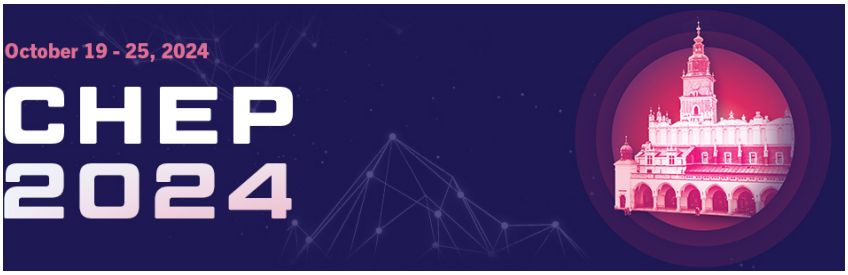

Contribution ID: **178** Contribution code: **TUE 28** Type: **Poster**

Deployment of inference as a service at the US CMS Tier-2 data centers

Tuesday 22 October 2024 16:00 (15 minutes)

Coprocessors, especially GPUs, will be a vital ingredient of data production workflows at the HL-LHC. At CMS, the GPU-as-a-service approach for production workflows is implemented by the SONIC project (Services for Optimized Network Inference on Coprocessors). SONIC provides a mechanism for outsourcing computationally demanding algorithms, such as neural network inference, to remote servers, where requests from multiple clients are intelligently distributed across multiple GPUs by a load-balancing service. This talk highlights the recent progress in deploying SONIC at selected U.S. CMS Tier-2 data centers. Using realistic CMS Run3 data processing workflows, such as those containing transformer-based algorithms, we demonstrate how SONIC is integrated into the production-like environment to enable accelerated inference offloading. We will present developments from both the client and server sides, including production job and data center configurations for NVIDIA and AMD GPUs. We will also present performance scaling benchmarks and discuss the challenges of operating SONIC in CMS production, such as server discovery, GPU saturation, fallback server logic, etc.

Primary authors: COLLABORATION, CMS; PEDRO, Kevin (Fermi National Accelerator Lab. (US))

Presenter: PEDRO, Kevin (Fermi National Accelerator Lab. (US))

Session Classification: Poster session

Track Classification: Track 4 - Distributed Computing