Contribution ID: **36** Type: **not specified**

Next Generation Coprocessors as a service

Tuesday, 4 October 2022 17:00 (15 minutes)

In the as-as-service paradigm, we offload coprocessors to servers to run dedicated algorithms at high rates. The use of as-a-service allows us to balance computation loads leading to a dynamically resource-efficient system. Furthermore, as-a-service enables the integration of new types of coprocessors easily and quickly. In this talk, we present next generation studies using as-a-service computing, and we show the most recent performance of Intelligence Processing Units (IPUs), FPGAs, and how parallelized rule-based algorithms can also be implemented as-a-service quickly. We also show how we can optimize as-a-service to take into account network efficient inference strategies, including ragged batching. Finally, we propose a set of benchmarks that present real challenges and can enable us to understand how the future as-a-service landscape will evolve and how it can be used in recent scientific developments.

Primary authors: RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); KHODA, Elham E (University of Washington (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); LIU, Miaoyuan (Purdue University (US)); TRAN, Nhan (Fermi National Accelerator Lab. (US)); THOMAS, Nirmal; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); ROTHMAN, Simon (Massachusetts Inst. of Technology (US)); PIPEROV, Stefan (Purdue University (US)); MCCORMACK, William Patrick (Massachusetts Inst. of Technology (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US)); PEDRO, Kevin (Fermi National Accelerator Lab. (US))

Presenter: RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US))

Session Classification: Contributed Talks