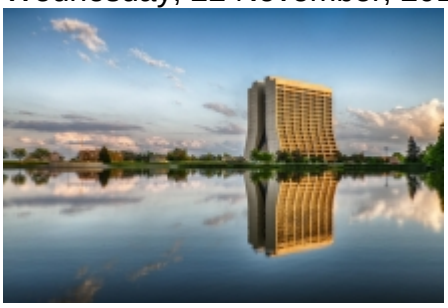
Published on *CERN openlab* (http://test-static-05.web.cern.ch)

Fermilab joins CERN openlab, works on ?data reduction? project with CMS experiment [1]

Wednesday, 22 November, 2017



[2]

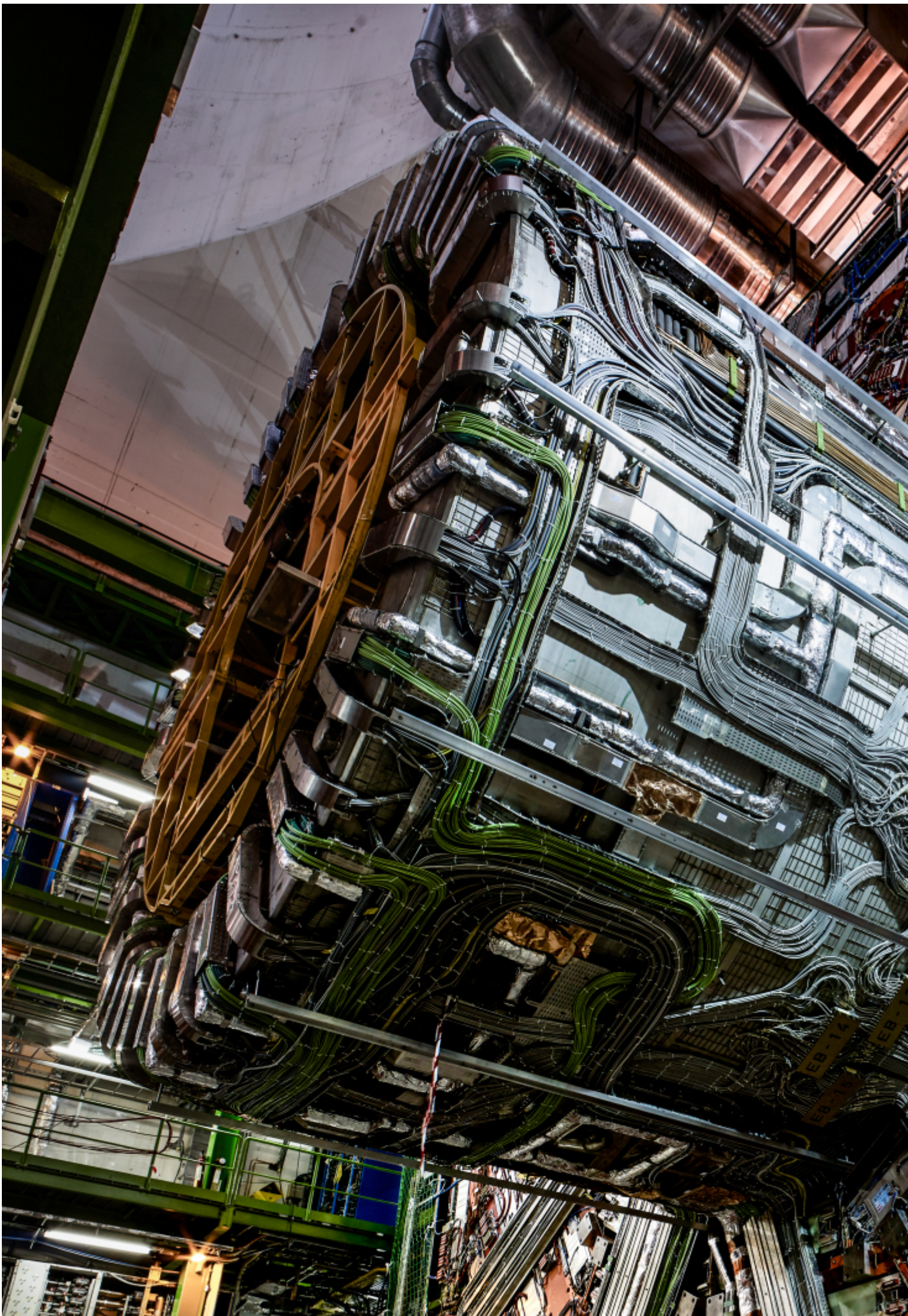
Fermilab, the USA?s premier particle physics and accelerator laboratory, has joined CERN openlab [3] as a research member. Researchers from the laboratory will collaborate with members of the CMS experiment [4] and the CERN IT Department [5] on efforts to improve technologies related to ?physics data reduction?. This work will take place within the framework of an existing CERN openlab project with Intel [6] on ?big-data analytics?.

?Physics data reduction? plays a vital role in ensuring researchers are able to gain valuable insights from the vast amounts of particle-collision data produced by high-energy physics experiments, such as the CMS experiment [7] on CERN?s Large Hadron Collider (LHC) [8]. The project?s goal is to develop a new system ? using industry-standard big-data tools ? for filtering many petabytes of heterogeneous collision data to create manageable, but rich, datasets of a few terabytes for analysis. Using current systems, this kind of targeted data reduction can often take weeks; but the aim of the project is to be able to achieve this in a matter of hours.

?Time is critical in analysing the ever-increasing volumes of LHC data,?says Oliver Gutsche, a Fermilab [9] scientist working at the CMS experiment. ?I am excited about the prospects CERN openlab brings to the table: systems that could enable us to perform analysis much faster and with much less effort and resources.? Gutsche and his colleagues will explore methods of ensuring efficient access to the data from the experiment. For this, they will investigate techniques based on Apache Spark, a popular open-source software platform for distributed processing of very large data sets on computer clusters built from commodity

hardware. "The success of this project will have a large impact on the way analysis is conducted, allowing more optimised results to be produced in far less time,? says Matteo Cremonesi, a research associate at Fermilab. "I am really looking forward to using the new open-source tools; they will be a game changer for the overall scientific process in high-energy physics."

The team plans to first create a prototype of the system, capable of processing 1 PB of data with about 1000 computer cores. Based on current projections, this is about 1/20th of the scale of the final system that would be needed to handle the data produced when the High-Luminosity LHC [10] comes online in 2026. Using this prototype, it should be possible to produce a benchmark (or ?reference workload?) that can be used evaluate the optimum configuration of both hardware and software for the data-reduction system.


EB-14