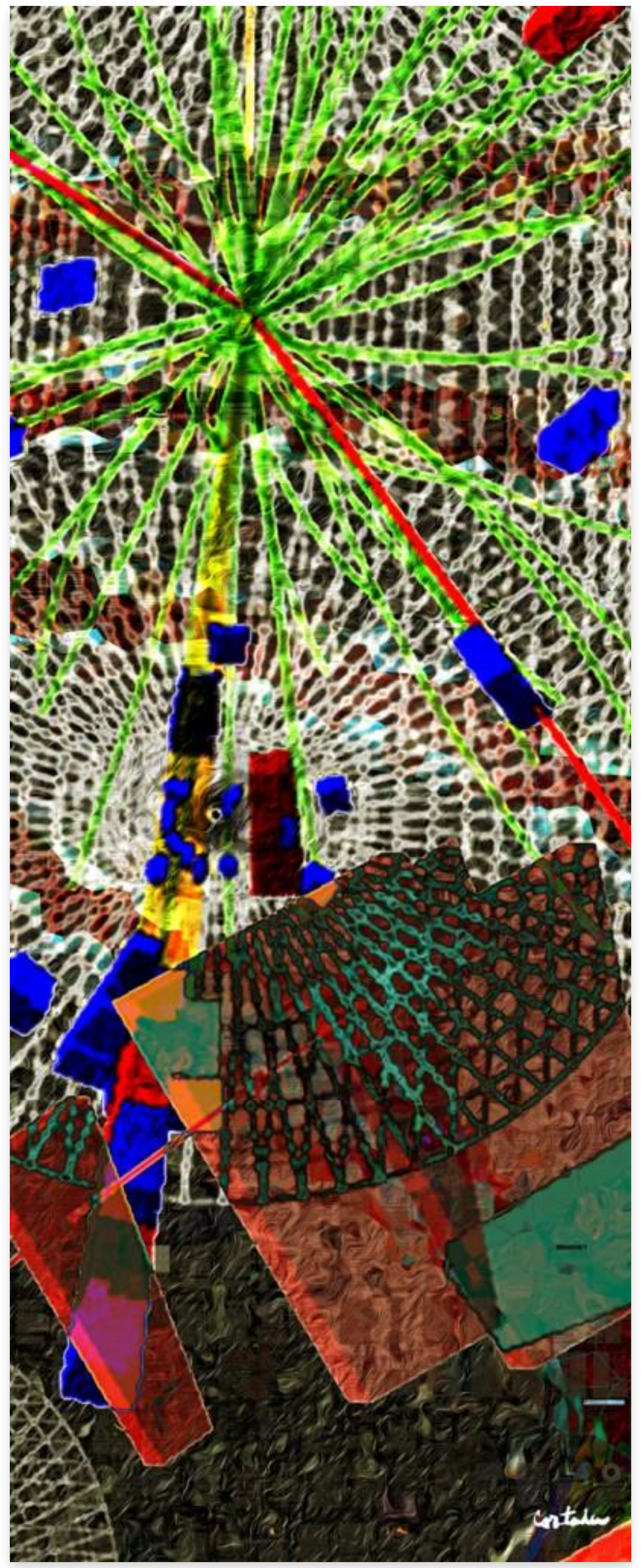
U.S. CMS Operations Program

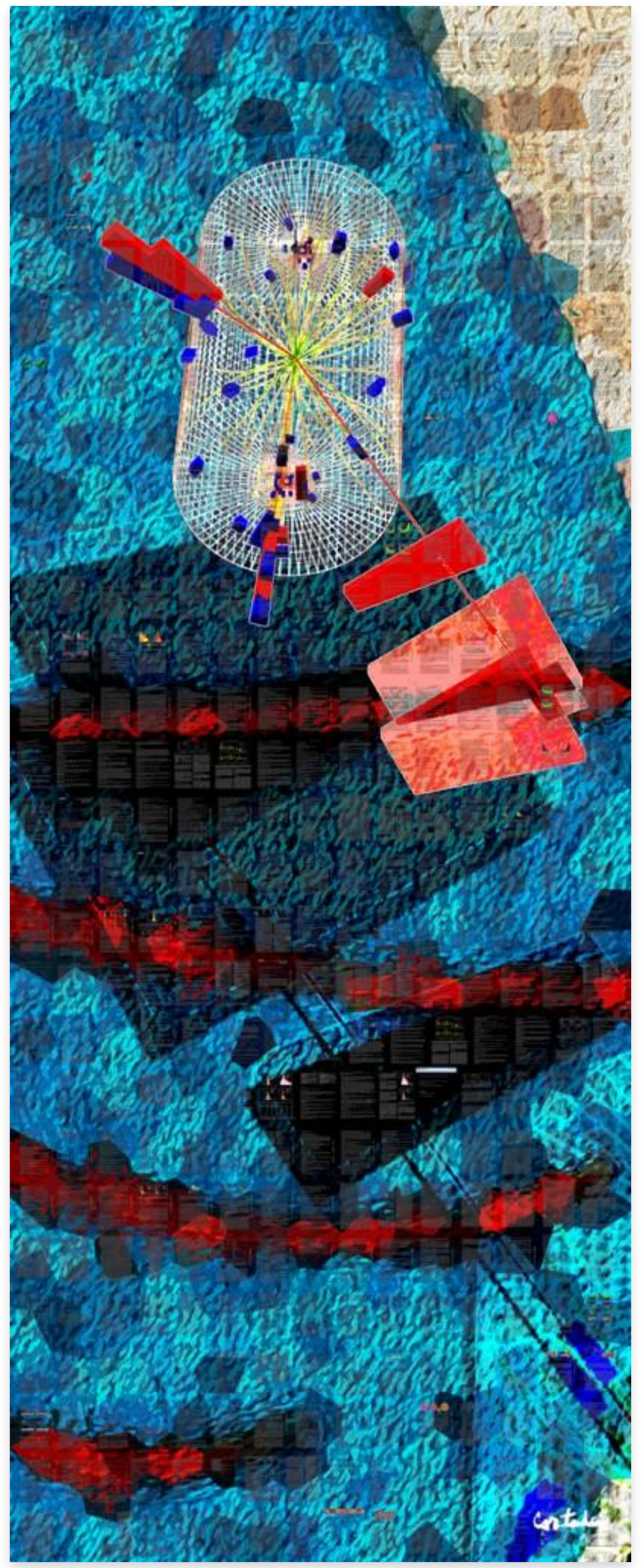
The U.S. CMS HL-LHC R&D Strategic Plan

Oliver Gutsche

26th International Conference on Computing in High Energy Physics (CHEP2023)

8. May 2023




CMS and U.S. CMS

3394	1102	282	247	57
PHYSICISTS (1228 STUDENTS)	ENGINEERS	TECHNICIANS	INSTITUTES	COUNTRIES & REGIONS

The Details

The CMS collaboration has around:

6288
ACTIVE PEOPLE
(PHYSICISTS, ENGINEERS, TECHNICAL, ADMINISTRATIVE, STUDENTS, ETC.)

Of these members there are about:

2166	1228	1102	1388
PHD PHYSICISTS (1769 MEN, 397 WOMEN)	PHYSICS DOCTORAL STUDENTS (919 MEN, 309 WOMEN)	ENGINEERS (951 MEN, 151 WOMEN)	UNDERGRADUATES (995 MEN, 393 WOMEN)

A typical CMS physics paper will be signed by the PhD physicists and a significant fraction of the doctoral students meaning it will typically have about 2100 signatures.

** As of October 2022 https://cms.cern/index.php/collaboration/people-statistics

- CMS collaboration
  - Over 200 Institutes from 57 countries
  - Over 3000 physicists with ⅔ having authorship privileges
- CMS experiment
  - Detector with multiple sub-components to record particle collisions
  - Over 1000 publications so far
  - Enabled by availability of sufficient computing resources to Store, Process and Analyze the Data
- U.S. CMS
  - 30% of the authors from ~50 U.S. institutions
  - U.S. researchers involved in physics and construction of major detector components
- U.S. CMS Operations Program
  - Centrally funded by DOE and NSF
    - Operate and maintain the U.S. detector components
    - Software and Computing needs of CMS

U.S. CMS Software & Computing Operations Program

Mission of the U.S. CMS Software & Computing Operations Program: “to develop and operate the software and computing resources necessary to process CMS data expeditiously and to enable U.S. physicists to fully participate in the physics of CMS.”

- Enable U.S. physicists to take on a leadership role in CMS physics
- Provide necessary resources (computing, personnel) for processing, storing, and analyzing collision and simulated data samples in a timely manner
- Develop and maintain performant software

U.S. CMS has major responsibilities for CMS Offline & Computing (CMS O&C)

- U.S. Contributions to CMS distributed computing infrastructure
  - Hardware resources and their administration/operation
- Production software
  - Infrastructure software like data management, workflow management, metadata bookkeeping, etc.
  - Core applications like the CMS Software Framework CMSSW
- Operations support for CMS
  - Contributions to operate services and distributed site infrastructure @CERN and in the U.S.
- **Forward looking R&D to enable HL-LHC**

High Luminosity LHC (HL-LHC)

High Luminosity LHC (HL-LHC)

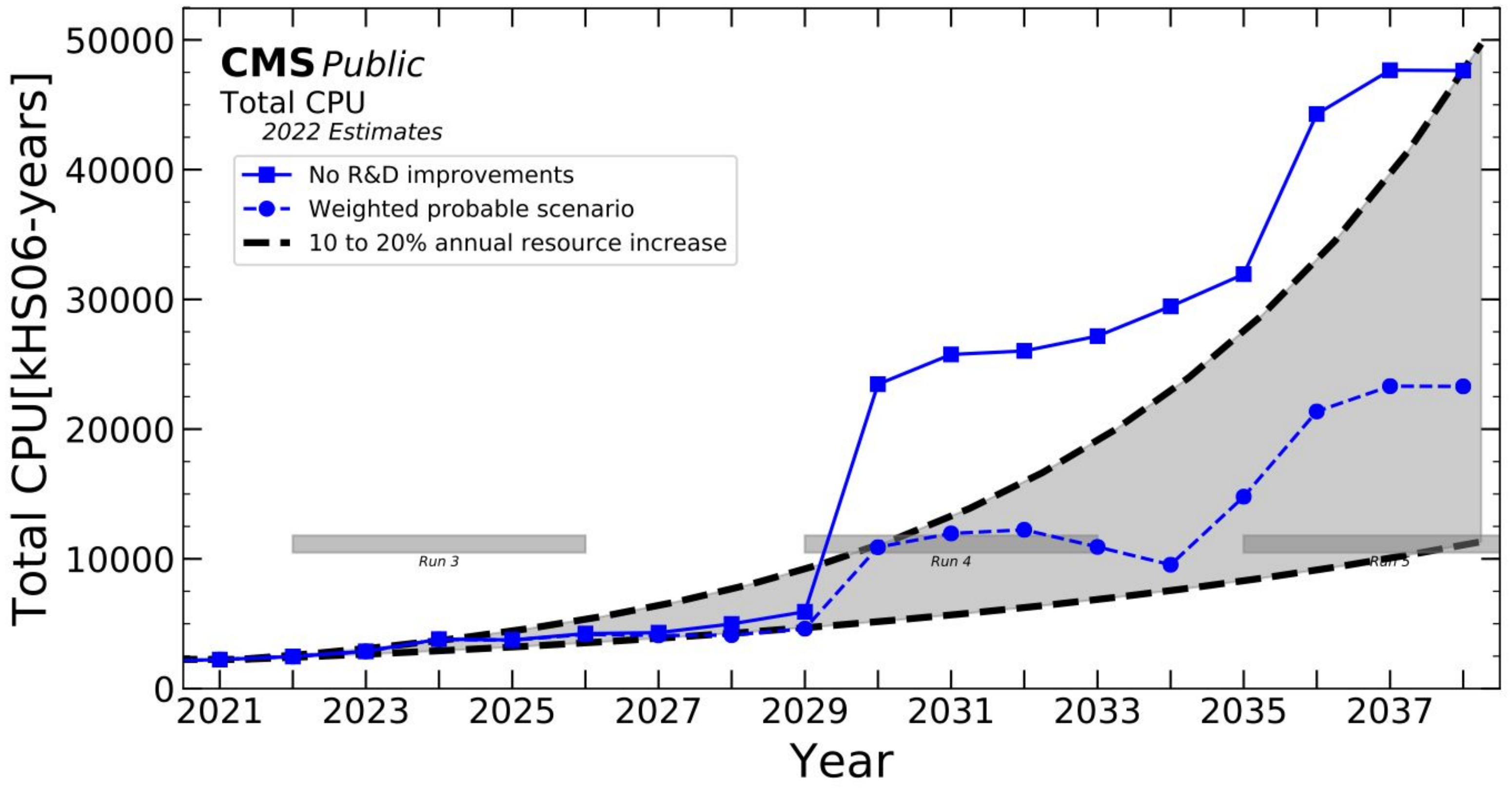
- Next phase of the science harvest @ CERN: 2029-2042
- Over 90% of the Integrated Luminosity of the LHC
- Higher Intensity Proton-proton collisions
- New CMS detector components with higher granularity and more channels

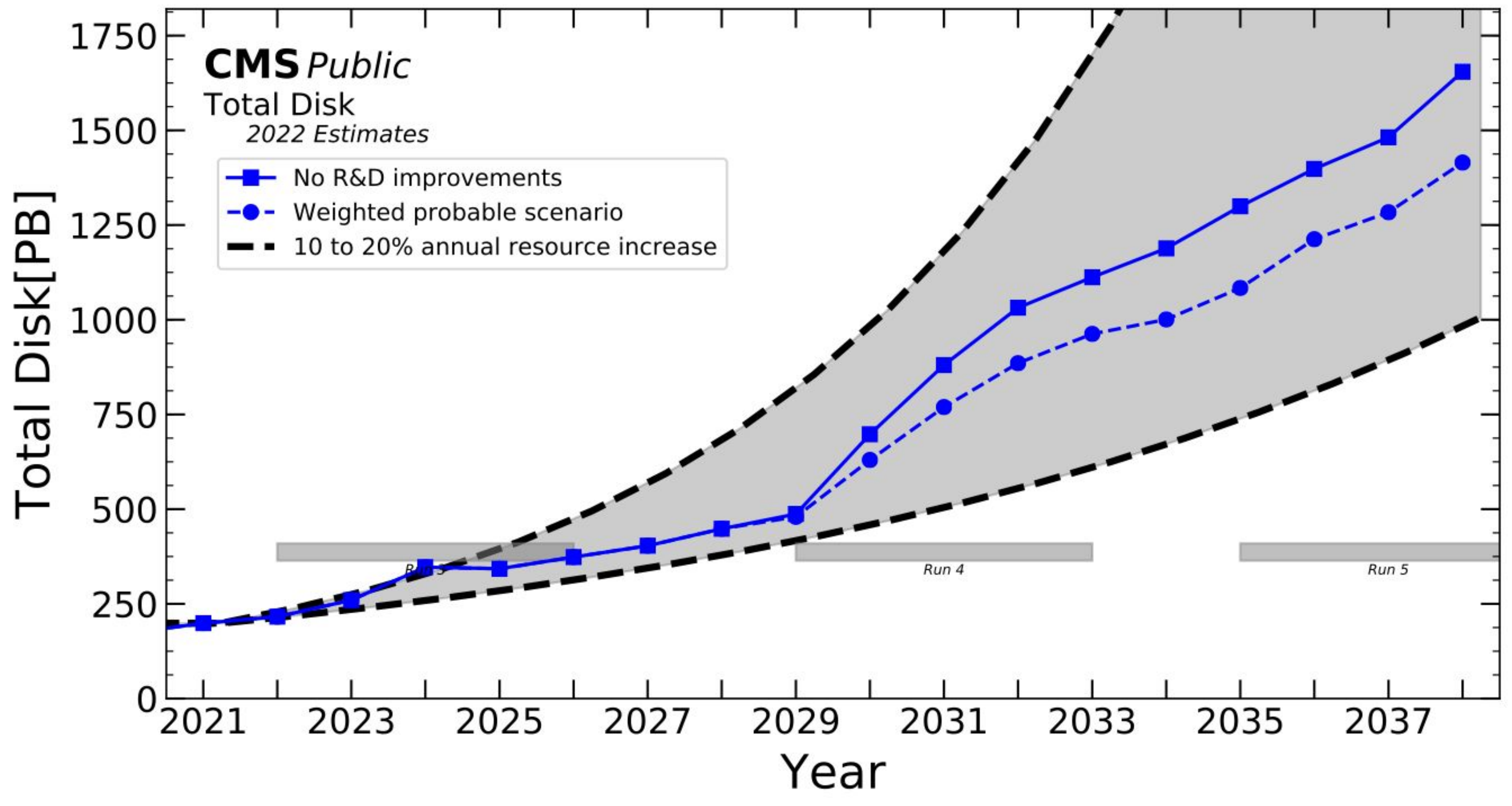
Unprecedented Computing Needs compared to today(*)

- Number of events to be processed each year larger by x3: 150 Billion events
- Size per event larger by x5: disk storage needs reaches ½ exabyte by 2030
- Most data is active: needs to be held on quasi-randomly accessible storage systems to be processed by hundreds of simultaneous processing pipelines
- Physics software: more than 4 million lines of highly specialized code with high algorithmic complexity and low computational intensity ➜ providing unique challenges to use accelerators.

(*) based on input from [“The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger” TDR]

Latest extrapolated resource needs of CMS for HL-LHC, based on an updated computing model described in https://cds.cern.ch/record/2815292





https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults

U.S. CMS HL-LHC R&D Strategic Plan

- To do physics in HL-LHC, we need to
  - Archive multiple-hundreds of PB of RAW data on tape
  - Process all the events and produce even more simulations
  - Utilize accelerators and advanced computing architectures efficiently
  - Integrate AI/ML on unprecedented scale
  - Provide access for analysis: more events analyzed in less time ➜ high statistics analyses

- The U.S. CMS S&C Operations Program defined 4 “Grand Challenges” (GC) that are tackling high priority areas and are embedded in the overall CMS effort:

  (1) Modernize physics software and improve algorithms
  (2) Build infrastructure for exabyte-scale datasets
  (3) Transform the scientific data analysis process
  (4) Transition from R&D to operations

(1) Modernize physics software and improve algorithms

Exploit novel algorithms, including ML/AI, reduce algorithmic complexity, increase computational intensity, and provide core software infrastructure to enable effective use of modern hardware and accelerators.

Work packages:

- Core Software Framework and Software Portability
- Establish Performance Metric and Performance Baseline for Physics Software
- U.S. Contributions to the Charged Particle Tracking Software
- U.S. Contributions to Software for High Granularity Calorimeter
- U.S. Contributions to CMS Advanced Algorithms Work

(2) Build infrastructure for exabyte-scale datasets

Build infrastructure to archive, store, transfer, and provide access to exabyte-scale datasets. Explore data lakes and custodial storage: establish a technology and cost model for custodial/archival storage facilities which manages operations costs, and optimizes hardware costs. Orchestrate computational services and data access, provide intelligent network services, and master the integration of national and international computing resources, including High Performance Computing centers (HPCs).

Work packages:

- Produce exabyte-scale central datasets
- Manage exabyte-scale network flows
- Store exabyte-scale central datasets
- Derive petabyte-size analysis datasets
- Provide access to petabyte-size analysis datasets
- Operate facilities to support production, storage and transfer of exabyte-scale datasets

(3) Transform the scientific data analysis process

Re-imagine how physicists interact with data when extracting science: seize on industry advances in data science, consider required performance parameters like throughput, latency, ease of use, and functionality, and develop the facility and software infrastructure to support thousands of physicists analyzing exabytes of data.

Work packages:

- Provide Column-Wise Analysis Software Infrastructure
- Provide Column-Wise Analysis Facility Infrastructure
- Provide Column-Wise Data Augmentation Infrastructure

(4) Transition from R&D to operations

The R&D program will contribute to several advances in infrastructure particularly in analysis facilities and networking/storage that will need to be rolled into operations. This transition needs to include adequate testing, monitoring, and training/documentation for a successful adoption by the operational staff.

Work packages:

- Develop sufficient monitoring capabilities
- Assess readiness of systems/services for HL-LHC for transition to user testing and operation
- Write Documentation for systems/services
- Provide training for operators and users

U.S. and International Partners

- U.S. CMS is part of the community's ecosystem for computing and software related research and developments.
  - Research partnerships
    - Host National Lab: Fermilab
    - 7 U.S. Tier-2 institutes and additional U.S. institutes
    - Other National Labs
    - CERN
  - National and international consortia
    - Open Science Grid (OSG)
    - HEP Software Foundation (HSF)
  - Joint and collaborative projects
    - IRIS-HEP
    - HEP-CCE
  - Community efforts
    - Joint Blueprint activities with U.S. ATLAS, OSG, ESnet, and IRIS-HEP Snowmass Computational Frontier

Summary & Conclusions

- U.S. CMS has established a HL-LHC Software & Computing R&D program
  - to provide the R&D necessary to plan for and build the computing services required for full scientific exploitation of the CMS HL-LHC upgrades
  - in partnership with its host lab Fermilab, its 7 Tier-2 institutes and the U.S.~CMS research and S\&C communities
- The program addresses the four Grand Challenges of HL-LHC software and computing:
  - (1) Modernizing Physics Software and Improving Algorithms
  - (2) Building Infrastructure for Exabyte-Scale Datasets
  - (3) Transforming the Scientific Data Analysis Process
  - (4) Transition from R&D to Operations
- Many of the projects addressing the challenges have presentations and status updated this week at CHEP2023.