Hybrid Edge-HPC Systems for Low-Latency Data-Driven Inference

Ryan Hartung and Douglas Thain
{rhartung, dthain}@nd.edu

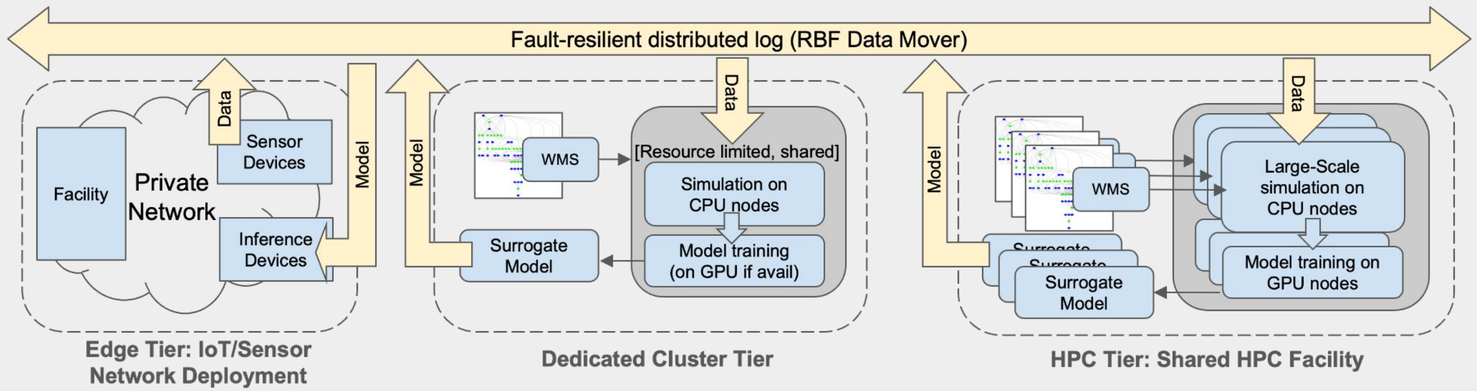
Background

- Edge applications require continuous, low-latency inference while model updates depend on remote HPC simulations
- High-fidelity CFD simulations take hours and batch scheduling causes unpredictable delays
- Outdated models degrade prediction accuracy

Solution

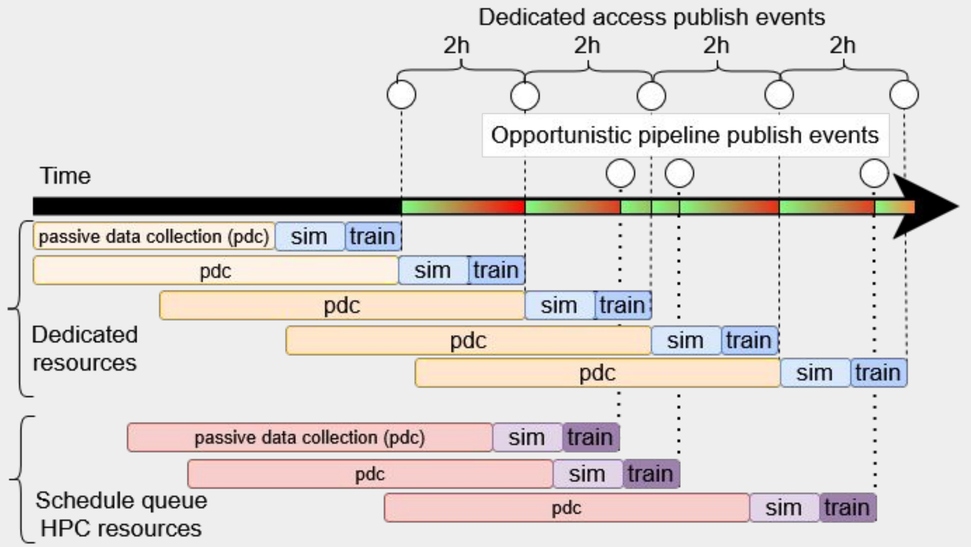
Reverse Backfill (RBF), a hybrid edge-HPC learning and inference architecture. RBF treats HPC resources not as synchronous backends, but as asynchronous model improvement engines. Edge nodes perform continuous inference using surrogate models, while simulation and training pipelines generate improved models as resources become available.

RBF Architecture



Decay Time

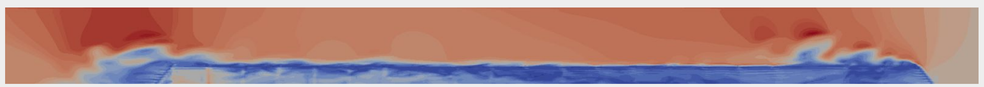
Passive data collection (pdc), simulation (sim), and training (train) stages overlap across multiple pipeline instances, while model updates are published opportunistically upon completion. This design enables continuous inference despite irregular and delayed HPC execution.

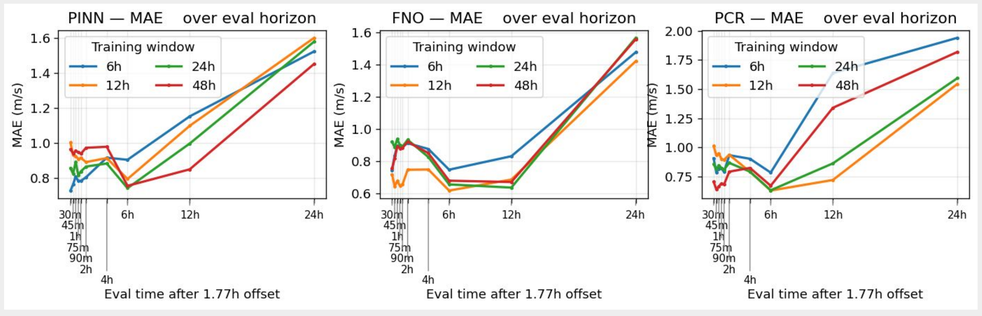


Surrogate Models

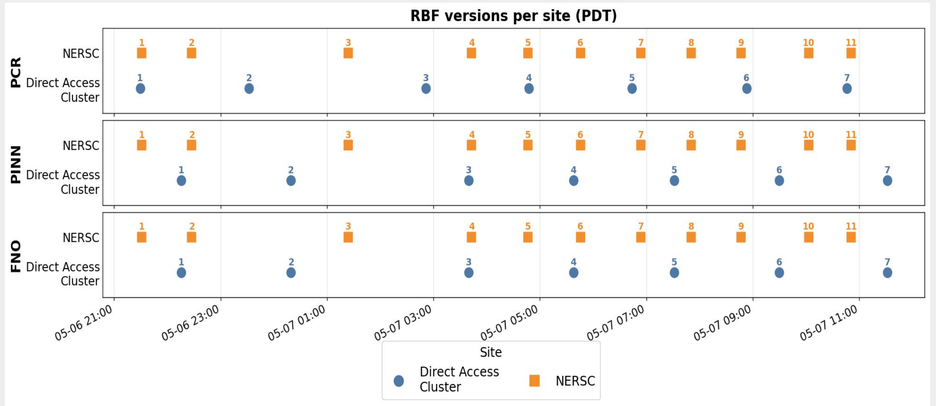
Simulation software: OpenFOAM

Training models: Physics-Informed Neural Network (PINN), Fourier Neural Operator (FNO), and Principal Component Regression (PCR).





Model Cadence



Combination	Min	Avg	Max	Std
dedicated cluster	113.4	134.8	200.4	32.9
NERSC	47.9	80.0	176.5	40.4
dedicated cluster + NERSC	3.3	50.0	135.8	34.3

Acknowledgements

On behalf of the RADICAL team: Liubov Kurafeeva, Ryan Hartung, Benjamin Carter, Alan Subedi, Avhishek Biswas, Michael Fay, Chandra Krintz, Rich Wolski, Andre Merzky, Douglas Thain, Mehmet Can Vuran, Shantenu Jha.

This work was supported by DOE Grant DE-SC0025541

UNIVERSITY OF NOTRE DAME

U.S. DEPARTMENT of ENERGY