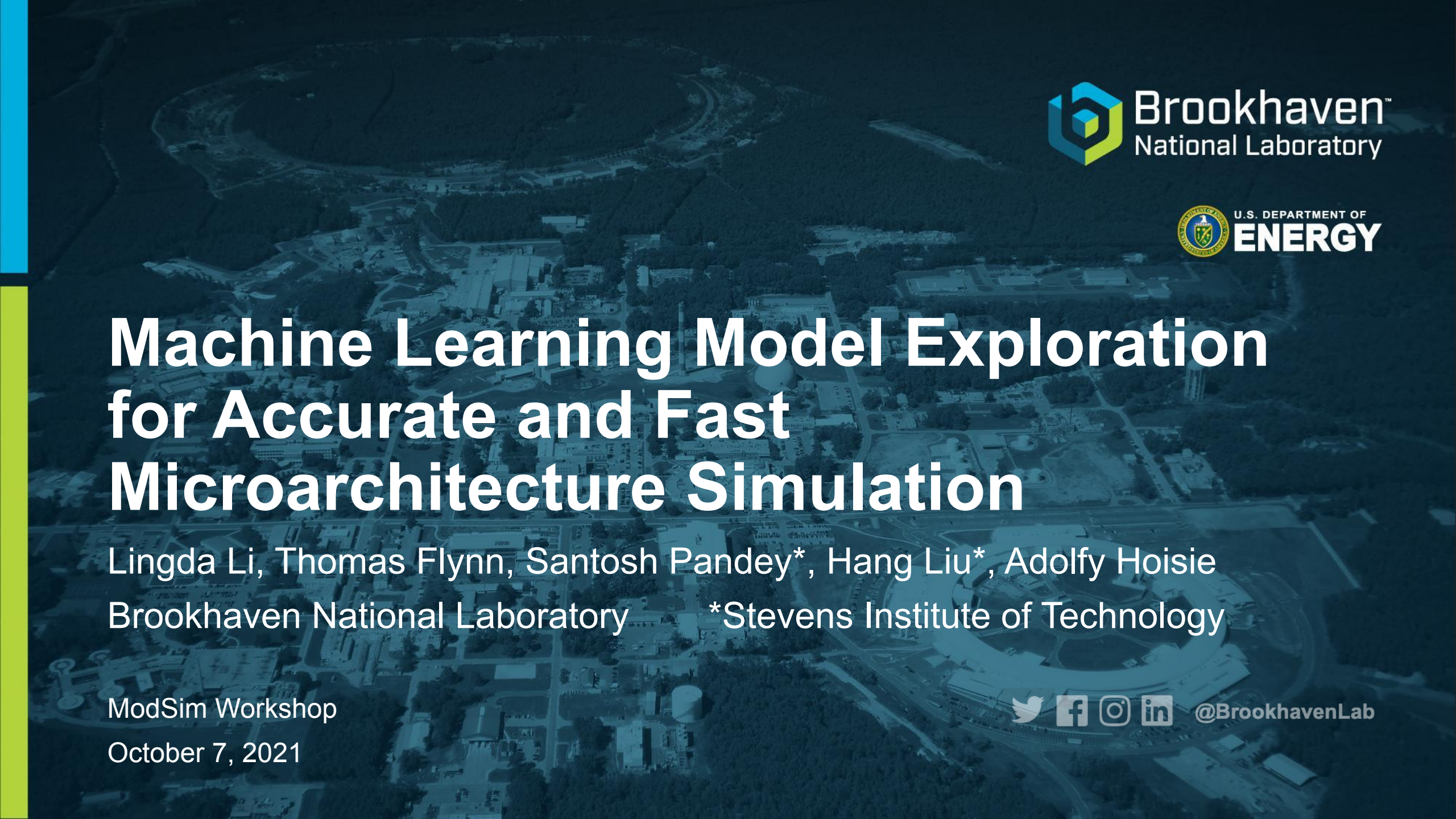


Machine Learning Model Exploration for Accurate and Fast Microarchitecture Simulation

Lingda Li, Thomas Flynn, Santosh Pandey*, Hang Liu*, Adolfy Hoisie

Brookhaven National Laboratory *Stevens Institute of Technology

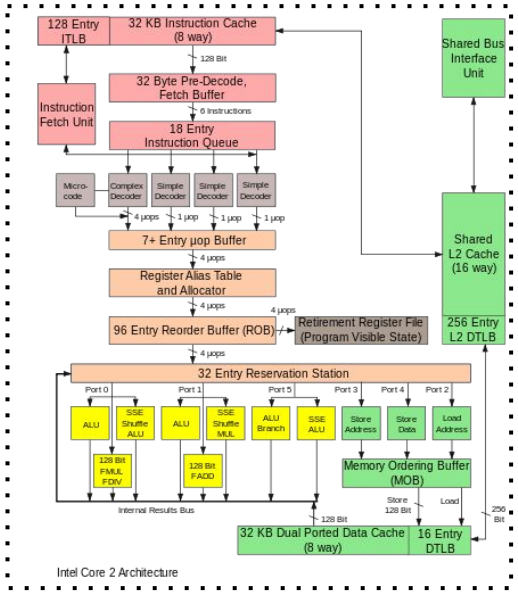
ModSim Workshop

October 7, 2021

ML-based Microarchitecture Simulation

Traditional Simulation

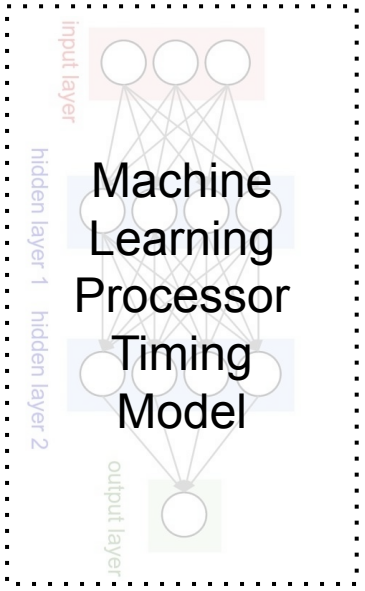
Instruction



Latency

ML-based Simulation

Instruction



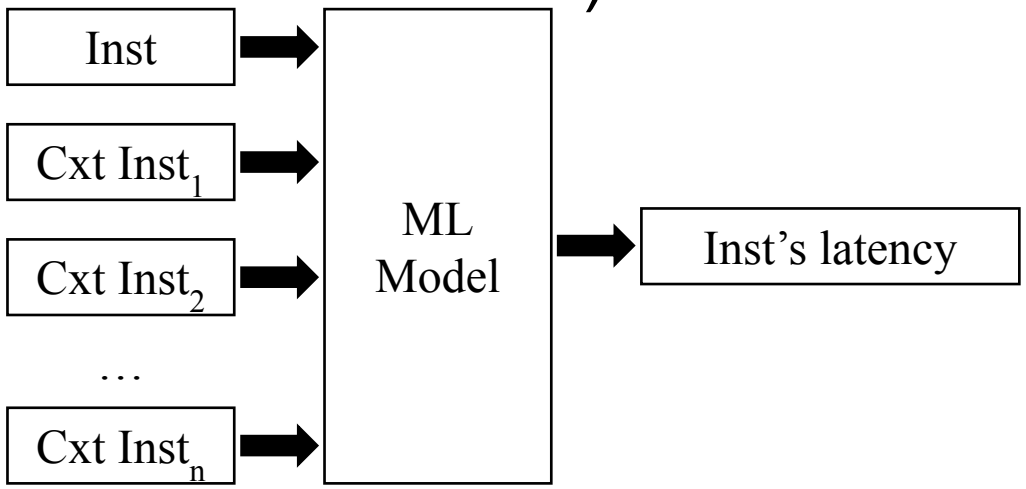
Latency

On-the-fly (context) instructions

- Traditional discrete-event simulators suffer from long simulation time
- Idea: decompose discrete-event simulation into instruction latency prediction, and use ML for the latter
- Advantage: change irregular discrete-event simulation to regular and highly parallel ML inference

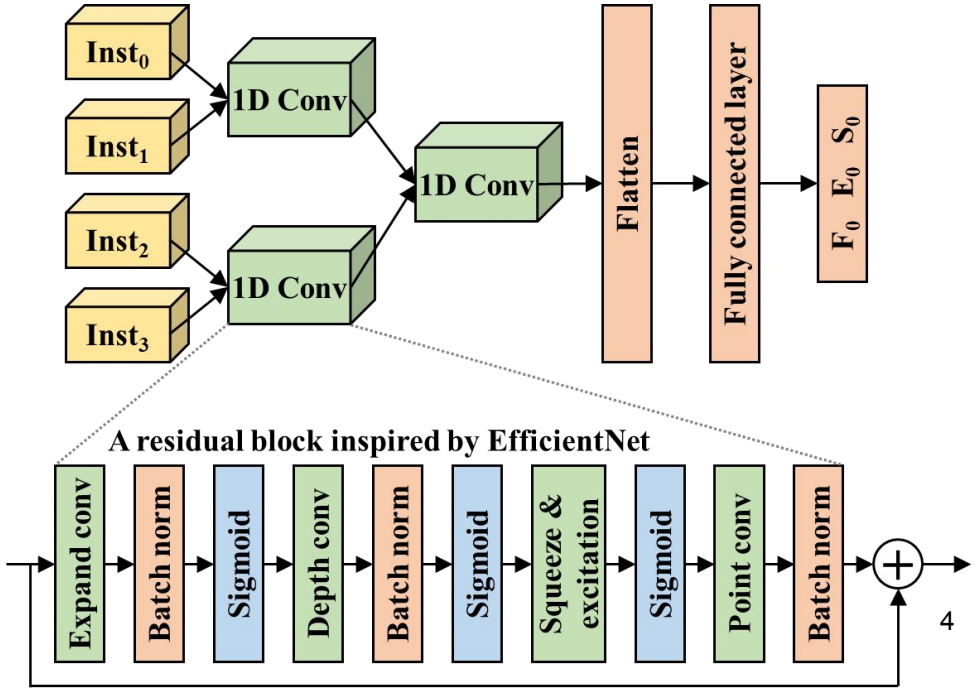
Instruction Latency Prediction Framework

- Instruction latency depends on processor states, which in turn depend on all instructions in the processor (i.e., context instructions)
- ML models to learn the impact of context instructions
- Input: to-be-predicted instruction and context instructions
  - Together they decide the current processor states
- Each instruction has the following features
  - Static properties: operation type, register indices, etc.
  - Dynamic properties: cache access level, memory address dependency, etc.



ML Model Architecture

- Sequence-oriented models
  - Long short-term memory (LSTM): integrate context instructions in order
  - Transformer: attention mechanism to reason instruction relationships
- Convolutional neural network (CNN) models
  - Convolution operations to summarize instructions hierarchically
  - Residual blocks to enable deep CNNs



ML Model Evaluation

- CNN models achieve better prediction accuracy while requiring less computation compared with LSTM and Transformer models
- 7RB+2F achieves the most accurate prediction, and other CNN models represent different trading points between computation and accuracy

CNN models | Sequence-oriented models

	2F	3C+2F	5C+2F	7C+2F	7RB+2F	LSTM	Transformer
Computation (MFlops)	5.7	8.1	21.4	50.8	93.3	119	1185
Latency error	57%	26%	16%	6.2%	5.1%	19%	17%
Simulation error	18%	1.9%	2.0%	2.3%	0.96%	2.4%	2.4%

C: convolution layer
F: fully-connected layer
RB: residual block

Phase-level Simulation Accuracy

- Compare the simulation CPI across execution phases
- The CPI curves across different phases are similar to that of gem5 which ML models are trained against