Efficient Digital Twin Training using Uncertainty-Guided Data Generation

Ana Gainaru

MS323, Neural Acceleration, Surrogate Models, and Learning Techniques for HPC Kernels
SIAM CSE25, March 06, 2025

Interactions between HPC and AI frameworks

- Research on HPC I/O focused on **modeling and simulation applications**
  - Handling large output and checkpointing the results
  - Write operation bursts commonly dominate traditional workloads
  - Analysis and viz typically access large portions of the data
- ML workloads perform **small I/O reads spread across a large number of random files**
  - Usually read-intensive and use many small files
- There is **no well-established consensus** on the preferred I/O stack for ML workloads
  - Many developers resort to developing their own custom solution

Interactions between HPC and AI frameworks

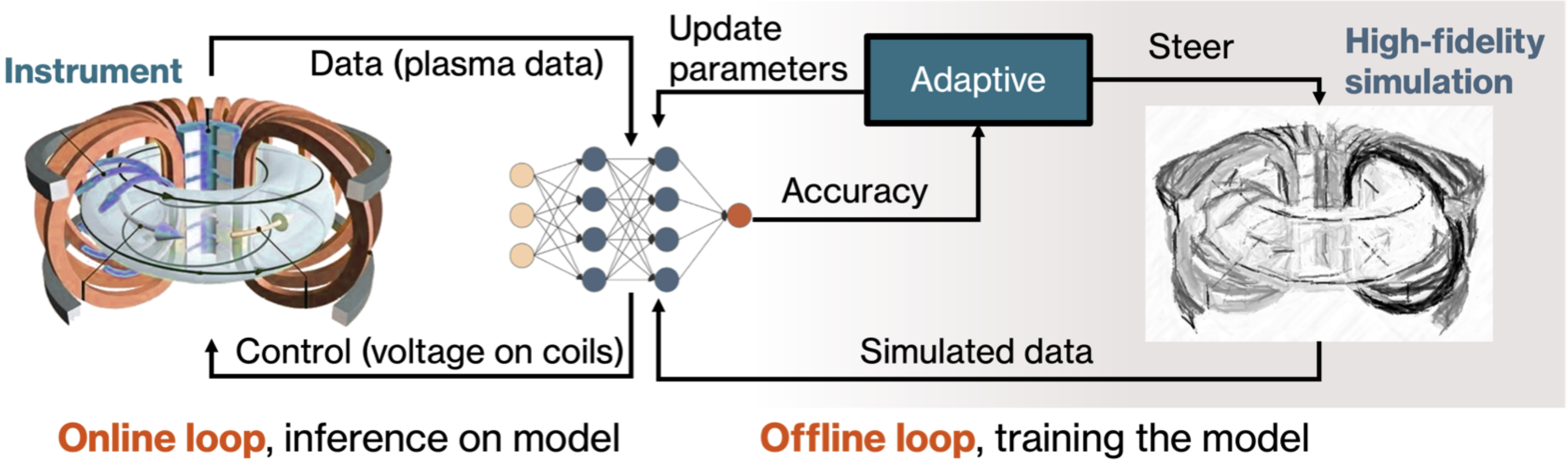
- Research on HPC I/O focused on **modeling and simulation applications**
  - Handling large output and checkpointing the results
  - Write operation bursts commonly dominate traditional workloads
  - Analysis and viz typically access large portions of the data
- ML workloads perform **small I/O reads spread across a large number of random files**
  - Usually read-intensive and use many small files
- There is **no well-established consensus** on the preferred I/O stack for ML workloads
  - Many developers resort to developing their own custom solution

1. What are the types of workflows that combine HPC and AI ?
2. What are the limitations integrating AI and HPC?
3. What are the benefits of unifying the AI and HPC solutions ?

Interactions between HPC and AI frameworks

- **AI-in-HPC**: The AI system is introduced instead of a component or a whole HPC simulation.
  - e.g. medical pipelines
- **AI-out-HPC**: the outside AI system dynamically controls the progression of the HPC workflow
  - e.g. control of computational campaigns via reinforcement learning
- **AI-about-HPC**: the AI systems are concurrent and coupled to the main HPC tasks
  - e.g., AI-based analysis and viz use the output of the HPC simulation to provide further insights

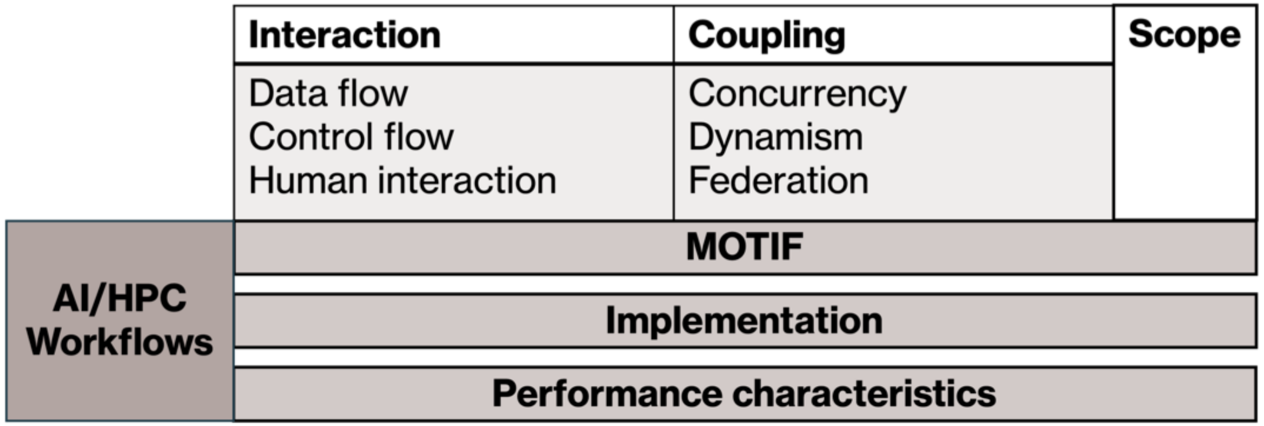
1. What are the types of workflows that combine HPC and AI ?

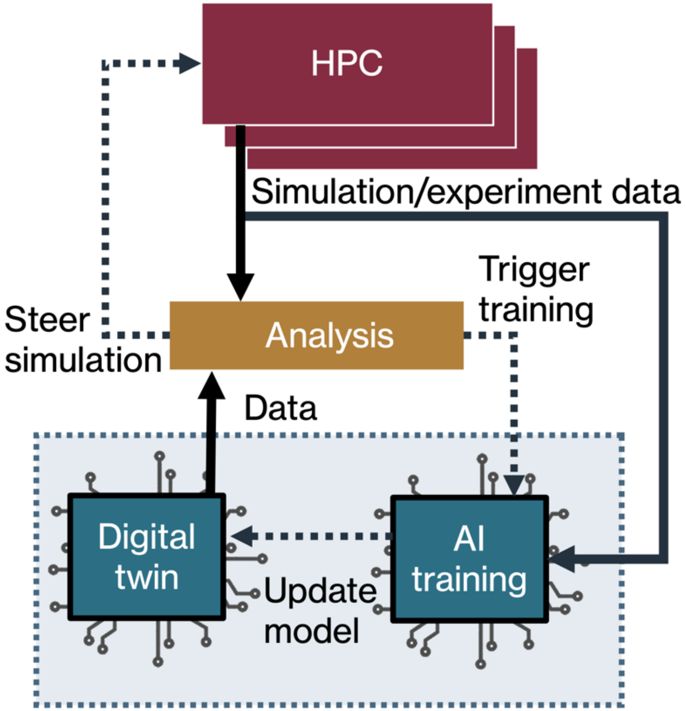


AI-coupled HPC Workflows, Shantenu Jha, Vincent Pascuzzi, and Matteo Turilli, 2022

Types of AI+HPC workflows

Breakdown of the main characteristics used to define the different behavioral motifs





- **AI-based steering ensembles of simulations**
  - HPC workflows, e.g. ensemble of simulations are "steered" by an AI system
- **Multistage pipeline**
  - Pipeline of HPCs with AI-based functions between stages
- **Inverse design**
  - AI-driven optimizations are used to iteratively identify causal factors from observational data
- **Digital replica**
  - HPC concurrent with AI digital replicas predictions and health monitoring

1. What are the types of workflows that combine HPC and AI ?

AI-coupled HPC Workflow Applications, Middleware and Performance, 2024, Wes Brewer, Ana Gainaru, Frédéric Suter, Feiyi Wang, Murali Emani, Shantenu Jha

Limitations of current solutions

- Testcases 3 types of AI/HPC workflows
  - Digital twin training for plasma physics simulation
  - Workflow pipelines in cancer research
  - Inverse design for plasma simulation*
- A data-centric approach to integrating AI in HPC
  - **Throughput limitations of the filesystem (in memory)**
  - **Code coupling (strong/loose coupling)**
  - **Large data space for training (steering the experiment)**

2. What are the limitations integrating AI and HPC?

- No consensus on the preferred I/O stack for ML workloads
  - **No schema**
  - **Not a consistent way of handling models and data**
  - **Custom solutions**

* The Artificial Scientist -- in-transit Machine Learning of Plasma Simulations, Jeffrey Kelling et all, IPDPS, 2025

Limitation: Raw formats are common in AI

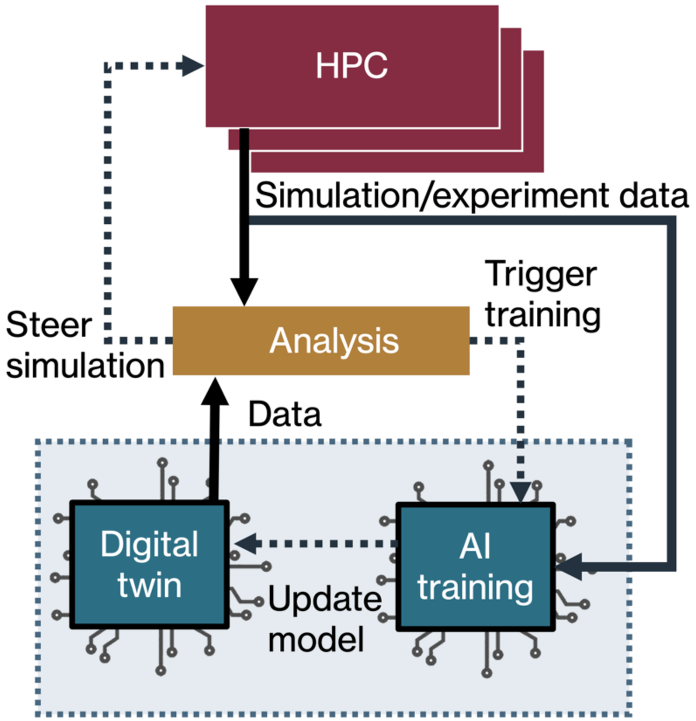
- Each solution uses different formats
  - Each with diverse properties that can affect performance when storing and retrieving ML datasets
  - Many applications still rely on unoptimized solutions (raw text and image formats)
  - Once a model is trained, the input data format is lost

File Format	Parallel I/O	Compression	Schema
TFRecords	✗	✗	✓
Apache Parquet	✓	✓	✓
RecordIO	✗	✓	✗
NPZ	✗	✗	✗
HDF5	✓	✓	✓
ADIOS2	✓	✓	✓
Raw formats (e.g. images, video, text)	✗	✗	✗

2. What are the limitations integrating AI and HPC?

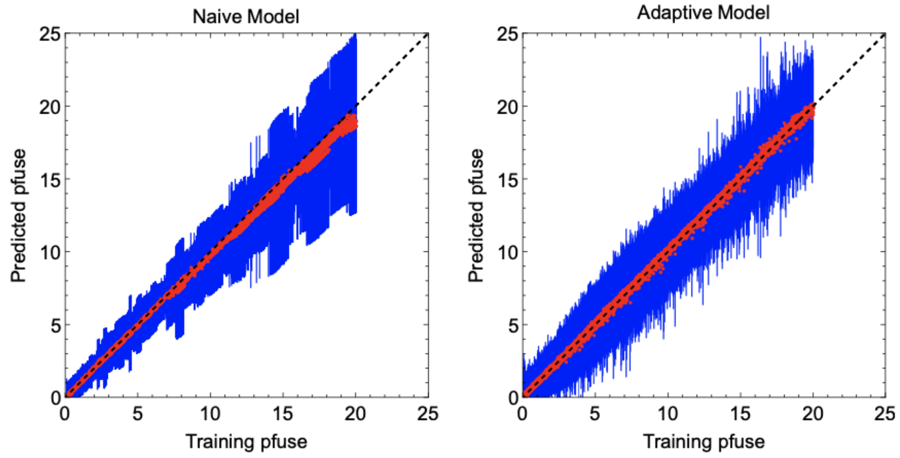
Limitation: Massive training datasets

- Naive uniform sampling of the input space
  - Imbalance data coverage
    - Some areas unnecessary over-sampled while others have sparse data coverage
  - Identifying deficit regions is a challenging task
    - Current solutions are increasing the amount of training data to eventually fill the coverage gaps -> **Massive datasets**
- Choosing the location of data samples
  - Bayesian approach to capture uncertainty in a deep neural network
- Steer an ensemble of simulations



3. What are the benefits of unifying the AI and HPC solutions ?

An adaptively trained model provides more accurate prediction at high electron fusion power output.



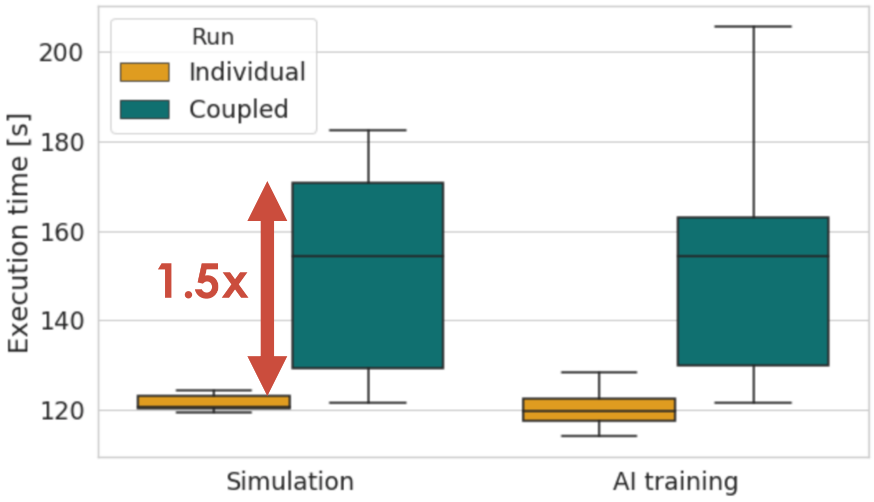
Adaptive Generation of Training Data for ML Reduced Model Creation, Mark Cianciosa et al, BTSD 2022

Oak Ridge National Laboratory

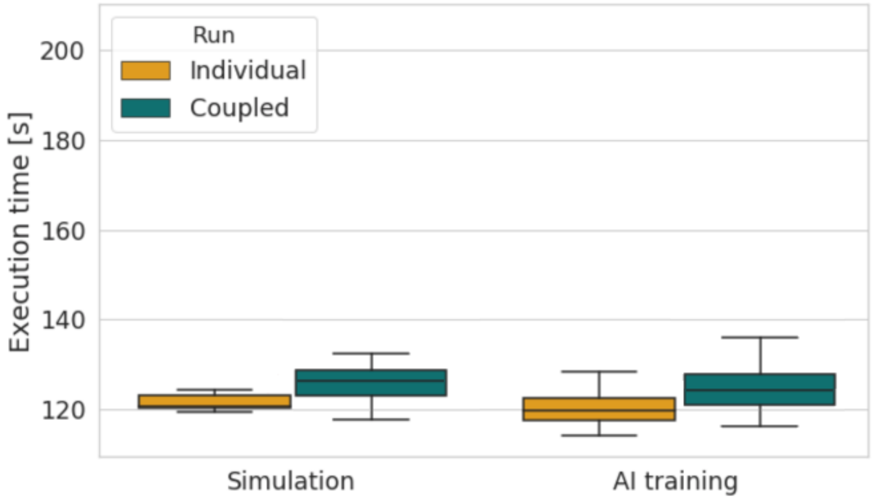
Limitation: Filesystem throughput

- Separate runs
  - Less than 3% performance degradation compared to separate runs
  - Less variation
  - If more models are needed
    - Overhead stays below 5% for 3 models
    - Variation increases with the number of nodes
- Throughput of **40 TFlops**
  - On Frontier

3. What are the benefits of unifying the AI and HPC solutions ?



Simulation and analysis execution time if ran separately or coupled

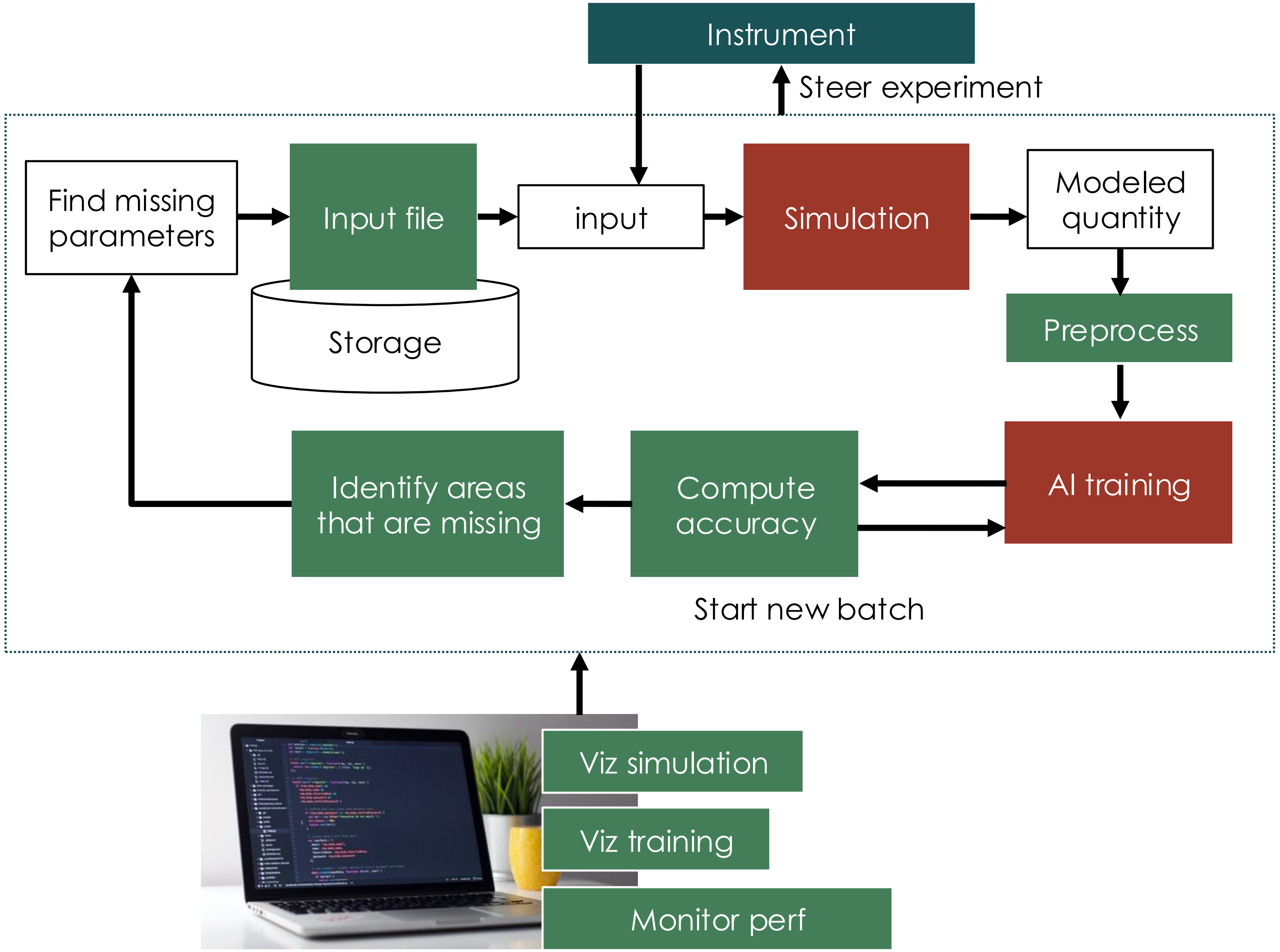


Execution time when streaming between coupled codes

OAK RIDGE National Laboratory

Limitation: Code coupling

- Coupling plasma codes
  - Streaming data between
  - Pre-process / post-process as part of the workflow
  - EFFIS workfow management system
    - Overhead of 1-5%
    - Command and control capabilities
- Near real time
  - Visualization
  - Performance monitoring



3. What are the benefits of unifying the AI and HPC solutions ?

Conclusions

- Many DOE projects are developing AI / HPC workflows
  - The I/O solution for AI and HPC have evolved separately
  - There is no well-established consensus on the I/O stack for AI-HPC workflows
- Our experience at large scale
  - Use **self describing formats** for data management
    - Enables querying, model/data tracking, reproducability
  - It's better to **avoid the filesystem**
  - Separate workflow into **units of work**
    - Offload data transfer to streaming libraries

Thank you !

gainarua@ornl.gov

Backup slides

Relevant publications

Junqi Yin et al. **Evaluation of pre-training large language models on leadership-class supercomputers**
The Journal of Supercomputing, June, 2023

Gainaru et al. **Understanding the Impact of Data Staging for Coupled Scientific Workflows**
IEEE Transactions on Parallel and Distributed Systems, 2022

Gainaru et al. **Framework for Automating the I/O of Deep Learning Methods**
In revision, Transactions on Computational Biology and Bioinformatics, 2022

Suchyta et al. **Hybrid Analysis of Fusion Data for Online Understanding of Complex Science on Extreme Scale Computers**, Cluster, 2022

Jean Luca Bez et al. **Access Patterns and Performance Behaviors of Multi-layer Supercomputer I/O Subsystems under Production Load**, HPDC 2022

Wang et al. **Improving I/O Performance for Exascale Applications through Online Data Layout Reorganization**,
IEEE Transactions on Parallel and Distributed Systems, 2021

Gainaru et al. **Profiles of upcoming HPC Applications and their Impact on Reservation Strategies**,
IEEE Transactions on Parallel and Distributed Systems, 2020

Gainaru et al. **Speculative scheduling for stochastic HPC applications**,
Proceedings of the 48th International Conference on Parallel Processing, 2019

Raghul Gunasekaran et al. **Comparative I/O Workload Characterization of Two Leadership Class Storage Clusters**, PDSW 2015