Contribution ID: **138** Type: **Standard 15 min talk**

SONIC: A Portable framework for as-a-service ML serving

Tuesday, October 15, 2024 2:45 PM (15 minutes)

Computing demands for large scientific experiments, including experiments at the Large Hadron Collider and the future DUNE neutrino detector, will increase dramatically in the next decades. Heterogeneous computing provides a solution enabling increased computing demands that pass the limitations brought on by the end of Dennard scaling. However, to effectively exploit Heterogeneous compute, software needs to be adapted, and resources need to be balanced. We explore the novel approach of Services for Optimized Network Inference on Coprocessors (SONIC) and present a strategy for optimized integration of heterogeneous coprocessors, including GPUs, FPGAs, Graphcore IPUs and others. Focusing on ML algorithms, we demonstrate how SONIC can be designed to dynamically allocate heterogeneous resources in an fully optimized mode. With the rapid adoption of deep learning models for core algorithms at big scientific experiments, we present a path towards rapid integration of deep learning models, and strategy for future large scale compute at big experiments including the CMS and ATLAS detectors at the Large Hadron Collider. We show our proposed path clears the way for substantially improved data processing by optimally exploiting resources while simultaneously increasing the bandwidth and overall computational power of these future experiments.

Focus areas

HEP

Primary authors: KONDRATYEV, Dmitry (Purdue University (US)); PASPALAKI, Garyfallia (Purdue University (US)); ZHAO, Haoran (University of Washington (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); MOHRMAN, Kelci Ann (University of Florida (US)); PEDRO, Kevin (Fermi National Accelerator Lab. (US)); LIU, Miaoyuan (Purdue University (US)); COCHRAN-BRANSON, Miles (University of Washington (US)); TRAN, Nhan (Fermi National Accelerator Lab. (US)); PALADINO, Noah (Massachusetts Inst. of Technology (US)); CHANG, Philip (University of Florida (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); PIPEROV, Stefan (Purdue University (US)); YAO, Yao (Purdue University (US)); FENG, Yongbin (Texas Tech University (US)); CHOU, Yuan-Tang (University of Washington (US))

Presenter: KONDRATYEV, Dmitry (Purdue University (US))

Session Classification: Contributed talks