

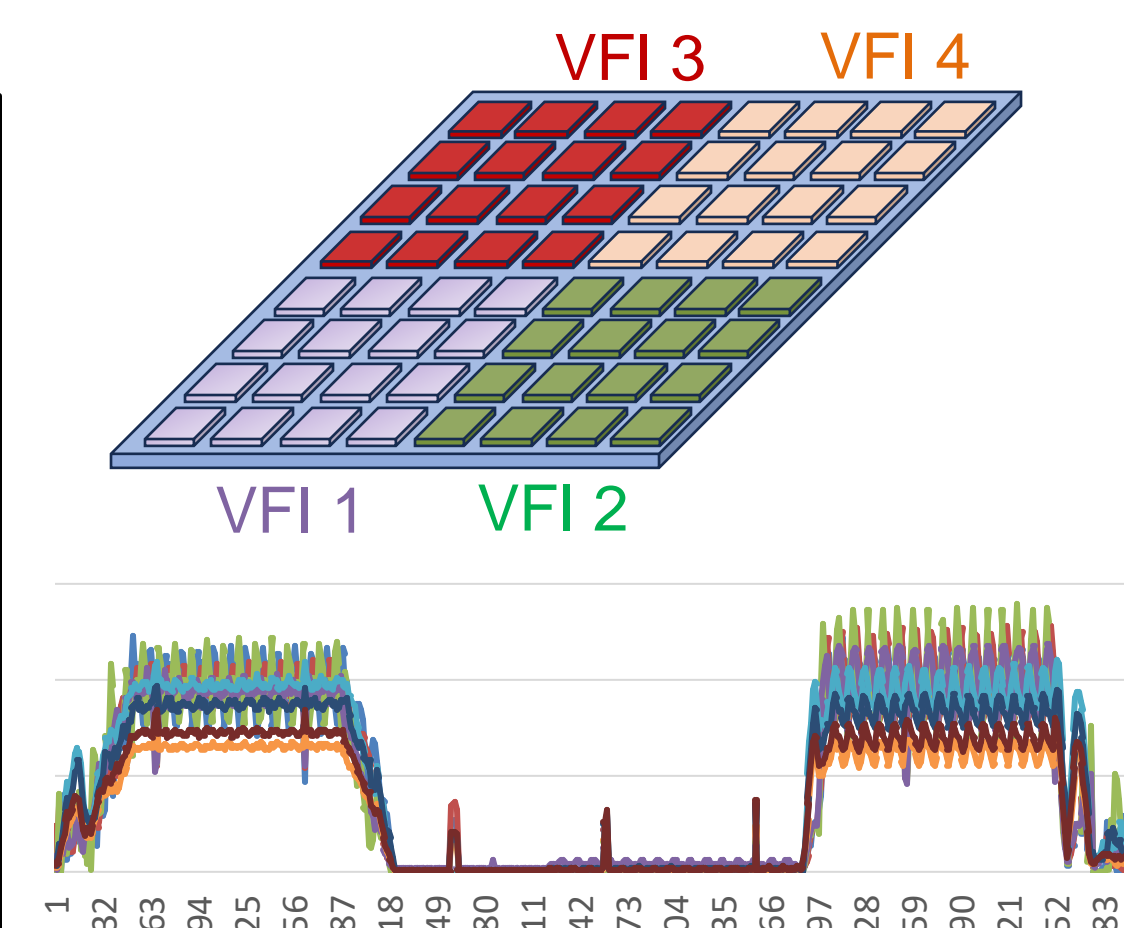
Power, Performance, and Thermal Trade-offs in Manycore Architectures

Synergies between ML and Computing systems

Presenter: Gaurav Narang, Advisors: Janardhan Rao Doppa, Partha Pratim Pande
Washington State University, Pullman, USA

Problem 1. Dynamic Power Management

- Voltage-Frequency Island (VFI) based large manycore systems
- Share Voltage/Frequency (V/F) among multiple cores and links
- Scalable solution
- Dynamically fine-tune V/F
- Time-varying workload features



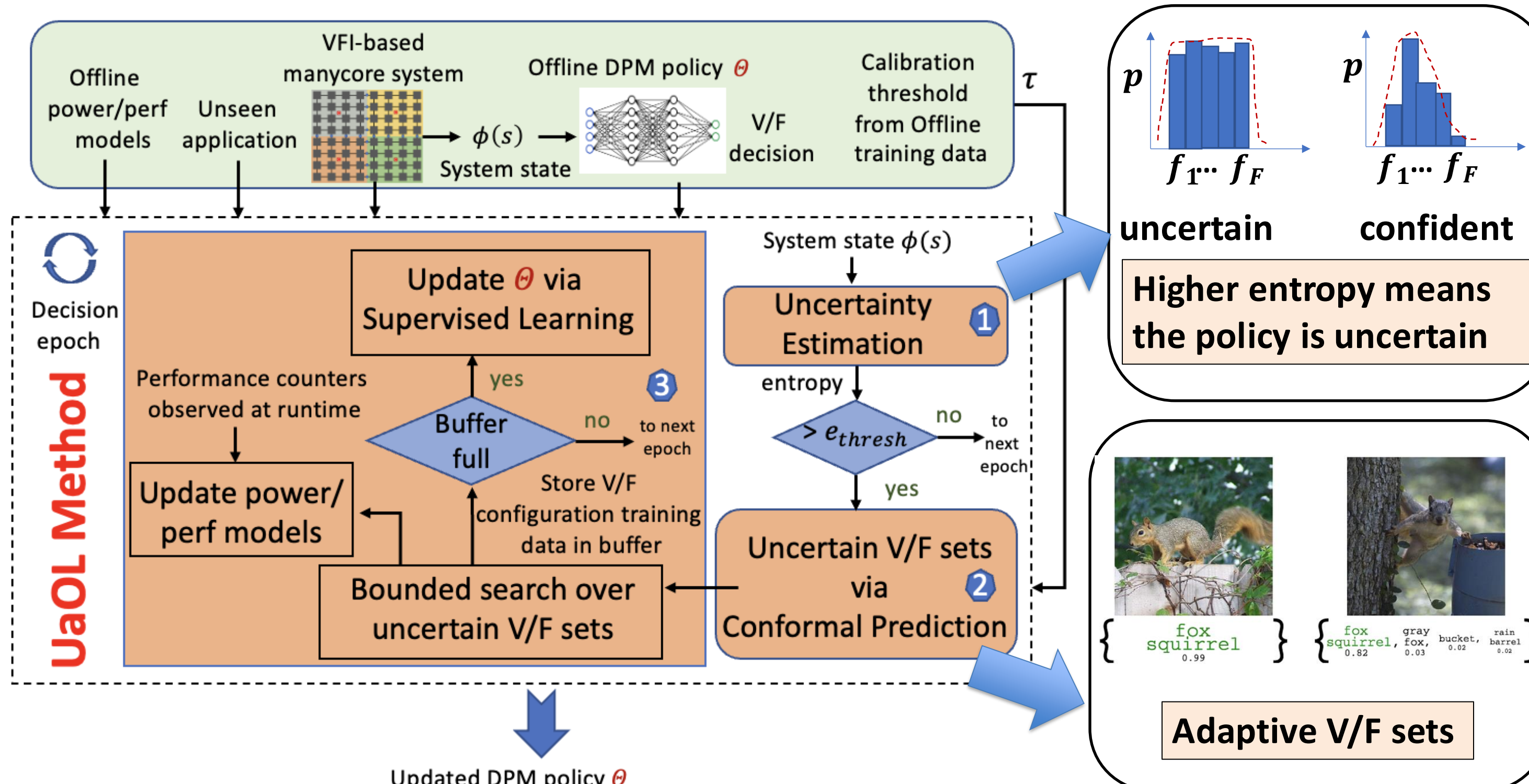
Time-varying workload features (Instruction per cycle and Inter-VFI traffic)

Challenges

- Large decision space
 - Exponential in the number of VFIs and V/F levels: $(\#V/Fs)^{\#VFIs}$
- Same application can have multiple phases
- Exponential growth of new and unknown applications

Problem: How do we learn dynamic power management policies for unseen applications at runtime?

Uncertainty aware Online Learning

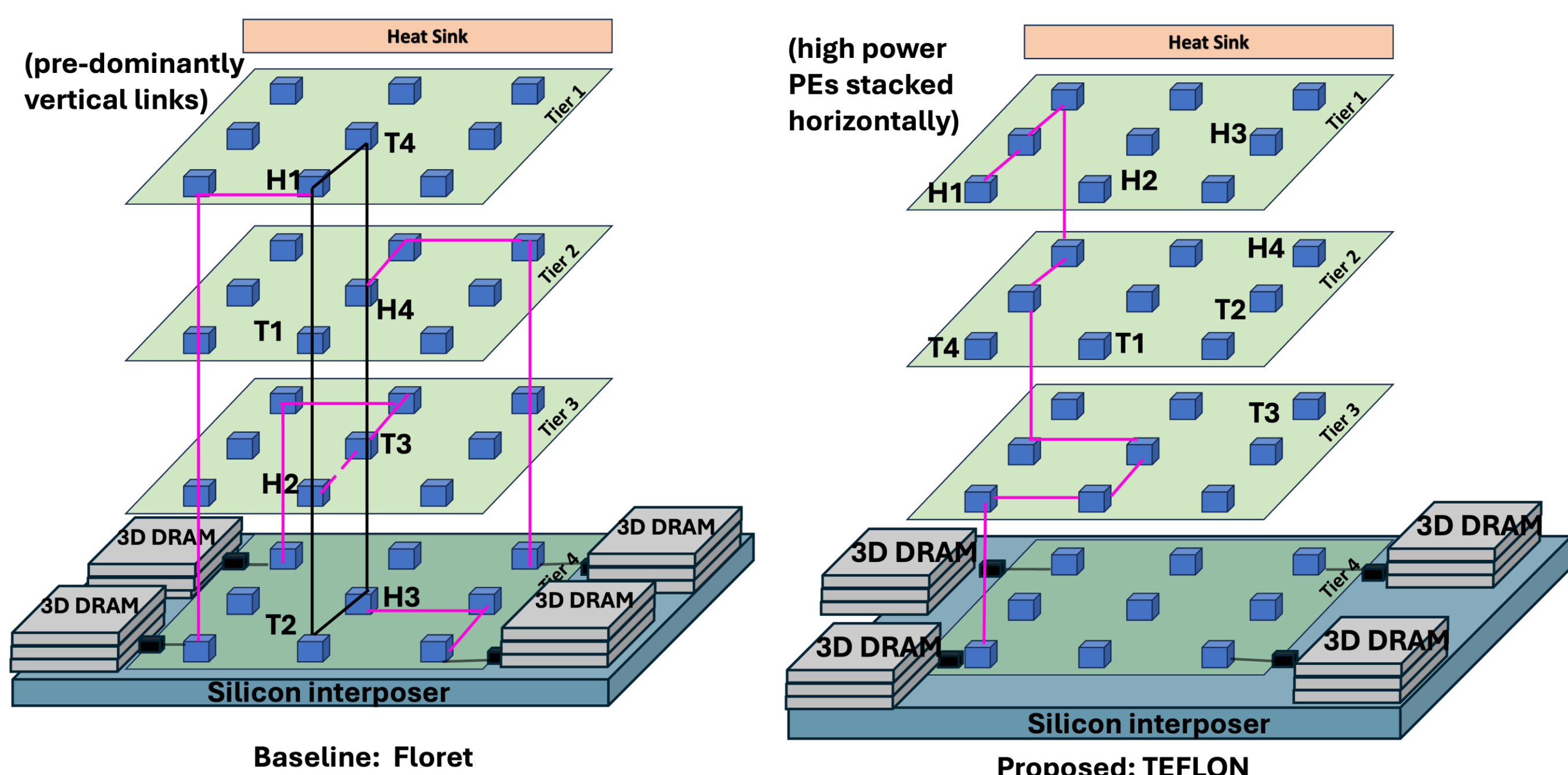


Key idea #1: Perform online learning only on uncertain/unknown system states ($entropy > e_{thresh}$) vs. ALL system states

Key idea #2: Perform bounded search over uncertain V/F sets vs. ALL V/F values to create supervised training data

Problem 2. Thermally Efficient Dataflow-aware NoC

How can neural layers be mapped to a 3D architecture to accelerate DNN inferencing without creating thermal hotspots?

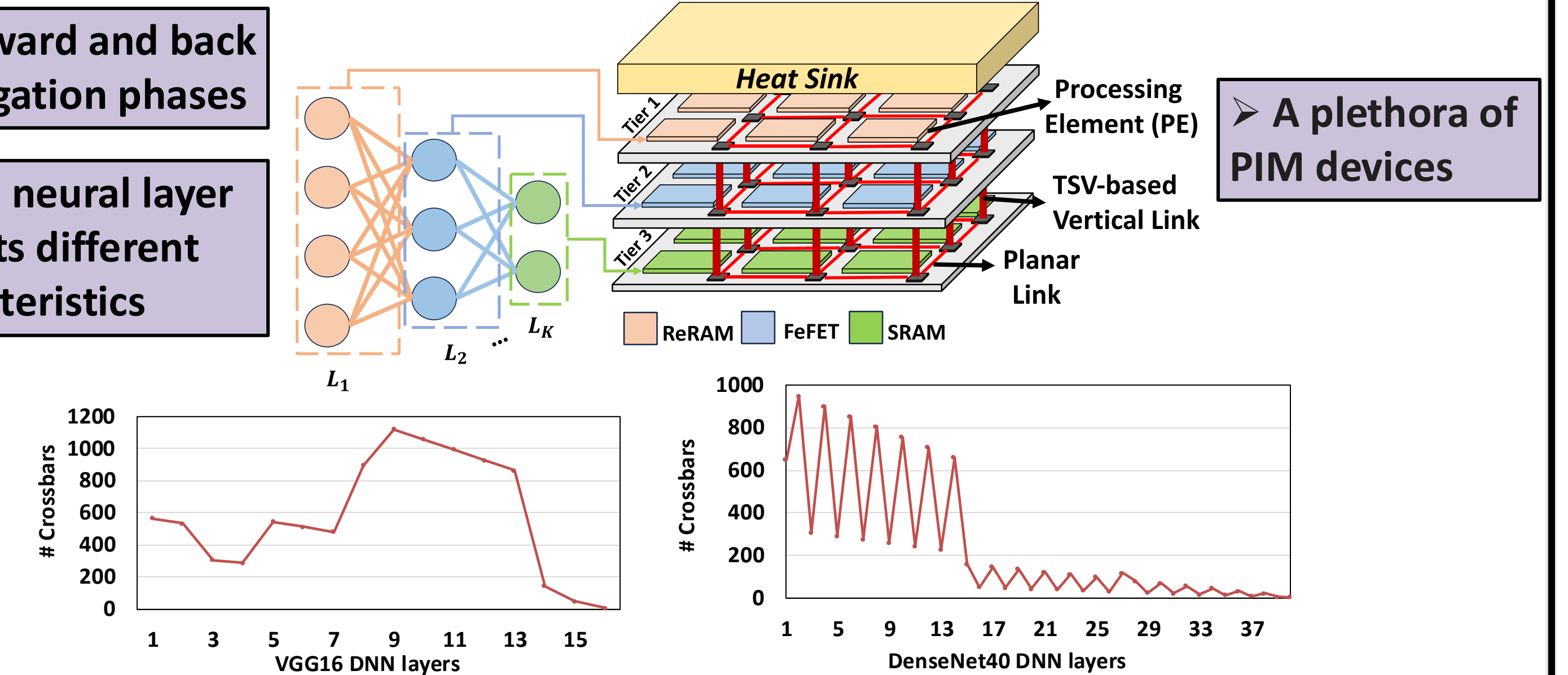


- Stacking contiguous neural layers along one specific vertical column will give high performance, but it will lead to thermal bottleneck
- We map the neural layers to
 - contiguously located PEs
 - not placing multiple high-power consuming PEs away from the heat sink and along one specific vertical column.

Problem 3. Heterogeneous PIM-enabled DNN Training

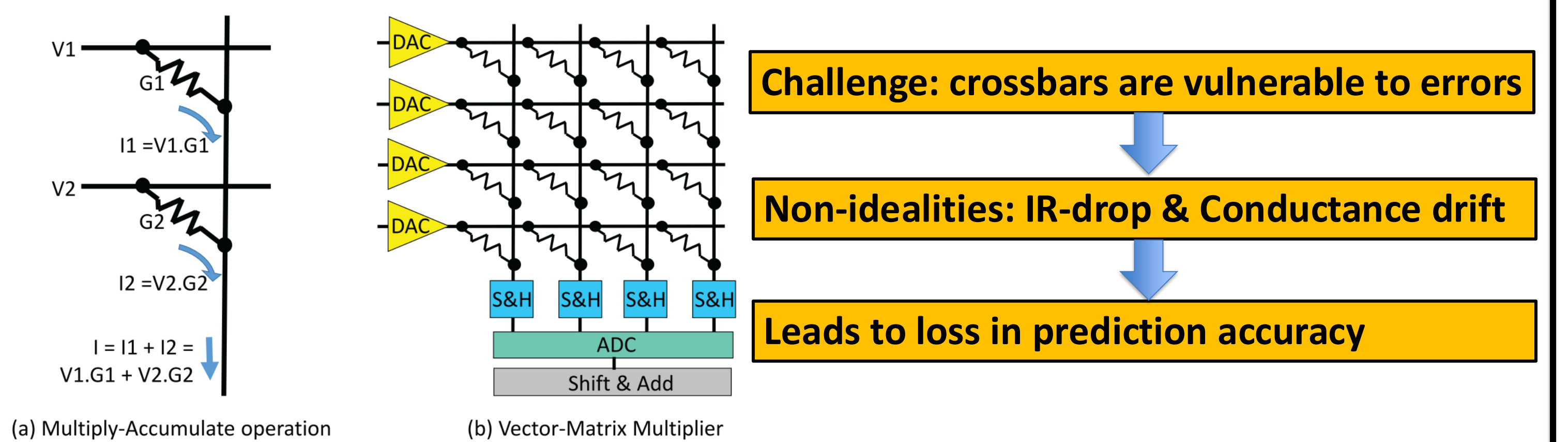
How can the varying characteristics of DNN neural layers and various PIM devices be exploited to accelerate DNN training?

- Forward and back propagation phases
- Each neural layer exhibits different characteristics

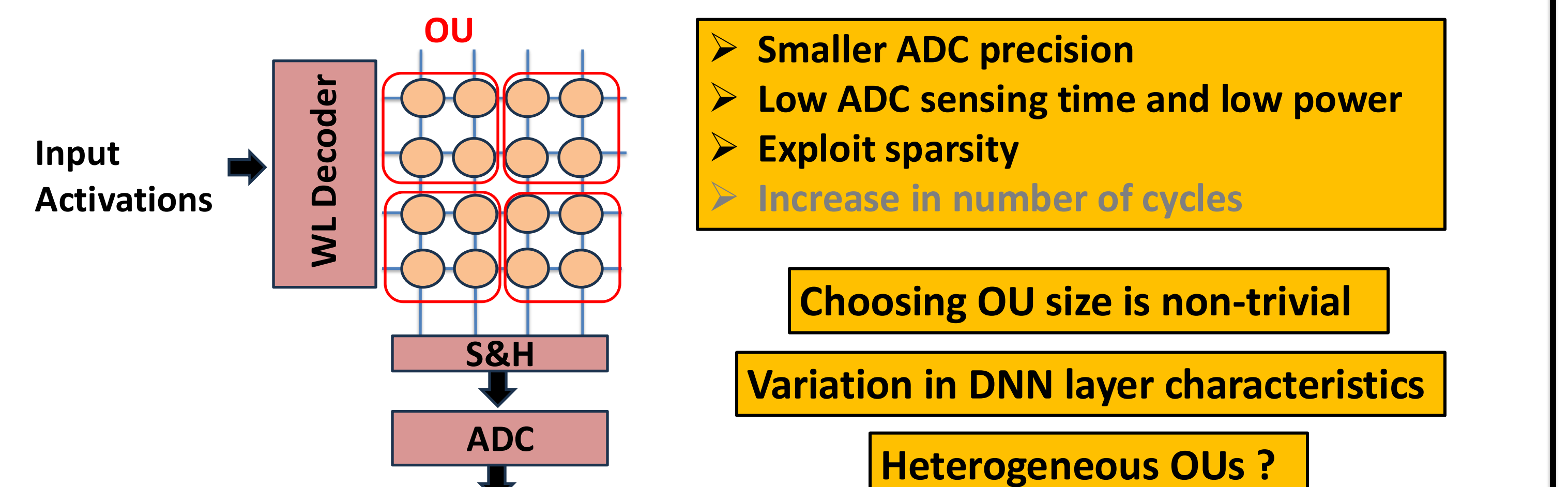


Key idea: A heterogeneous architecture that combines the benefits of multiple devices in a single platform can enable energy-efficient and high performance DNN training

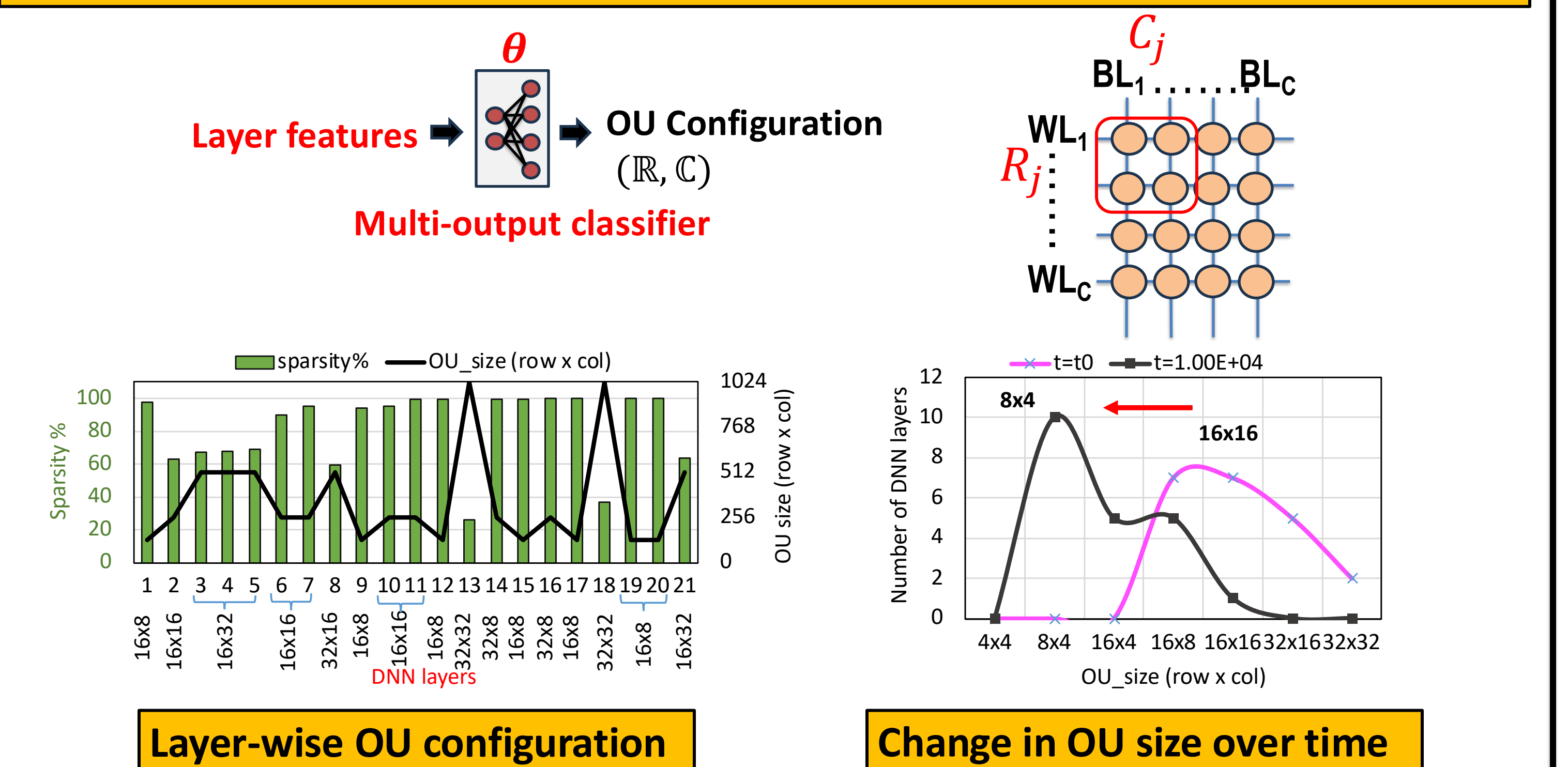
Problem 4. Learning to Optimize Layer-wise OU Sizes



Operation Units (OU) based Computation



How to learn to optimize DNN layer-wise OU configurations for non-ideal NVM crossbars at runtime?



Thesis Contribution

- Goal: High-performance, energy-efficient, and reliable computing systems
- Exploit the synergies between Machine learning (ML) and computing systems
- Exploit the heterogeneity in the computational kernels behind deep learning models
- Novel ML-based dynamic resource management algorithms

