Use case for the Compact Muon Solenoid at U.S. high performance computing facilities

Particle physics is the study of the most elementary components of matter and their interactions. The key experimental tools of the field are particle accelerators, which bring everyday particles such as protons into collisions at energy scales that are equivalent to the temperature of the very early universe. The Large Hadron Collider (LHC) is currently the world's highest-energy accelerator, colliding protons and creating heavy particles that existed in the early universe but are rarely observed in our world today. The Compact Muon Solenoid (CMS) is one of the two general-purpose particle physics detectors at the LHC. The CMS Collaboration co-discovered the Higgs boson in 2012, has provided constraints on many models of new physics, and has made precise measurements of the properties of known particles. CMS is currently analyzing data recorded from 2016 through 2018, with an emphasis on searches for dark matter and supersymmetric particles and measurements of the properties of the Higgs boson. The full suite of CMS physics measurements requires the simulation of billions of proton-proton collision events, including both processes that would arise from new, speculative physics models and those that arise from standard model processes that would be the background to the new physics. In addition, the collaboration is planning for the operation of the High Luminosity LHC (HL-LHC) in 2026. The HL-LHC will operate at much higher beam intensities than the LHC, leading to more complex collisions, and CMS will record data at a rate an order of magnitude larger than it did in 2018. CMS seeks to use all manner of resources to meet the corresponding computing challenges to pursue groundbreaking discoveries.

The CMS experiment has identified so-called “production” workflows as good fits for current and future High Performance Computing facilities. These workflows are centrally planned and managed by a small core team of experienced operators (as opposed to “analysis” workflows which are activities carried out by dozens of disparate teams of scientists simultaneously).

The production workflows are divided (at a roughly 1:4 ratio) between the processing of experimental data from the detector, and the generation and processing of simulated data in a Monte Carlo (MC) model. The data are naturally divided into uncoupled chunks called “events”, which are suitable for perfectly parallelized computing. In order to optimize memory usage, events are processed by a multithreaded framework (typically utilizing 8-16 cores, requiring at most 500 MB of memory per core). Both workflows are mostly CPU-bound.

Data processing workflows read the raw data from the CMS detector as input, reconstruct physics quantities in a compute-intensive step, and write output data in a reduced format. For simulation workflows, data is created by generating events and simulating the physical detector response; reading input from a specialized dataset (“pile-up”); reconstructing physics quantities; and writing output data. Both workflows will consume input data at ~500 kbit/s per core and

write output data at ~50 kbit/s per core. Because the data are processed event by event, the experiment has the ability to arbitrarily tune the length of job by scaling the number of events that are serially processed. The I/O rates are an estimate based on current Intel Xeon cores, and are projected to scale linearly with CPU performance.

The experiment has invested significant effort in time in developing, validating, and optimizing the software framework and processing code on the x86 architecture. The capability for execution on other architectures (i.e. PowerPC) and offloading work to co-processors/accelerators (i.e. GPUs) has been developed, but requires significant further development. We also expect new, different workflows that use accelerators for the training of machine learning algorithms that are currently under investigation in the community.

Storage requirements depend on the execution scale and connectivity (both local and wide-area) of the compute farm. The processing is capable of streaming data from a diverse set of remote experiment-owned sites over the wide-area network, but a large workflow running in parallel on HPC facilities has been shown to saturate intermediate network translation layers and peer interconnects. Local storage of 500-1000 TB should be set aside for caching the pile-up dataset at the facility if streaming is not feasible.

The computing environment is based on CentOS 6/7. This environment can be provided natively or with Singularity or Shifter containers. The experiment software is published using CVMFS[1] - mounting the software via a FUSE mount (pointing to a local web cache) is preferred, but some HPC facilities operate NFS servers with periodically synchronized snapshot of the software.

[1] https://cvmfs.readthedocs.io/en/stable/