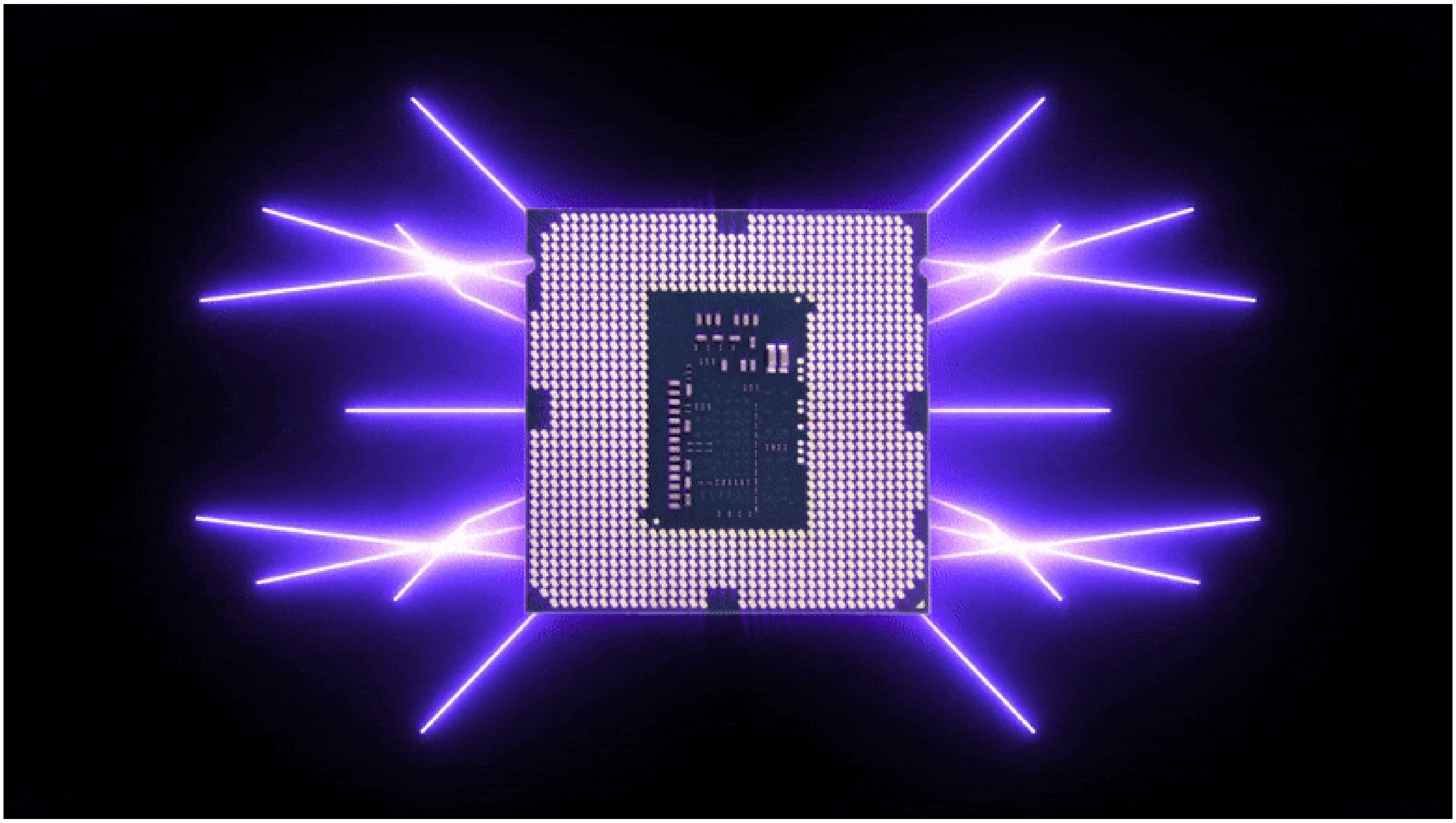
Large Hadron Collider Experiments Step Up the Data Processing Game With GPUs

TOPICS: CERN Large Hadron Collider

By CERN FEBRUARY 7, 2022



While data processing demand is rocketing for LHC's Run 3, the four large experiments are increasing their use of GPUs to improve their computing infrastructure.

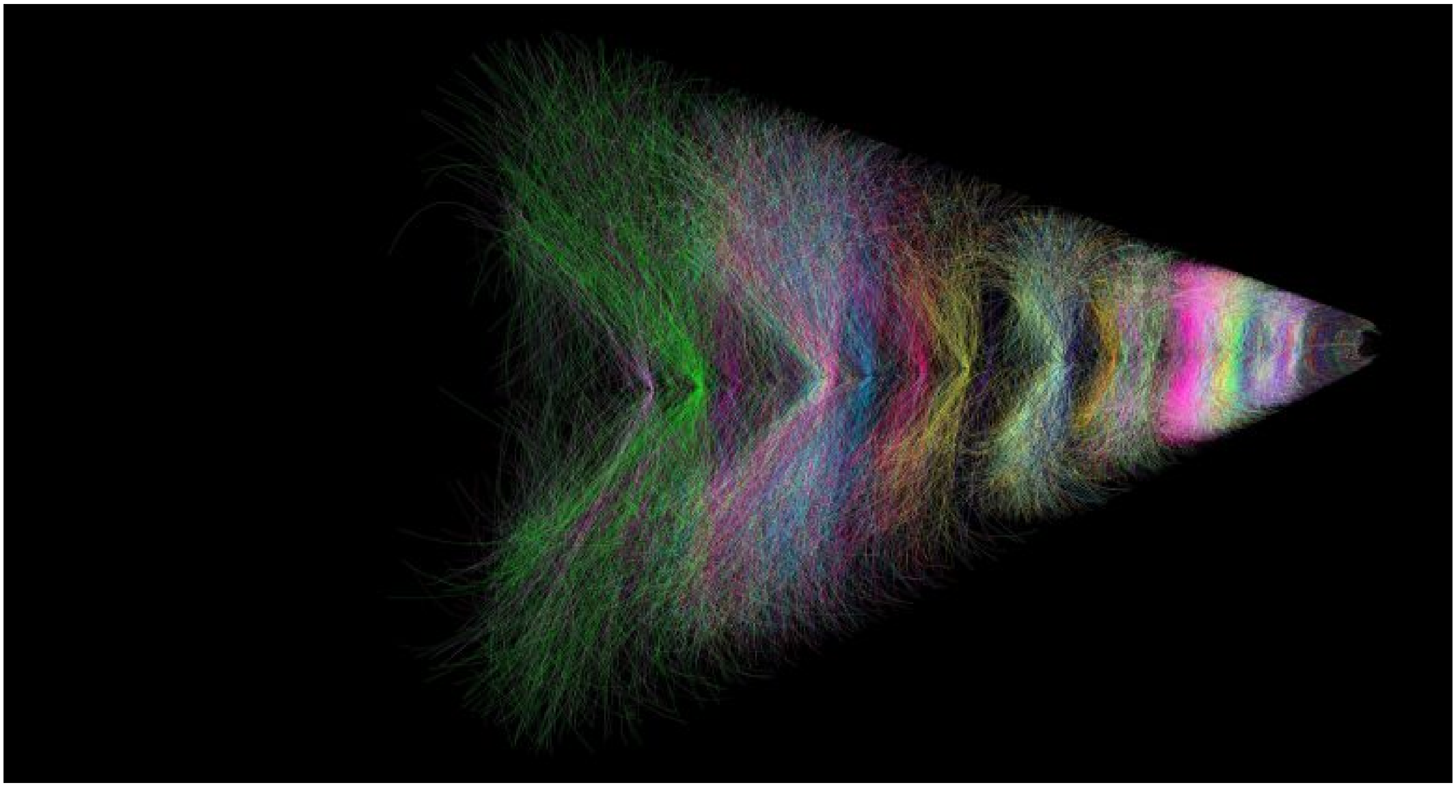
Analyzing as many as one billion proton collisions per second or tens of thousands of very complex lead collisions is not an easy job for a traditional computer farm. With the latest upgrades of the LHC experiments due to come into action next year, their demand for data processing potential has significantly increased. As their new computational challenges might not be met using traditional central processing units (CPUs), the four large experiments are adopting graphics processing units (GPUs).

GPUs are highly efficient processors, specialized in image processing, and were originally designed to accelerate the rendering of three-dimensional computer graphics. Their use has been studied in the past couple of years by the LHC experiments, the **Worldwide LHC Computing Grid** (WLCG), and **CERN openlab**. Increasing the use of GPUs in high-energy physics will improve not only the quality and size of the computing infrastructure, but also the overall energy efficiency.

A candidate HLT node for Run 3, equipped with two AMD Milan 64-core CPUs and two NVIDIA Tesla T4 GPUs. Credit: CERN)

“The LHC’s ambitious upgrade program poses a range of exciting computing challenges; GPUs can play an important role in supporting machine-learning approaches to tackling many of these,” says Enrica Porcari, Head of the CERN IT department. “Since 2020, the CERN IT department has provided access to GPU platforms in the data center, which have proven popular for a range of applications. On top of this, CERN openlab is carrying out important investigations into the use of GPUs for machine learning through collaborative R&D projects with industry, and the Scientific Computing Collaborations group is working to help port – and optimize – key code from the experiments.”

ALICE has pioneered the use of GPUs in its high-level trigger online computer farm (HLT) since 2010 and is the only experiment using them to such a large extent to date. The newly upgraded ALICE detector has more than 12 billion electronic sensor elements that are read out continuously, creating a data stream of more than 3.5 terabytes per second. After first-level data processing, there remains a stream of up to 600 gigabytes per second. These data are analyzed online on a high-performance computer farm, implementing 250 nodes, each equipped with eight GPUs and two 32-core CPUs. Most of the software that assembles individual particle detector signals into particle trajectories (event reconstruction) has been adapted to work on GPUs.



Visualisation of a 2 ms time frame of Pb-Pb collisions at a 50 kHz interaction rate in the ALICE TPC. Tracks from different primary collisions are shown in different colors. Credit: ALICE/CERN

In particular, the GPU-based online reconstruction and compression of the data from the Time Projection Chamber, which is the largest contributor to the data size, allows ALICE to further reduce the rate to a maximum of 100 gigabytes per second before writing the data to the disk. Without GPUs, about eight times as many servers of the same type and other resources would be required to handle the online processing of lead collision data at a 50 kHz interaction rate.

ALICE successfully employed online reconstruction on GPUs during the LHC pilot beam data taking at the end of October 2021. When there is no beam in the LHC, the online computer farm is used for offline reconstruction. In order to leverage the full potential of the GPUs, the full ALICE reconstruction software has been implemented with GPU support, and more than 80% of the reconstruction workload will be able to run on the GPUs.

From 2013 onwards, LHCb researchers carried out R&D work into the use of parallel computing architectures, most notably GPUs, to replace parts of the processing that would traditionally happen on CPUs. This work culminated in the **Allen project**, a complete first-level real-time processing implemented entirely on GPUs, which is able to deal with LHCb’s data rate using only around 200 GPU cards. Allen allows LHCb to find charged particle trajectories from the very beginning of the real-time processing, which are used to reduce the data rate by a factor of 30–60 before the detector is aligned and calibrated and a more complete CPU-based full detector reconstruction is executed. Such a compact system also leads to substantial energy efficiency savings.

Starting in 2022, the LHCb experiment will process 4 terabytes of data per second in real time, selecting 10 gigabytes of the most interesting LHC collisions each second for physics analysis. LHCb's unique approach is that instead of offloading work, it will analyze the full 30 million particle-bunch crossings per second on GPUs.

Together with improvements to its CPU processing, LHCb has also gained almost a factor of 20 in the energy efficiency of its detector reconstruction since 2018. LHCb researchers are now looking forward to commissioning this new system with the first data of 2022, and building on it to enable the full physics potential of the upgraded LHCb detector to be realized.

CMS reconstructed LHC collision data with GPUs for the first time during the LHC pilot beams in October last year. During the first two runs of the LHC, the CMS HLT ran on a traditional computer farm comprising over 30 000 CPU cores. However, as the **studies for the Phase 2 upgrade of CMS** have shown, the use of GPUs will be instrumental in keeping the cost, size, and power consumption of the HLT farm under control at higher LHC luminosity. And in order to gain experience with a heterogeneous farm and the use of GPUs in a production environment, CMS will equip the whole HLT with GPUs from the start of Run 3: the new farm will be comprised of a total of 25 600 CPU cores and 400 GPUs.

The additional computing power provided by these GPUs will allow CMS not only to improve the quality of the online reconstruction but also to extend its physics program, running the online **data scouting analysis** at a much higher rate than before. Today about 30% of the HLT processing can be offloaded to GPUs: the calorimeters local reconstruction, the pixel tracker local reconstruction, the pixel-only track and vertex reconstruction. The number of algorithms that can run on GPUs will grow during Run 3, as other components are already under development.

ATLAS is engaged in a variety of R&D projects towards the use of GPUs both in the online trigger system and more broadly in the experiment. GPUs are already used in many analyses; they are particularly useful for machine learning applications where training can be done much more quickly. Outside of machine learning, ATLAS R&D efforts have focused on improving the software infrastructure in order to be able to make use of GPUs or other more exotic processors that might become available in a few years. A few complete applications, including a fast calorimeter simulation, also now run on GPUs, which will provide the key examples with which to test the infrastructure improvements.

"All these developments are occurring against a backdrop of unprecedented evolution and diversification of computing hardware. The skills and techniques developed by CERN researchers while learning how to best utilize GPUs are the perfect platform from which to master the architectures of tomorrow and use them to maximize the physics potential of current and future experiments," says Vladimir Gligorov, who leads LHCb's Real Time Analysis project.