AI Infrastructure Workshop Outcomes

Background

The principal goal of this workshop series is to bring stakeholders (experiments, divisions) and service/facility providers together and understand what AI for production/operation workflows will look like in the near term for both training and inference. In particular, stakeholders should discuss their software and hardware needs, including resource requirements if possible. CSAID will use the product of this mini-workshop to plan resource acquisition, deployment, and developer effort.

The first workshop was held April 6, 2023: https://indico.fnal.gov/event/58099

The first workshop included an overview of Fermilab facilities and infrastructure; inputs from many users from CMS, Accelerator Directorate (AD), intensity frontier experiments including DUNE, cosmic frontier experiments, theory, Emerging Technology Directorate (ETD).

Outcomes

After the first workshop, the AI project office AI infrastructure coordinators (Burt Holzman, Nhan Tran) collected materials and identified common themes and potential next steps. Figure 1 shows the AI Infrastructure work plan including current specific charges and stakeholder reporting.

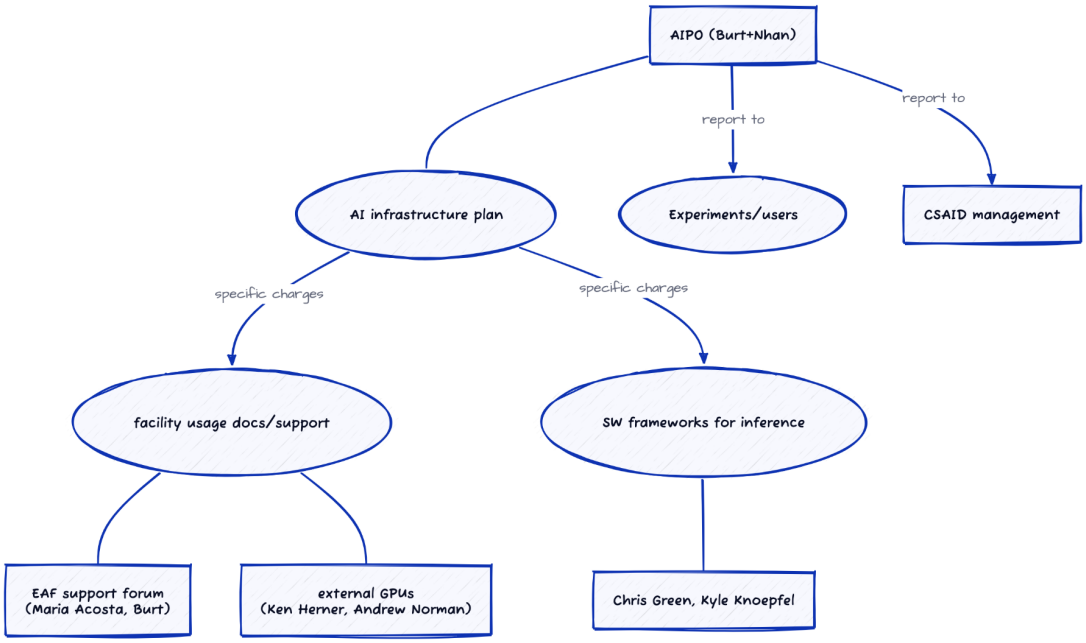

Figure 1: AI Infrastructure work plan including coordinators and reports to stakeholders

The AI Infrastructure work plan coordinators (POCs) have identified the following focus areas, below. For each of these focus areas, we define a charge and a focus area point-of-contact (POC) that will report on progress. The focus area POC can create a task force as needed with experiment stakeholders and CSAID (include division reps, users, postdocs, speakers from workshop).

Thus far, the three focus areas identified are:

- **Establish support forum (focused on EAF) with users helping other users – POC: Maria Acosta, Burt Holzman**
  - As EAF popularity increases and the facility scales, support for users will become more challenging to handle by only experts. *Build a user-support forum for EAF where other users can provide input*
  - This model can be expanded to other resources at Fermilab once a pattern of usage has been establish – e.g. Wilson Cluster, external resources (see next focus area)

- **Evaluate and document GPU usage from outside resources - OSG, HEPCloud, etc. – POC: Ken Herner, Andrew Norman**
  - There are a broader pool of GPU resources available to Fermilab users beyond those on-premises at Fermilab. *Evaluate those options, document their usage modes, and (if possible) make them easier to use*

- **Easier integration of inference with experiment frameworks – POC: Chris Green, Kyle Knoepfel**
  - We identified a pain-point where many different experiments are interfacing with ML software frameworks and inference modes (e.g. PyTorch, Tensorflow, ONNX-runtime) through a wide array of methods. This makes support for fast-evolving ML frameworks extremely challenging, especially when scaled out for large production campaigns
    - Increase cross-section for interaction by bringing framework experts such as Chris G, Kyle K, Matti K, Chris J, Giuseppe, DAQ team (DUNE + others) and others together
    - Define a priority list for SciSoft, LArSoft modes to support ML inference in experimental software frameworks

This will be reported to CSAID management for initial feedback on the focus area list and if any other items are of high(er) priority. Additionally, we will circulate this list to the registrants and ask for brief feedback via a form.

Findings from these focus areas as well as overall findings from the first workshop will be presented in a future forum, such as the CSAID Roadmap meetings or a 2nd workshop in the series.

Additional focus areas

Based on feedback from the community, we also enumerate here other areas identified during the workshop that we would like to highlight but place lower priority on than the highest priority focus areas discussed above.

- **Make commercial cloud access for analysis more programmatic** - for example, the Google Cloud projects where users deploy compute for quantum computing simulations. Formalizing this program should include a sustainable but lightweight operations and support model; a more systematic tie-in to grants/awards/task codes; etc. [from Gabe Perdue]
- **Fermilab processes related to smaller projects/teams** - not directly related to AI infrastructure, but affecting AI researcher teams disproportionately; i.e. Fermilab processes for giving access to site and computing can be burdensome and slow for small teams
- **Give AI users a global picture of computing resources at FNAL** - this could be derived from the initial CREST report, which provides an overview of all lab computing capabilities. We should build off it to make an AI-specific picture for folks.