RUTGERS

Diamond – Democratizing Large Neural Network Model Training for Science

Haotian Xie, Gengcong Yang, Jingxin Wang, Song Bian, Minu Mathew, Rohan Marwaha, Yadu Baduji, Ian Wang, Volodymyr Kindratenko, Kyle Chard, Shivaram Venkataraman, Ian Foster, Zhao Zhang

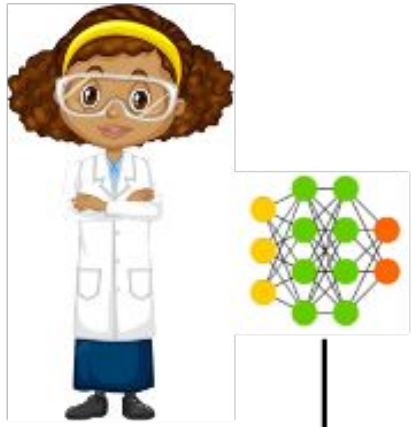
Motivation

- Researchers with diverse backgrounds and levels of expertise seek to adopt DL methods to advance their research.
- However, the application of cutting-edge DL methods often requires high-performance computing (HPC) clusters
- Regardless of their HPC expertise, researchers encounter usability, performance, portability, and reproducibility obstacles.
- Existing HPC clusters expose a minimal job and data management and assume user proficiency in both DL and HPC.

Diamond Approach

- Development: Web-based service for remote programming and containers to support configuration of complex environments
- Deployment: Suite of tools such as automatic scaler, task planner, and cross-cluster executor to enable scaling, optimized execution, data management, and provenance across HPC clusters
- Dissemination: Integration with external services, such as Garden/DLHub and HuggingFace, to publish and deploy models for on-demand inference.

Diamond Software Components



Teams

- Rutgers
- TACC
- UChicago
- UIUC
- UWisc-Madison

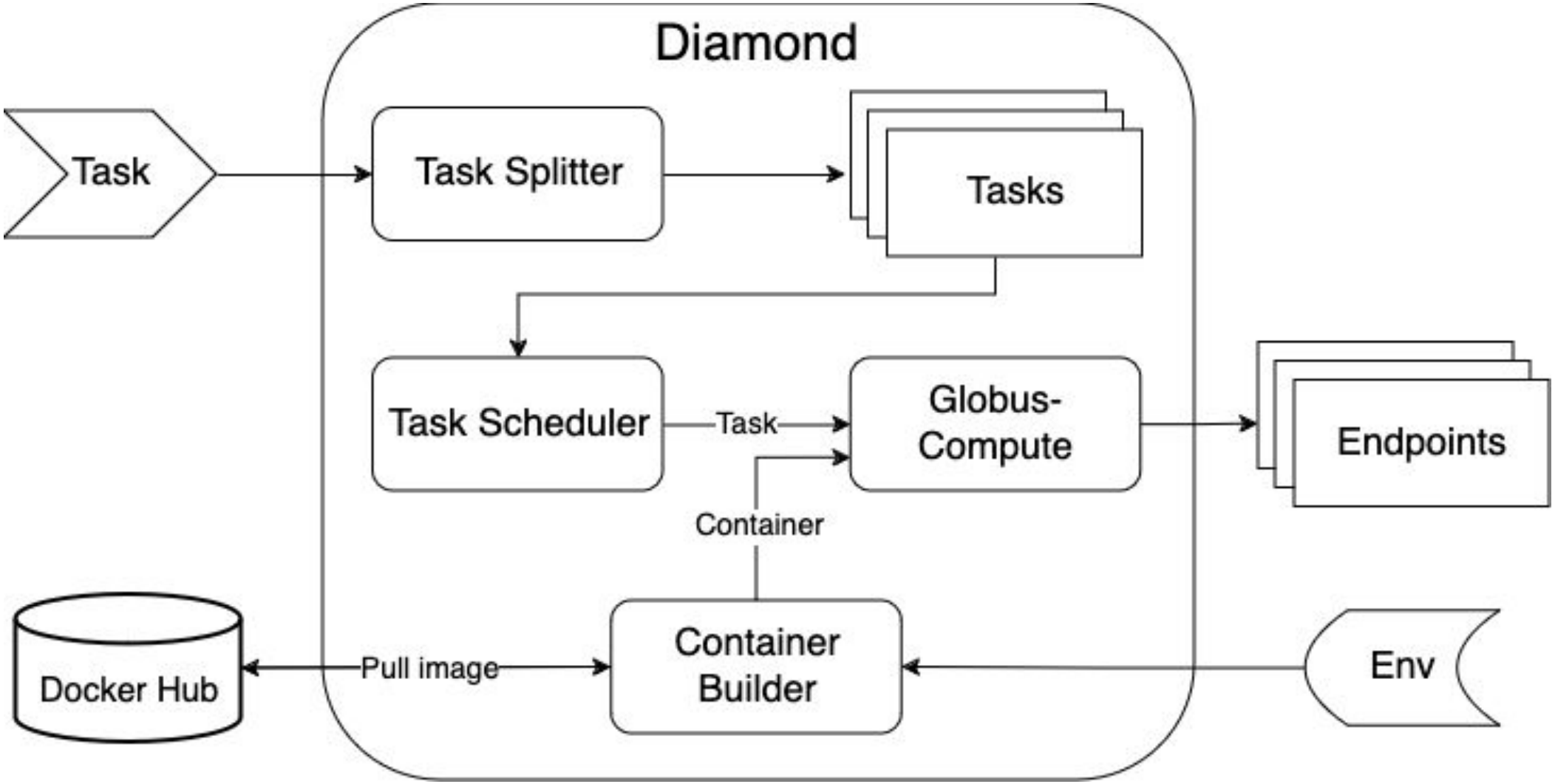
App Program
GPT-NeoX
OpenFold
HuggingFace
ViT
DeepSpeed
Ray
Horovod
Pillow
numpy
Tokenizers
Container Image
FuncX

Dev Interface
Diamond Develop
Provenance Tracking
Publication
Diamond Disseminate
Auto-Scaler
Hyper-Parameter Tuner
Job Planner
Placement Planner
Cross-cluster Executor
Diamond Deploy
Auth Module
Container Conf
Data Fabric
Diamond Core
Diamond

FuncX
Slurm
MPI
CUDA
CUDNN
NCCL
MAGMA
cuSolver
cuBLAS
File System
GPU Cluster

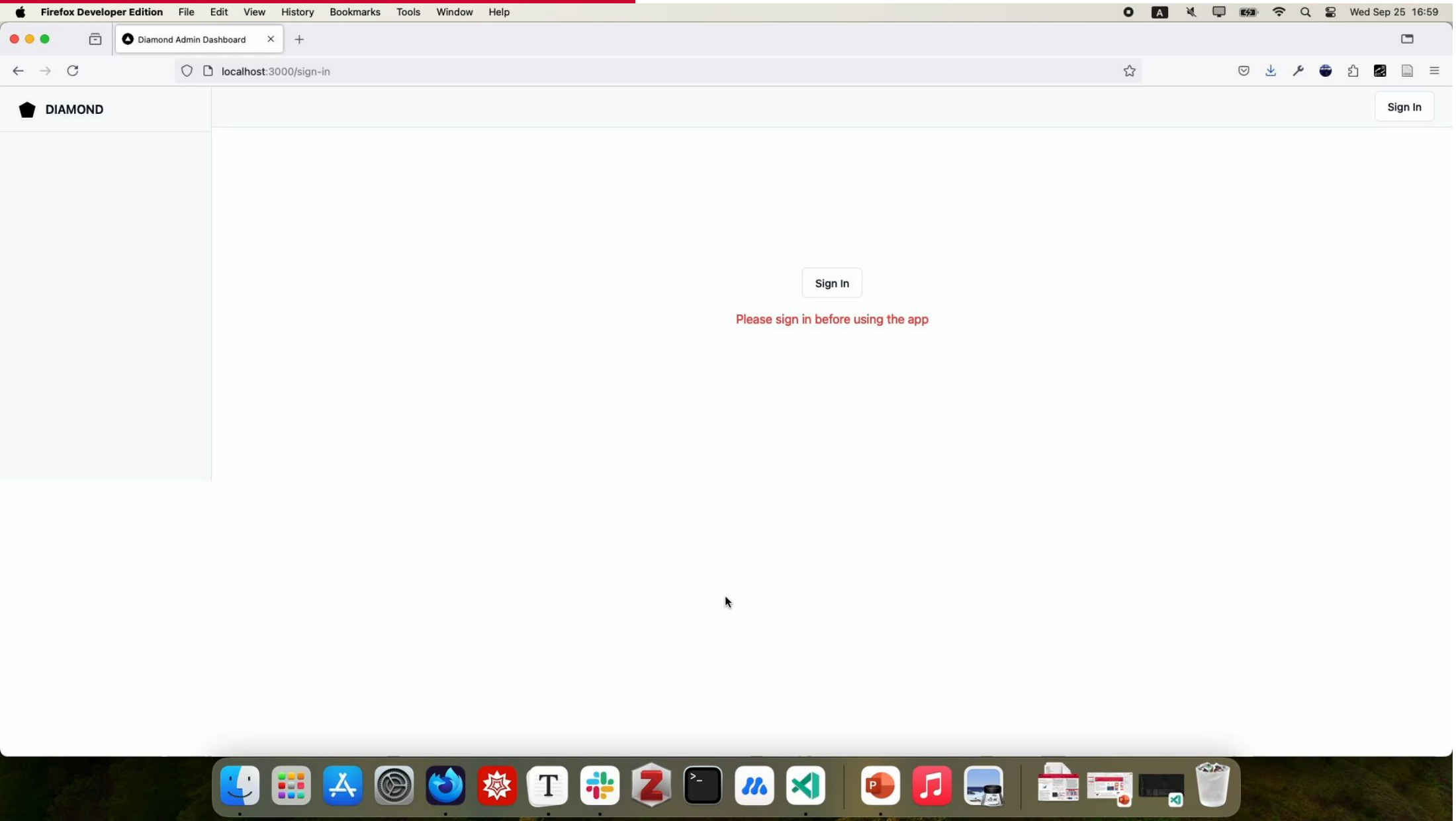
FuncX
Determined AI
wandb
Garden
DLHub
External Service

Diamond Workflow



Run with Diamond

https://diamond.ncsa.illinois.edu/



RUTGERS

Thank you!