NPP FY 2022 LDRD Type A Proposal - 1st Full Draft Presentation

Towards a Scalable and Distributed Machine Learning Service for Data-Intensive Applications

May 25, 2021

NPP - FY 2022 Draft LDRD Type A Presentation

Proposal Title (proposal Type A): Towards a Scalable and Distributed Machine Learning Service for Data-Intensive Applications

Principal Investigator: Meifeng Lin

Department/Division: CSI and NPP

Other Investigators: Brett Viren, Torre Wenaus

Indicate if this is a cross -directorate proposal. Yes _X__ No___

If yes, Identify submitting directorate: CSI

Proposal Term: From - 10/2021 To – 9/2024

May ask for a delayed start to ensure new hires are onboard.

Proposal Overview

- **Motivation:** Current ML development focuses on algorithm R&D; translating it into actual production will require improved ML software infrastructure to facilitate deployment and reduce turnaround time.

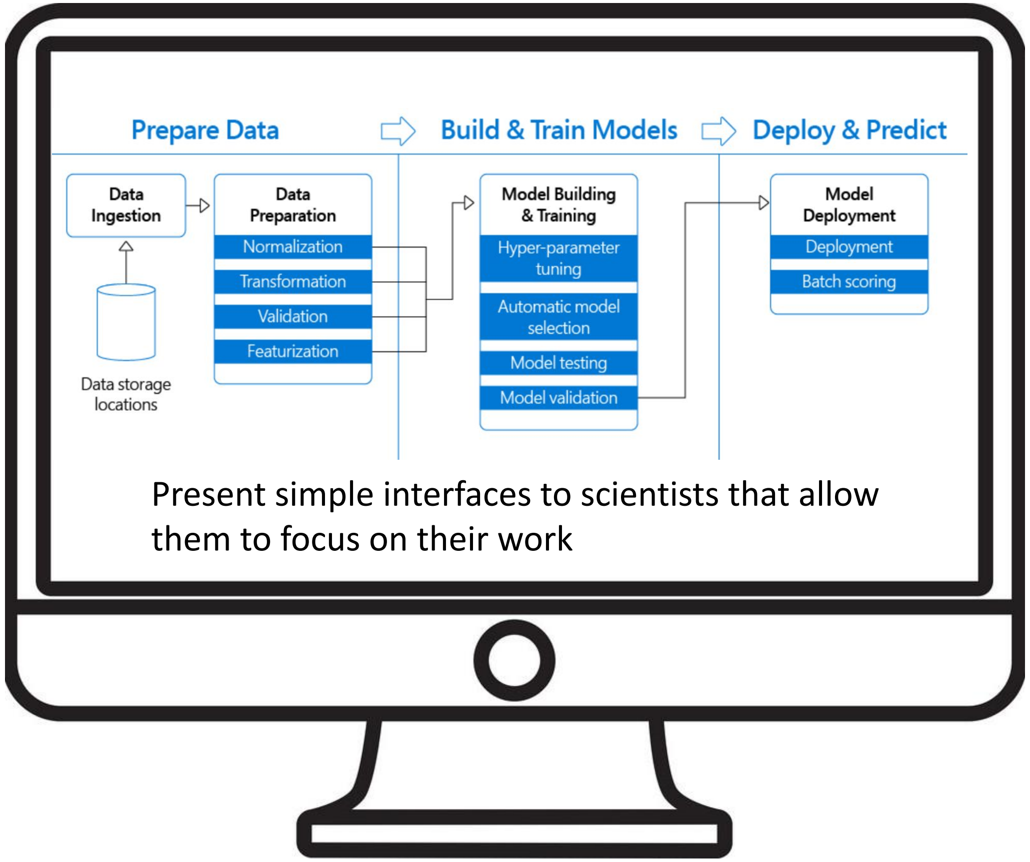
- **Project Description:** We propose a collaboration between CSI's High Performance Computing (HPC) group and the Physics Department's NPPS group to combine expertise towards developing and applying large scale distributed machine learning services for hyperparameter optimization and training that transcend any single facility. CSI's HPC expertise combined with NPPS's experience building large scale ML services and working with physicists in identifying and supporting physics analysis use cases will enable us to effectively leverage large scale HPC resources for ML to a much greater extent than either CSI or the NPPS team can on their own. The first application used to define requirements will be the AI/ML applications of DUNE and the Wire-Cell toolkit group.

- **Expected Results:** A service delivering large scale processing resources to AI/ML applications in the BNL science program, drawing on HPCs, distributed grids, clusters and commercial clouds through a coherent suite of services and scientist-friendly interfaces. The service will be an enabler for expanding scientific creativity in AI/ML applications towards models of a scale and complexity otherwise impractical, or with prohibitive development/training time requirements.

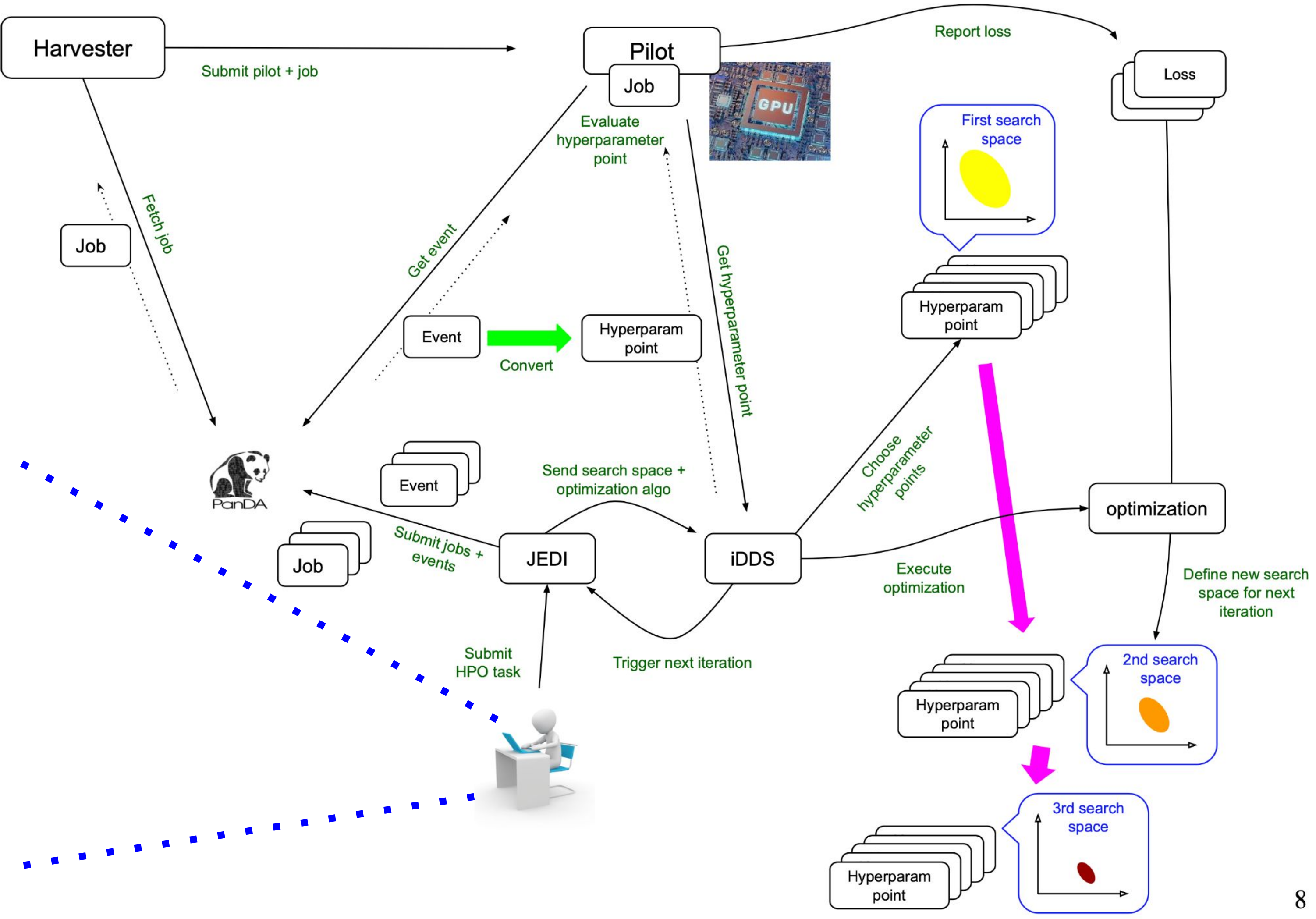
- **Information on how this proposal aligns with the NPP strategic long-term plan:** The proposal delivers an enabler for the Laboratory’s strategic growth plan to use AI/ML to create solutions to experiment-driven computing challenges across the Laboratory, leveraging Laboratory and wider computing resources to do so.

Hyperparameter Optimization with PanDA and iDDS

Focus on elements of the AI/ML workflow that are processing intensive for large, complex, data intensive models, such as hyperparameter optimization



Present simple interfaces to scientists that allow them to focus on their work



Challenges we intend to address

- Our goal is to deliver a ML software service with three key features:
    - **Portability:** Can we deploy it across different hardware architectures and facilities?
    - **Scalability:** How do we ensure the service delivers optimal performance for the applications?
    - **Agility:** Can the service be agile and intelligent to dispatch the applications to the most suitable available resources?
- To accomplish this, we need to address the associated research questions:
    - **Portability:** Containerization + portable ML frameworks? Will the architecture-specific backends be efficiently supported within container environment?
    - **Scalability:** Different ML algorithms/applications may scale differently on different platforms. Can we design a tool that can intelligently decide which platform will be the most suitable for the target application? Performance analysis and modeling of the ML app may be needed.
    - **Agility:** Can PanDA and iDDS be able to dispatch the ML workloads to suitable resources in real time? Are there other tools we can develop/utilize?
- Issues related to **cybersecurity** will also need to be investigated.

Milestones and Success Criteria

- **Year 1:** Collect user requirements from scientific applications at BNL, making use of the newly established AI/ML working group. Investigate single-node portability solutions.
- **Year 2:** Investigate scalability solutions, and develop analysis and modeling tools.
- **Year 3:** Investigate ways to integrate portability and scalability components with real-time workload management.
- **Success Criteria:** Working prototype to be deployed on distributed HPC resources.

Return on Investment:

- Our project will pave the way for the development of a scalable and distributed machine learning service framework.
- Will allow more HEP/NP scientists to apply ML training and optimization more efficiently.
- Will place us in a strong position to respond to future ASCR and HEP/NP ML funding calls.

NPP FY 2022 LDRD Draft Presentation – Funding Requirements (Preliminary).

Fiscal Year	FY 2022	FY 2023	FY2024
CSI - HPC	260k	255k	265k
• 1 PD for R&D (1st year relo+new computer)			
• 0.1 FTE Lin for project management and coordination, PD supervision			
NPP	150k	140k	145k
• 0.5 PD for data preparation and testing (1st year relo+new computer)			
• 0.05 FTE each for Viren and Wenaus			
Travel $5000/year			
Total Funds*	410k	395k	410k

*Make sure appropriate M,S,T, dept. burdens and lab Overhead are included.