EDEN

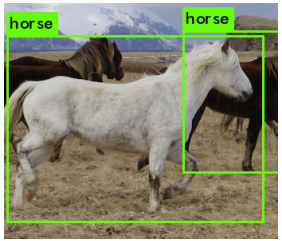
Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM

Skanda Koppula Lois Orosa A. Giray Yaglikci
Roknoddin Azizi Taha Shahroodi Konstantinos Kanellopoulos Onur Mutlu

ETH zürich SAFARI

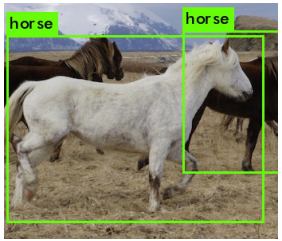
Motivation

Deep neural networks (DNNs) are critical
in computer vision, robotics, and many other domains



Motivation

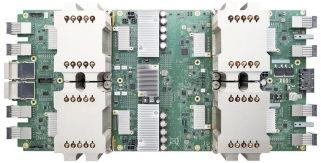
Deep neural networks (DNNs) are critical in computer vision, robotics, and many other domains



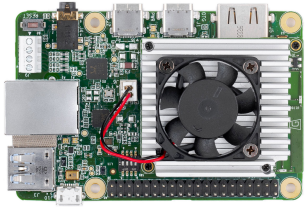
Modern DNN inference platforms use DRAM

Mobile CPUs

GPUs



Data Center Accelerators



Edge-device Accelerators

Challenges of DNN Inference

DRAM has high energy consumption

- **25% to 70% of system energy** is consumed by DRAM in common DNN inference accelerators

DRAM can bottleneck performance

- Potential **19% speedup** by **reducing DRAM latency** on CPU for some DNNs

Observations

1. DNNs have an **intrinsic robustness to errors** in the input, weight, and output data types

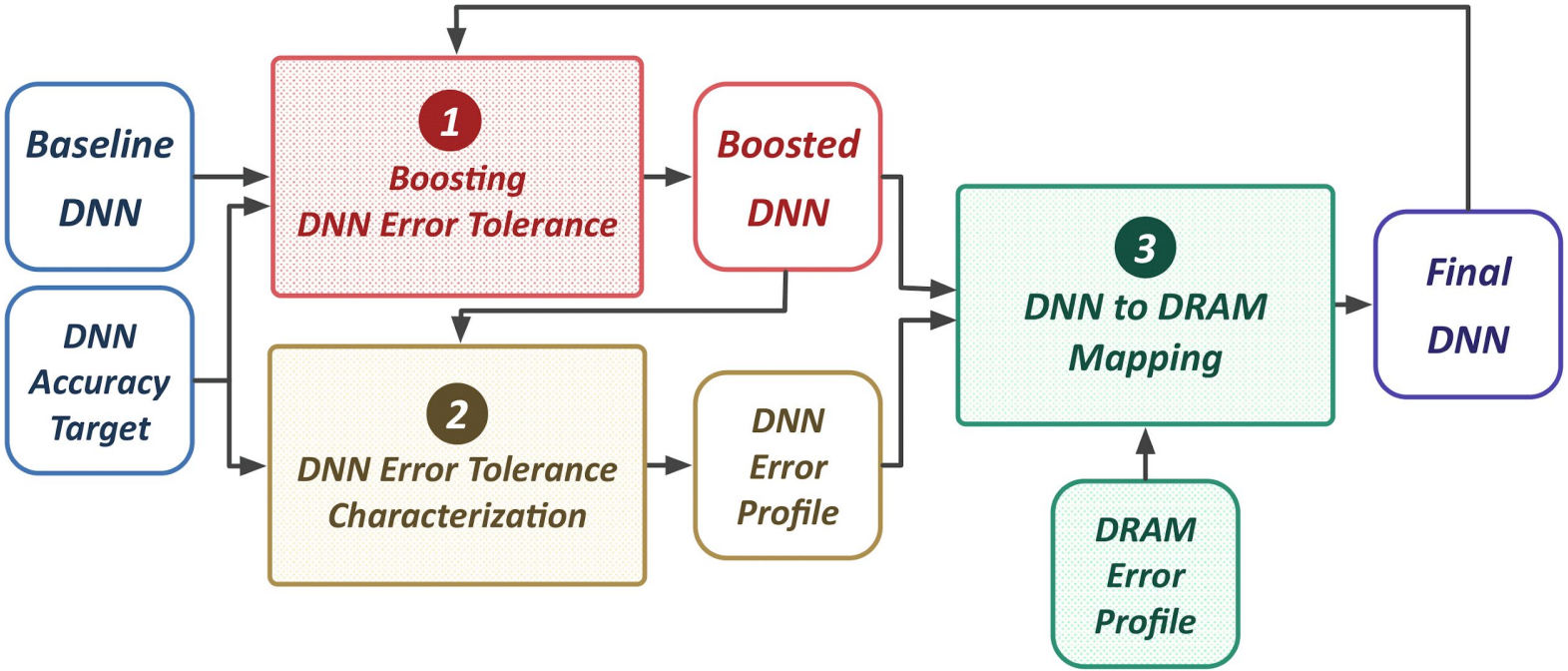
2. We can **reduce DRAM energy consumption** and **latency** if we tolerate **more bit errors**

Insight

1. DNNs have an **intrinsic robustness to errors**

Approximate DRAM (voltage and latency-scaled DRAM) can provide **higher energy-efficiency** and **performance** for **error-tolerant DNN inference** workloads

We propose **EDEN**, a mechanism to enable **accurate DNN inference** on **approximate DRAM**



Key Results

- CPU: **21% DRAM energy reduction, 8% speedup**
- GPU: **37% DRAM energy reduction**
- DNN Accelerators: **31% DRAM energy reduction**

While maintaining a **user-specified accuracy target** within **1% of the original DNN accuracy**

EDEN

Enabling Energy-Efficient, High-Performance Deep Neural Network Inference Using Approximate DRAM

Skanda Koppula Lois Orosa A. Giray Yaglikci
Roknoddin Azizi Taha Shahroodi Konstantinos Kanellopoulos Onur Mutlu

ETH zürich SAFARI