THE UNIVERSITY OF CHICAGO

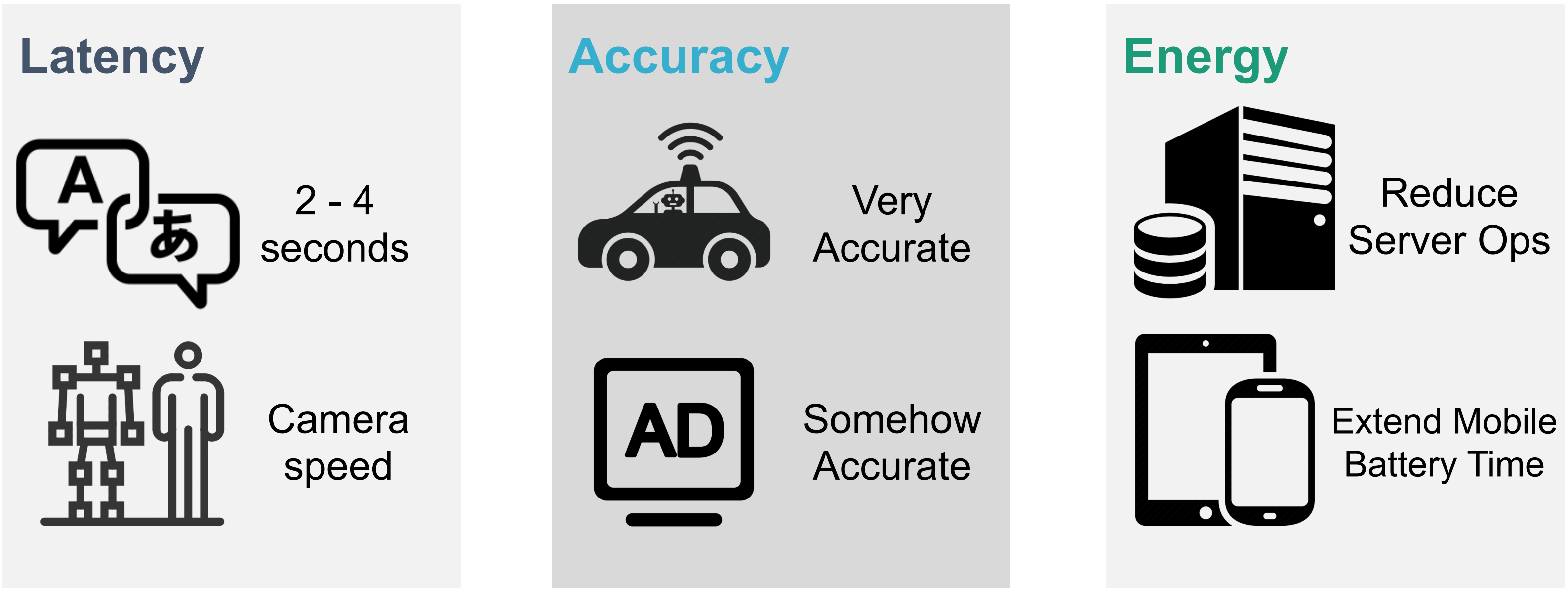
# Accurate Anytime Learning for Energy and Timeliness in Software Systems

Chengcheng Wan (cwan@uchicago.edu)

## Overview

### Motivation

- Many software systems incorporate DNN inference
  - ○ DNN inference consumes substantial time and resources
  - ○ DNN inference accuracy affects software functionality
- Software often faces strict accuracy-latency-energy requirements
  - ○ Violating requirements may lead to severe failures!

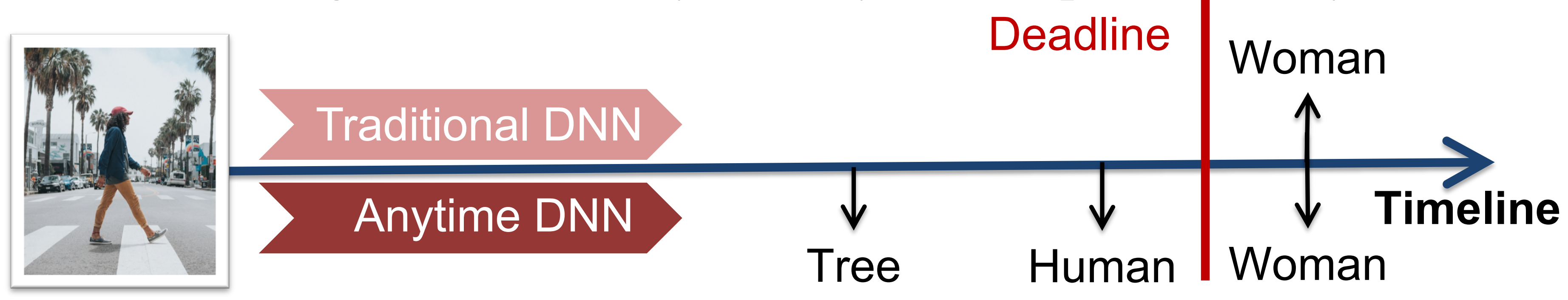

Latency
2 - 4 seconds
Camera speed
Accuracy
Very Accurate
Somehow Accurate
Energy
Reduce Server Ops
Extend Mobile Battery Time

### Problem definition

- How to adjust DNN and system power setting to
  - ○ Maximize accuracy with energy budget and inference deadline
  - ○ Minimize energy with accuracy goal and inference deadline

### Our solution

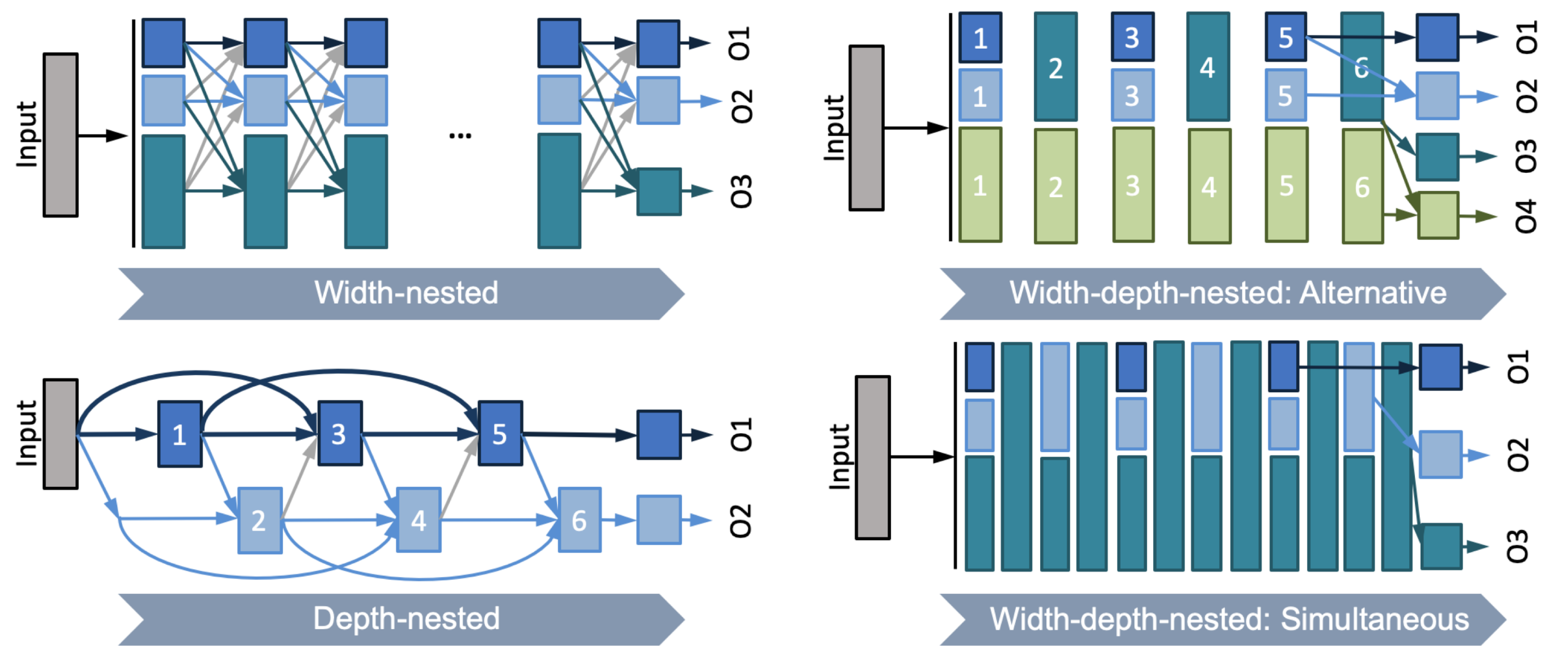
- Anytime neural networks
  - ○ Generating increasingly more accurate results as time goes
  - ○ Balance goals of accuracy and anytime output flexibility


Deadline
Traditional DNN
Anytime DNN
Timeline
Tree
Human
Woman
Woman

- Adaptive DNN inference management
  - ○ Run-time feedback-based estimation and control
  - ○ Coordinate application and system adaptation

## Anytime Neural Networks [1]

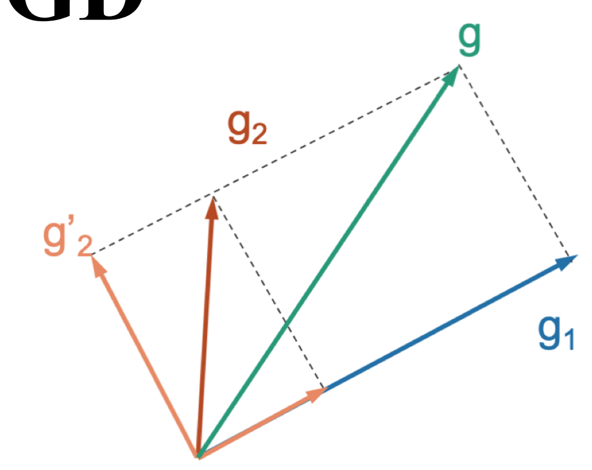
### Network design: nested architecture


Width-nested
Width-depth-nested: Alternative
Depth-nested
Width-depth-nested: Simultaneous

- Inference conducted in stages (marked in different colors)
- Every stage getting deeper or wider or both than previous ones
- Results, O1, O2, O3, …, output after each stage
- Maximizing re-use of intermediate state across successive stages

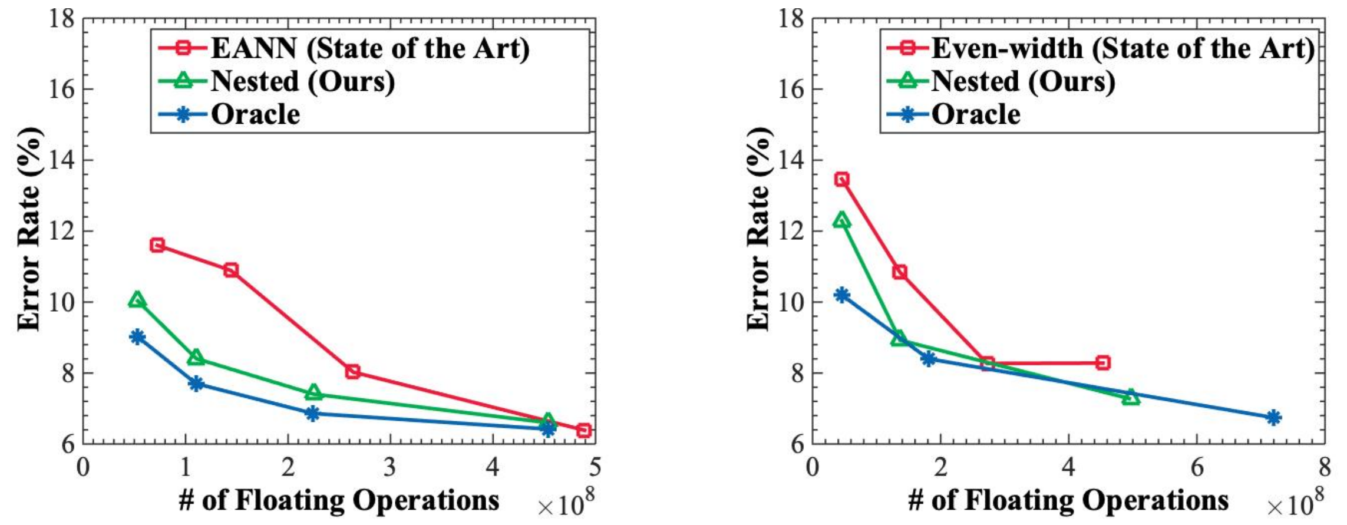
### Network training: orthogonalized SGD

- Re-balances task-specific gradients
- Subsequent outputs do not interfere with how earlier outputs desire to move parameters



### Experimental Results

We offer much better accuracy-FLOPs trade-offs than previous state of the art, and come close to the infeasible Oracle.

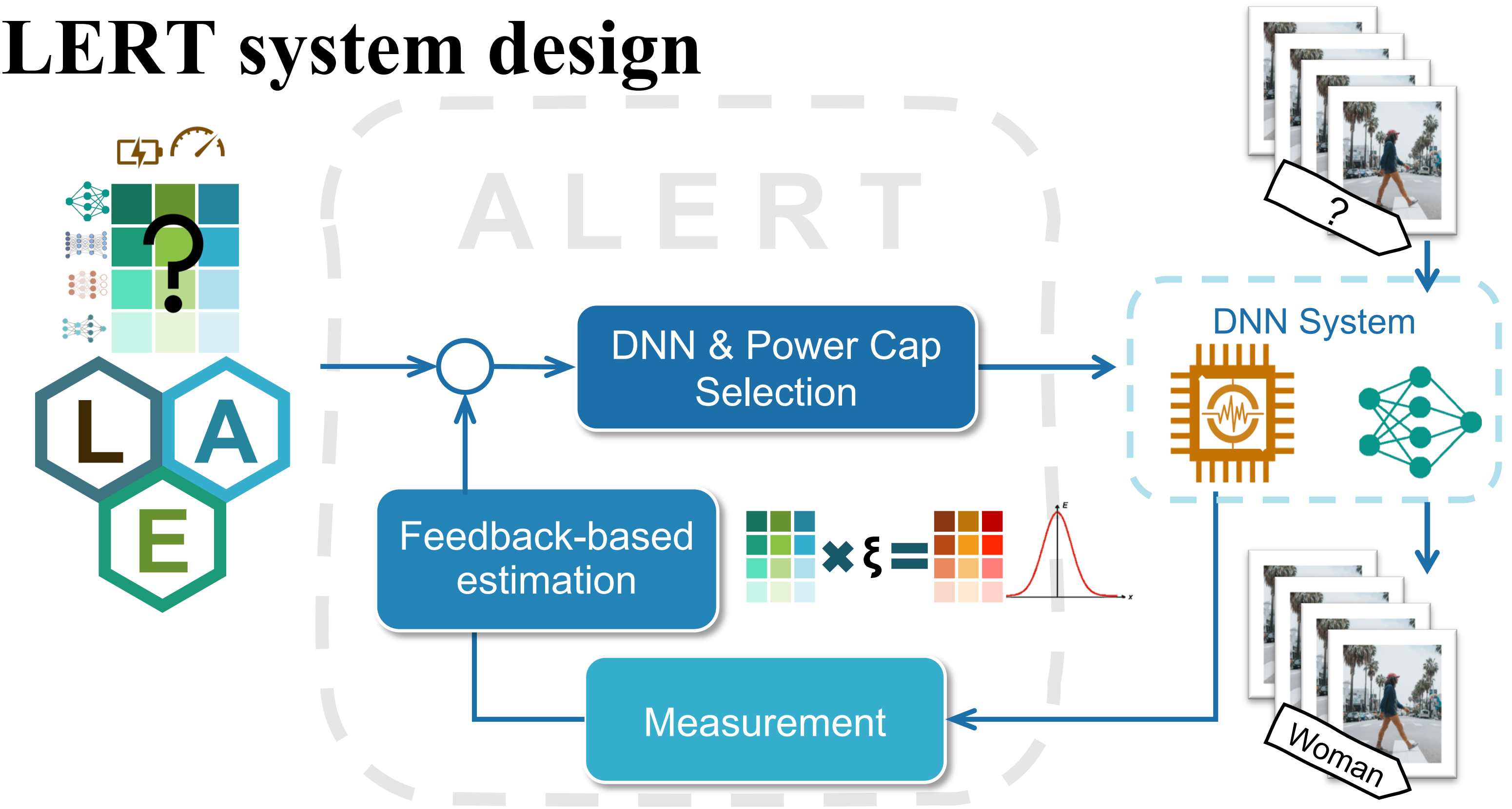


CIFAR image classification with nested ResNet. (Smaller is better)

[1] **Chengcheng Wan**, Henry Hoffmann, Shan Lu, and Michael Maire. Orthogonalized SGD and Nested Architectures for Anytime Neural Networks. ICML, 2020.

## Adaptive DNN Inference Management [2]

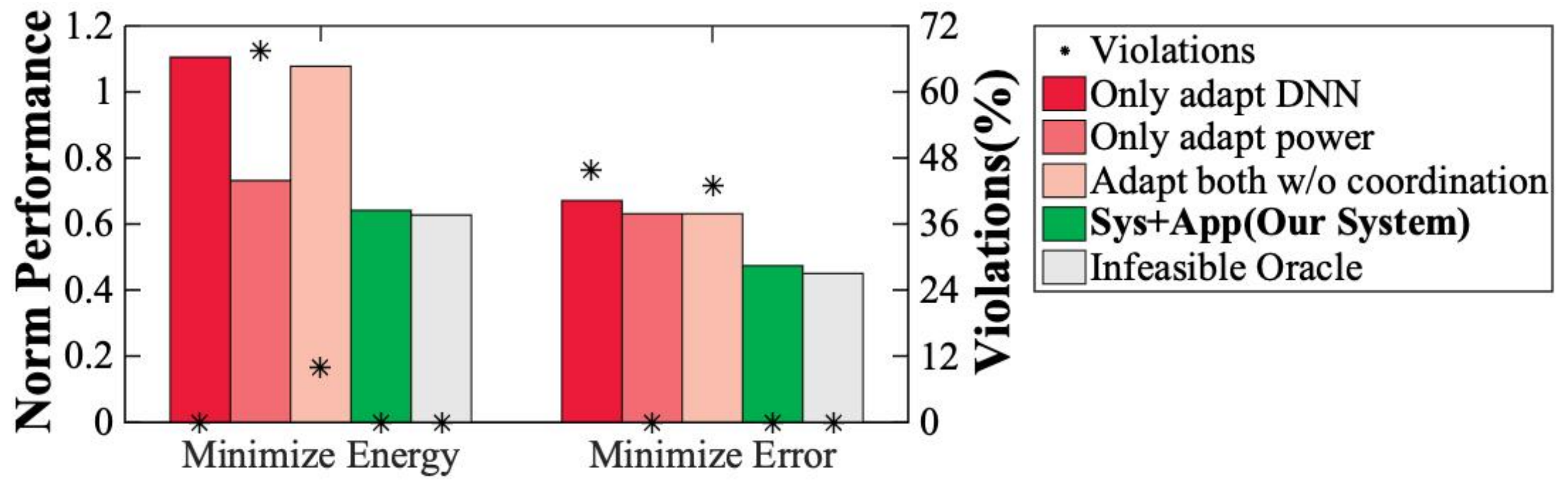
### ALERT system design


ALERT
DNN & Power Cap Selection
DNN System
Feedback-based estimation
Measurement

- Two knobs
  - ○ Application level: DNN, including anytime and traditional ones
  - ○ System level: discrete system power cap setting
- Global slowdown factor ξ
  - ○ The environmental difference between profiling and runtime
  - ○ Estimated by Kalman filter with all recent history
  - ○ Modelled as a random variable to reflect the system volatility
- Runtime estimation
  - ○ Latency: profiling data x ξ
  - ○ Accuracy: expectation with estimated latency distribution
  - ○ Energy: estimate idle power with Kalman filter

### Experimental Results

Our System outperforms state of the art and static oracle.



Average performance normalized to static oracle. (Smaller is better)

[2] **Chengcheng Wan**, Muhammad Santriaji, Eri Rogers, Henry Hoffmann, Michael Maire, and Shan Lu. ALERT: Accurate Learning for Energy and Timeliness. USENIX ATC, 2020.