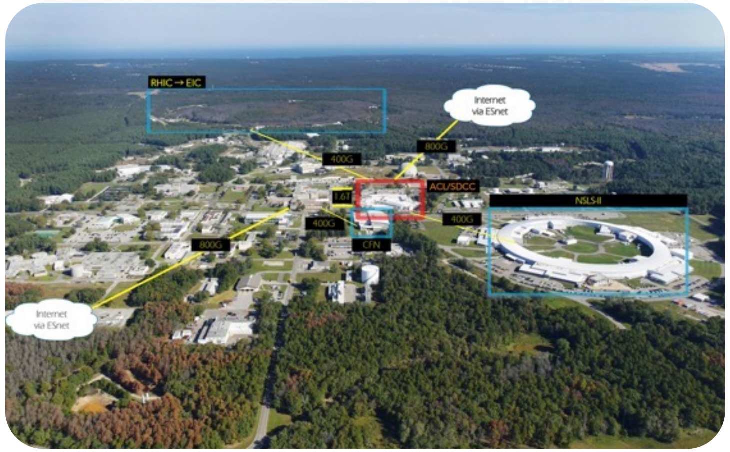
Novel AI Architectures

Adolfy Hoisie, Chair
Systems, Architectures, and Emerging Technologies Department
Computational Science Initiative

Brookhaven Lab-Stony Brook University Workshop on Artificial Intelligence
May 07, 2024

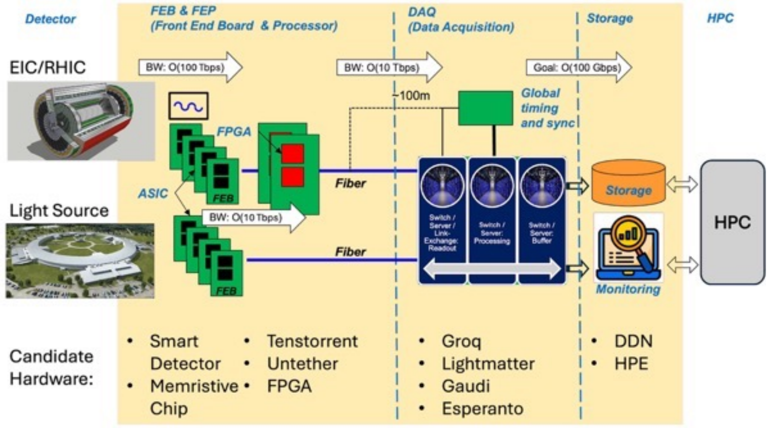
Codesign in Action for Experimental Science Computing: Architectures, Systems, and Testbeds

Advanced Computing Lab

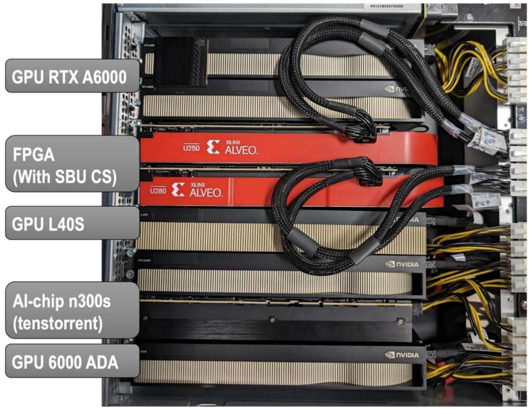


Unique collaborative testbed facility with access to live, actual data from diverse experiments, such as CFN (microscopy), NSLS-II, and RHIC, for codesign of architectures and experimental workflows.

Experimental Science Workflows

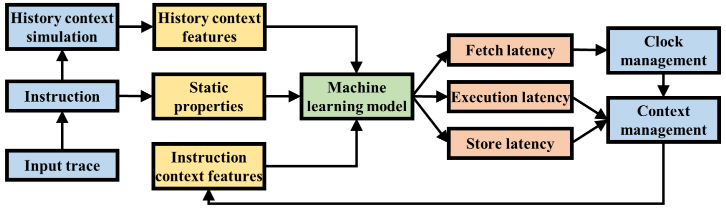


Extreme data challenges, heterogeneity, large spectrum of spatial and temporal computing scales from the edge to the extreme – real-time to long computational campaigns.

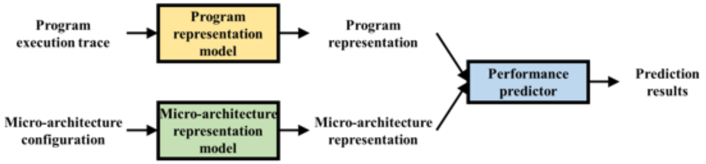


AI-based Modeling and Simulation

Leading the charge with SimNet and PerfVec
Accurate simulation faster by orders of magnitude compared with Discrete-Event Simulation



SimNet: AI-based architecture simulation https://github.com/lingda-li/simnet



PerfVec: AI-based Architecture modeling https://github.com/PerfVec/PerfVec

Li, L. S. Pandey, T. Flynn, H. Liu, N. Wheeler, and A. Hoisie. 2022. SimNet: Accurate and High-Performance Computer Architecture Simulation using Deep Learning. POMACS 6(2):Article 25. DOI: 10.1145/3530891.

Pandey, S., L. Li, T. Flynn, A. Hoisie, and H. Liu. 2022. Scalable Deep Learning-Based Microarchitecture Simulation on GPUs. SC22, pp. 1-15. DOI: 10.1109/SC41404.2022.00084.

Brookhaven National Laboratory

Performance Prediction Methods: Speed versus Accuracy

Smart Modeling and Simulation for HPC (SMaSH) is an intricate challenge because of the complexity of the design space.

Methodologies exist that lack either practicality or accuracy.

	Speed	Accuracy	Flexibility
Analytical Modeling	Fast	Low	Low
Emulation	Fast	High (?)	Very low
Discrete Event Simulation	Slow	High	High
Machine Learning (ML)-based Simulation	Medium	High	Medium
ML-based Modeling	Fast	High	High

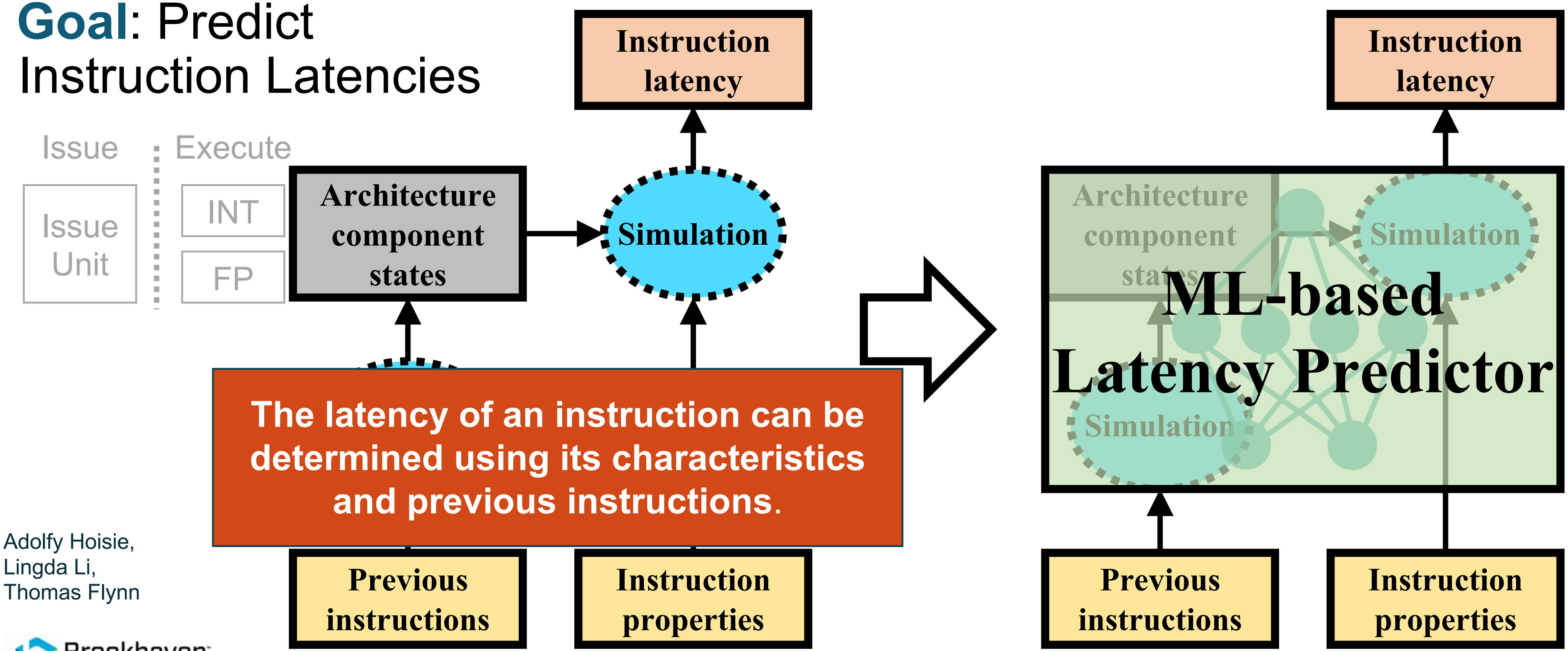
Discrete event (DE) simulation is slow:

- For example, gem5 simulates a modern microprocessor at several hundreds of KIPS.
- Not practical for realistic architectures and workloads.

GOAL: Accelerate Accurate Architecture Simulation by ~~Two~~ Many Orders of Magnitude

Machine Learning (ML)-based Simulation Foundation

Goal: Predict Instruction Latencies



The latency of an instruction can be determined using its characteristics and previous instructions.

Adolfy Hoisie,
Lingda Li,
Thomas Flynn

Brookhaven National Laboratory

A New Path: ML-based Simulation

Explore ML's application in computer architecture simulation:

ML has shown great success in many domains.

- ML models are excellent function approximators.

ML is highly regular and parallel.

- Modern accelerators (e.g., GPUs and TPUs) are well optimized for ML.

We offer the first ML-based computer architecture simulator: ***SimNet***

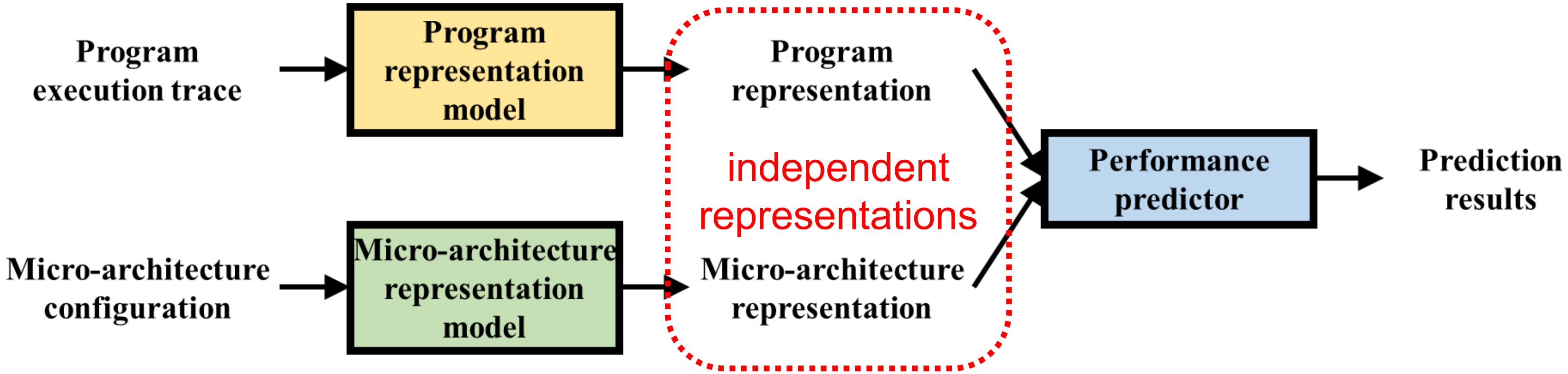
Li et al. SimNet: Accurate and High-Performance Computer Architecture Simulation using Deep Learning. SIGMETRICS, 2022.
Pandey et al. Scalable Deep Learning-Based Microarchitecture Simulation on GPUs. SC22, 2022.

Generic Performance Modeling

A generic performance model should separate the impact of program and microarchitecture.

When one party changes, there is no need to remodel the other.

PerfVec isolates the performance impact of program and microarchitecture using separate ML models.

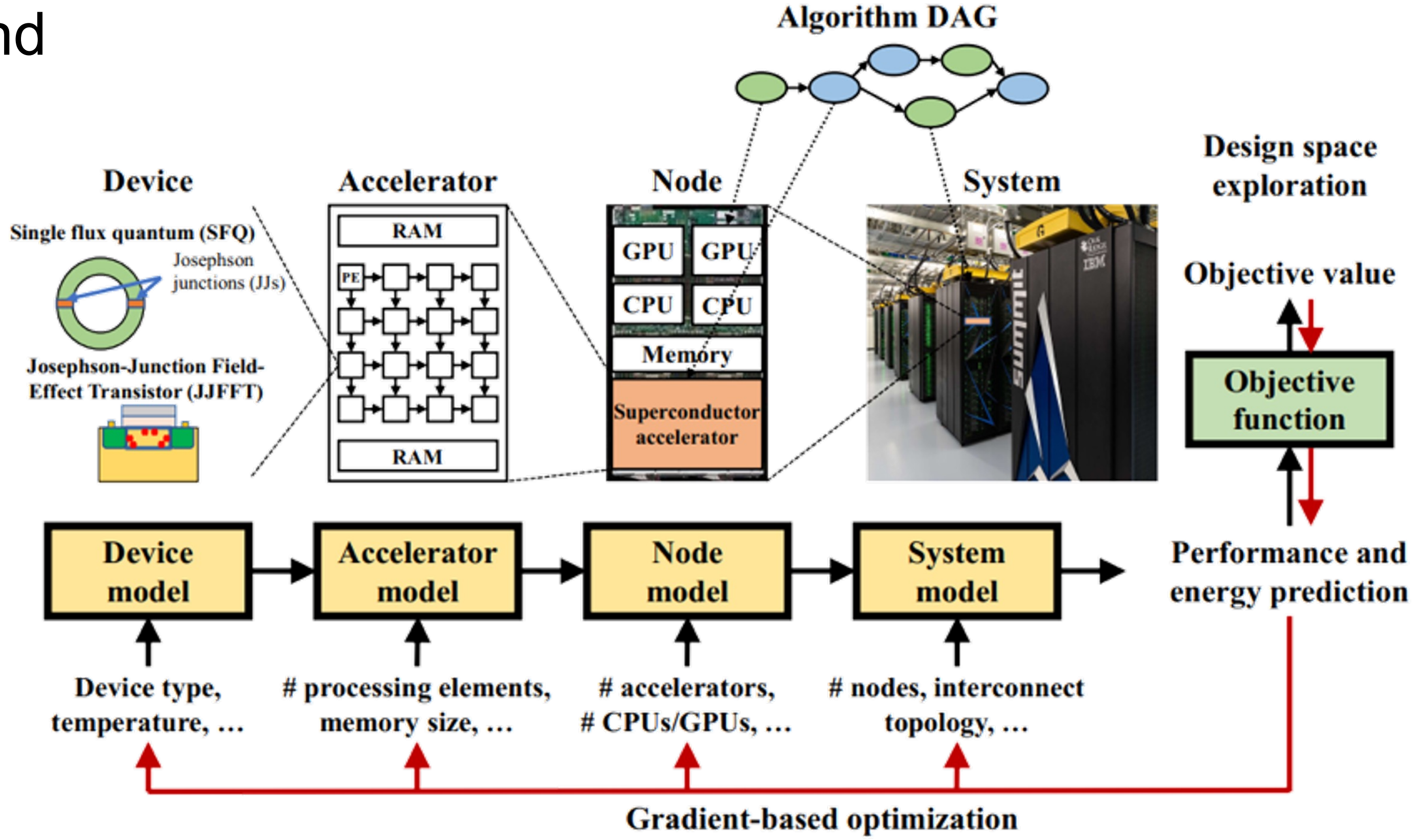


Li, Flynn, Hoisie. Learning Independent Program and Architecture Representations for Generalizable Performance Modeling. https://arxiv.org/abs/2310.16792.

Hierarchical, AI-enabled Modeling and Optimization of Future Supercomputers

Goal: Develop a modular and hierarchical modeling framework to explore and optimize system-level impacts of beyond-CMOS technologies

- Superconducting accelerators
- Dense linear algebra applications
- AI-enabled analytical modeling and simulation across abstraction levels



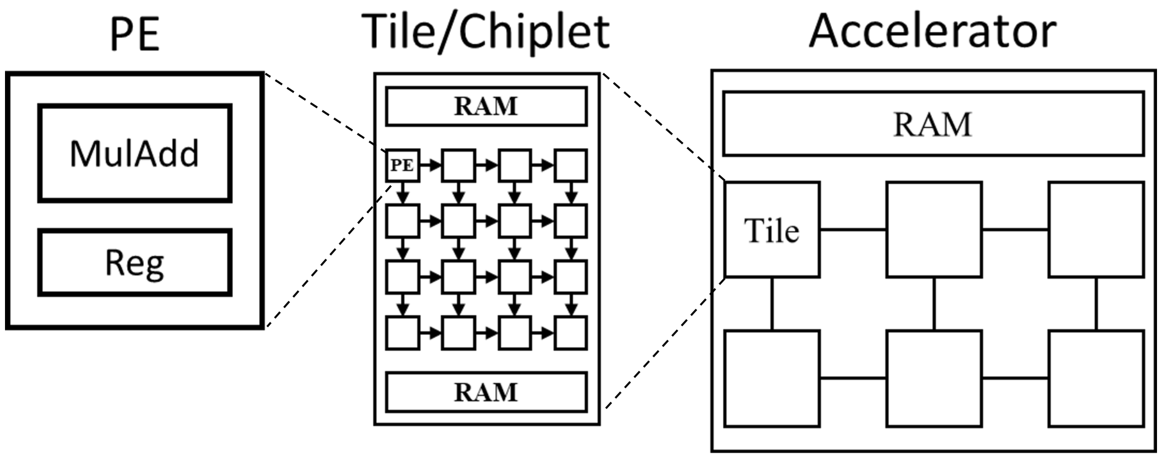
Lingda Li and Adolfy Hoisie with collaborators from the University of Texas at Austin

ML-enhanced Modeling

ML-based primitive-level surrogate accelerator models

- Train fast and accurate surrogate using traditional simulators
- Primitive operations directly executed by the accelerator, e.g., fixed-size matrix multiplication, data movement
- Large operations (e.g., GEMM) are decomposed into primitives

Integrate with ML-based and/or analytical models of other node/system-level components

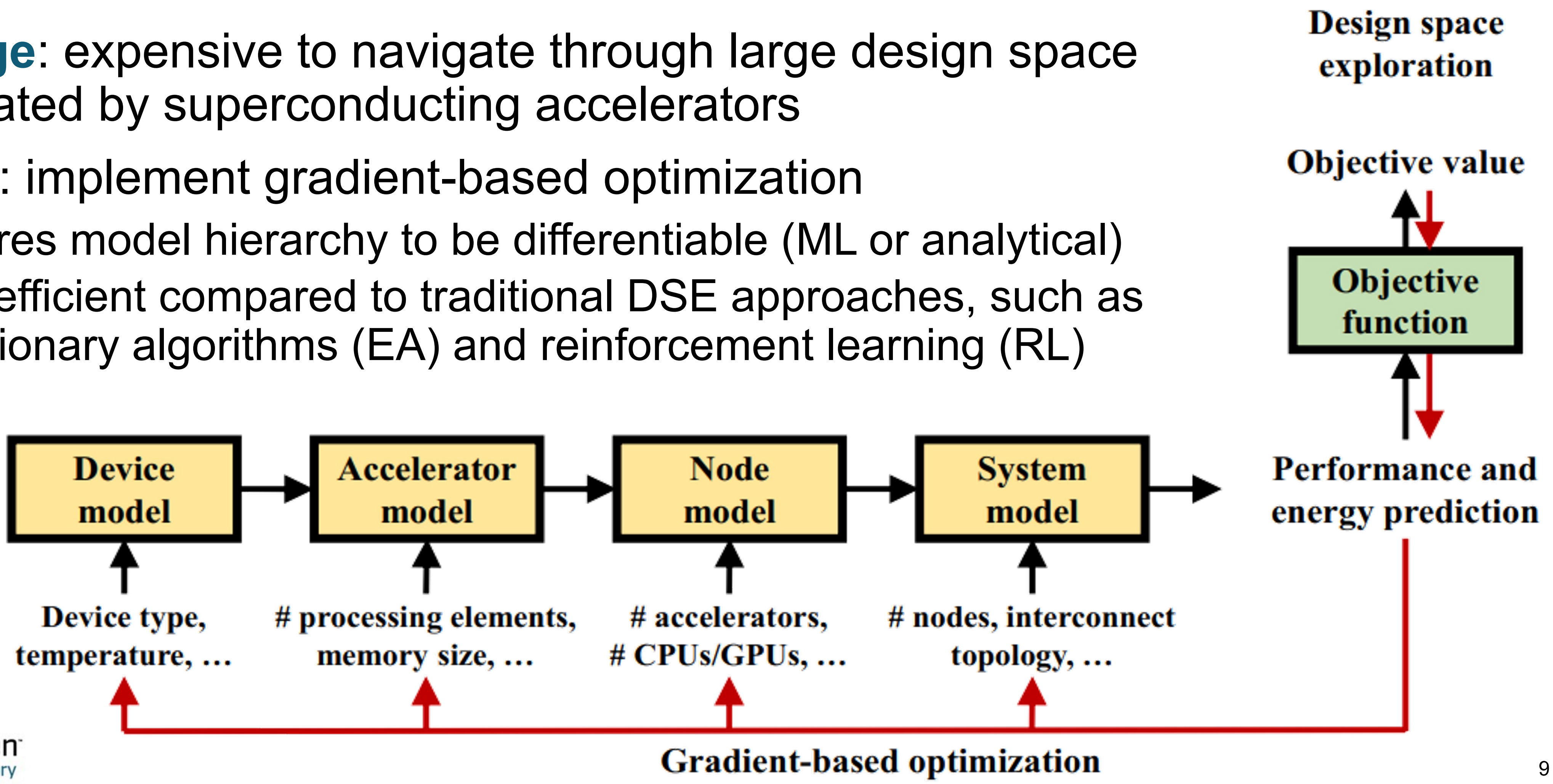


ML-enhanced Design Space Exploration (DSE)

Challenge: expensive to navigate through large design space aggravated by superconducting accelerators

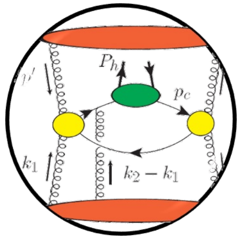
Solution: implement gradient-based optimization

- Requires model hierarchy to be differentiable (ML or analytical)
- More efficient compared to traditional DSE approaches, such as evolutionary algorithms (EA) and reinforcement learning (RL)

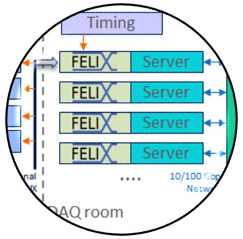


Real-time Data Reduction

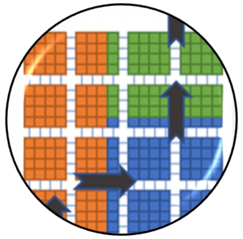
Challenge: The Electron-Ion Collider (EIC) subsystem has a high noise/background rate that requires real-time data reduction computationally: dRICH, far detectors, calorimeters
Solution: Provide a specialized algorithm and hardware for efficient and high-throughput real-time AI data reduction



NP Physics
- Diverse topology
- Stringent system Ctrl
- Max data preservation



Streaming DAQ
- New physics capability accessible only via streaming DAQ
- Example: adopted for sPHENIX and EIC
- Requires data reduction computationally



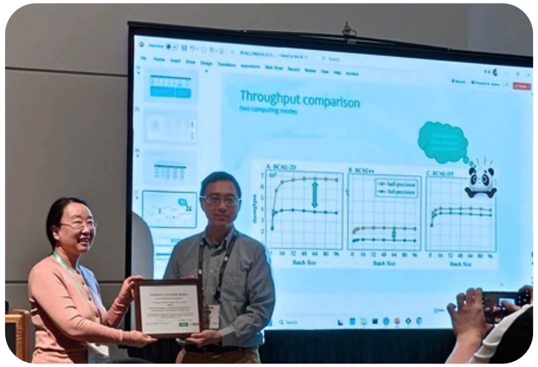
Opportunities for AI Enhancement
- Specialized AI algorithm for reliable and high-performance data reduction
- Novel hardware emerging for high-throughput AI computing

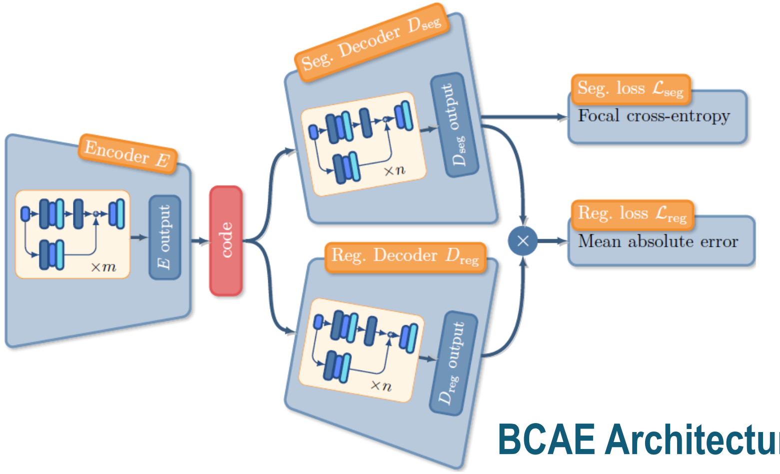
Physics need → Streaming DAQ → Opportunity for real-time AI → Enhanced physics program

Step 1: Real-time AI Algorithm

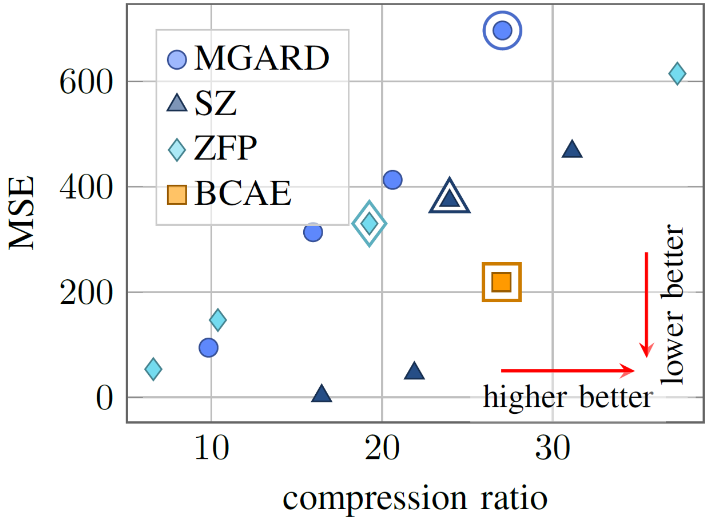
Bicephalous Convolutional Autoencoder (BCAE) that performs data compression and noise filtering in one step:

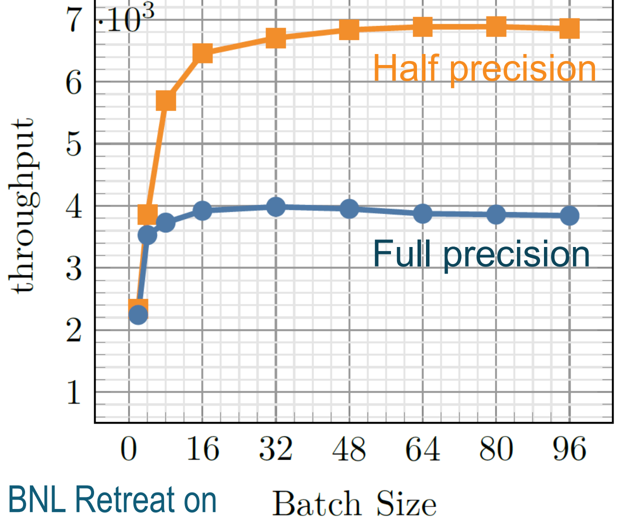
- Validating on (simulated) sPHENIX TPC 3D voxel data
- *Paper award at Data Reduction Workshop (SC23)





BCAE Architecture





Jin Huang
Yuhui (Ray) Ren

BNL Retreat on AI/ML for EIC

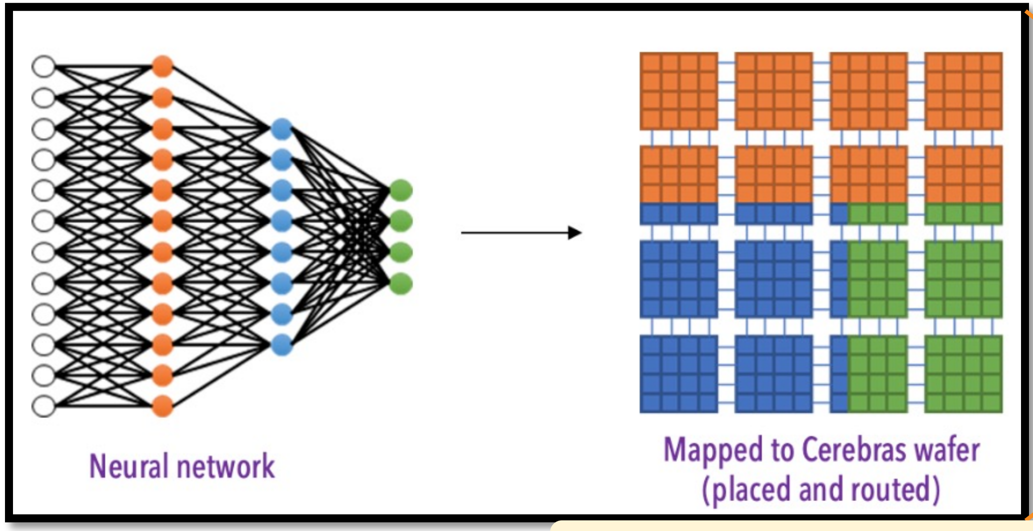
*Huang et al. SC23, DOI: 10.1145/3624062.3625127 arXiv:2310.15026

Step 2: Real-time AI Accelerator

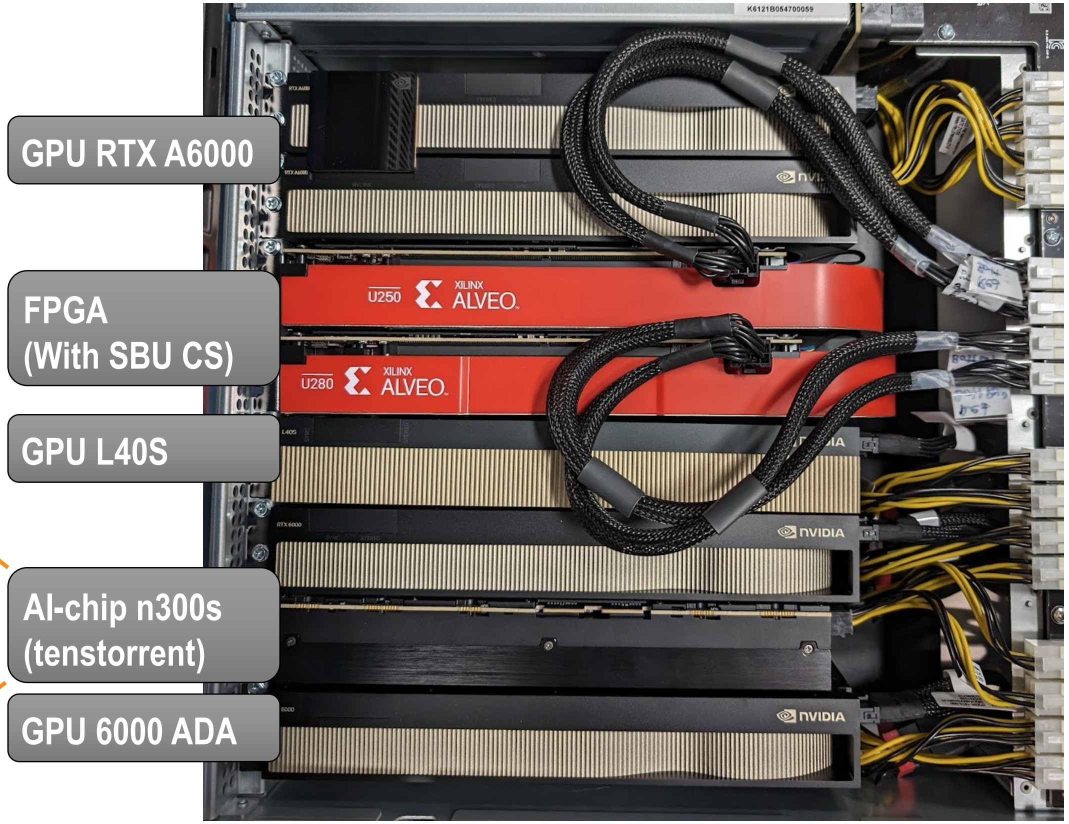
A new family of AI chips is emerging with non-von Neumann architectures.

BCAE test on GraphCore IPU and Groq Card

First installation of AI chip in experiment: tenstorrent n300s, 240 MB SRAM, 540 TFLOPS for FP8, 160 RISC-V cores



Real-time AI Test Stand
Deployed in sPHENIX IR on DAQ network



*Huang et al. SC23, DOI: 10.1145/3624062.3625127 arXiv:2310.15026

Summary

- Vibrant portfolio of activities in multiple dimensions of the **Novel Architectures for AI** space
- Research funded by multiple agencies and sources: Department of Energy, Department of Defense and Laboratory Directed Research and Development.
- Motivated by challenges posed by the experimental science workflows
- Synergy with SBU's research as evidenced by collaborations and high-bandwidth interactions – with significant room to expand