Contribution ID: **125** Type: **Parallel talk**

Harnessing AI and ML Innovations for High-Luminosity LHC: Transitioning from R&D to Production

With the upcoming High-Luminosity Large Hadron Collider (HL-LHC) and the corresponding increase in collision rates and pile-up, a significant surge in data quantity and complexity is expected. In response, substantial R&D efforts in artificial intelligence (AI) and machine learning (ML) have been initiated by the community in recent years to develop faster and more efficient algorithms capable of managing this deluge of data. Several projects focused on triggering and offline reconstruction are currently underway. These initiatives have demonstrated highly promising results, offering physics performance comparable to existing algorithms but with the ability to run on modern hardware architectures ("not just CPU") with substantially reduced computation times. As we approach the start of the HL-LHC, the time is ripe to transition these models from the R&D phase to prototyping and eventually to full-scale deployment within the production systems of the experiments. This presentation will discuss ongoing integration efforts that address new questions and challenges within an environment with high throughput and stringent constraints on timing, energy consumption and total cost per collision event. These include both the high-level trigger and offline reconstruction, aiming for the acceleration of model inferences and the execution of complex pipelines on hybrid heterogeneous CPU/GPU and CPU/inference server architectures. Successful integration of these advances will capitalise on recent AI and ML R&D efforts, helping experiments to efficiently process their data during the HL-LHC era.

AI keywords

deployment; fast inference; high throughput; production; heterogeneous computing environments

Primary authors: BISCARAT, Catherine (L2I Toulouse, CNRS/IN2P3, Université de Toulouse); STARK, Jan (L2I Toulouse, CNRS/IN2P3, Université de Toulouse); CAILLOU, Sylvain (L2I Toulouse, CNRS/IN2P3, Université de Toulouse)

Presenter: STARK, Jan (L2I Toulouse, CNRS/IN2P3, Université de Toulouse)

Track Classification: Inference & Uncertainty